REVISITING RECOGNIZING TEXTUAL ENTAILMENT FOR EVALUATING NATURAL LANGUAGE PROCESSING SYSTEMS

by

Adam Poliak

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

October, 2020

Abstract

Recognizing Textual Entailment (RTE) began as a unified framework to evaluate the reasoning capabilities of Natural Language Processing (NLP) models. In recent years, RTE has evolved in the NLP community into a task that researchers focus on developing models for. This thesis revisits the tradition of RTE as an evaluation framework for NLP models, especially in the era of deep learning.

Chapter 2 provides an overview of different approaches to evaluating NLP systems, discusses prior RTE datasets, and argues why many of them do not serve as satisfactory tests to evaluate the reasoning capabilities of NLP systems. Chapter 3 presents a new large-scale diverse collection of RTE datasets (DNC) that tests how well NLP systems capture a range of semantic phenomena that are integral to understanding human language. Chapter 4 demonstrates how the DNC can be used to evaluate reasoning capabilities of NLP models. Chapter 5 discusses the limits of RTE as an evaluation framework by illuminating how existing datasets contain biases that may enable crude modeling approaches to perform surprisingly well.

The remaining aspects of the thesis focus on issues raised in Chapter 5. Chapter 6 addresses issues in prior RTE datasets focused on paraphrasing and presents a high-quality test set that can be used to analyze how robust RTE systems are to paraphrases. Chapter 7 demonstrates how modeling approaches on overcoming biases, e.g. adversarial learning, can enable RTE models overcome biases discussed in Chapter 5. Chapter 8 applies these methods to the task of discovering emergency needs during disaster events.

Keywords: Recognizing Textual Entailment, Natural Language Inference, Natural Language Understanding, Computational Semantics, Natural Language Processing

Thesis Committee

Primary Reader:

Benjamin Van Durme
Associate Professor
Department of Computer Science
Johns Hopkins University

Secondary Readers:

Aaron Steven White
Assistant Professor
Department of Linguistics
University of Rochester

João Sedoc
Assistant Research Professor
Department of Computer Science
Johns Hopkins University

Acknowledgments

> Without [a team], it would be like going on a trip in your car if you knew where you wanted to go but didnt know how to get there. You might correctly be described as going nowhere.
>
> John Wooden

The team around me during graduate school was integral to my completion of this journey, and helped me figure out both where to go and how to get there. I am indebted to my colleagues, mentors, friends, and family for their support, encouragement, and constant feedback during this journey.

I am immensely thankful to Benjamin Van Durme, my Ph.D. advisor, who helped me develop my own research voice, taught me how to lead collaborations, and encouraged me to prioritize the important things in life. During my Ph.D. Ben knew exactly when and how to either tighten the ropes or give me the independence I needed. After our weekly meetings, I would often be more calm, excited, and inspired for the week ahead. I learned an incredible amount during those meetings, especially when the latex-ed agendas could not keep us from going off track.

I am also grateful to my other committee members Aaron White and João Sedoc. Their ideas impacted this thesis and long discussions with each greatly influenced my

career aspirations. Aaron helped supervise one of my first projects in Ben's lab, which planted the seeds for many ideas in this thesis. I am happy João chose to spend a year at JHU before starting at NYU. I enjoyed TA-ing his course and appreciate the many life lessons he taught me. João, I still owe a round of tennis. I would also like to thank my Graduate Board Oral committee members, Ido Dagan, David Yarowsky, and Kyle Rawlins. Your questions during my thesis proposal sharpened this thesis and led to many insights.

I was very fortunate to spend my time as a graduate student as a member of The Center for Language & Speech Processing (CLSP). My colleagues and labmates at the CLSP are fantastic and I am excited to hear about everyone's future success. The colleagues include: Adam Teichert, Adrian Benton, Andrew Blair-Stanek, Annabelle Carrell, Anton Belyy, Brian Leonard, Cash Costello, Chandler May, Courtney Napoles, Craig Harman, Dee Ann Reisinger, Elias Stengel-Eskin, Elliot Schumacher, Elizabeth Salesky, Felicity Wang, Francis Ferraro, Greg Vorsanger, Guanghui Qin, Hongyuan Mei, Huda Khayrallah, Jason Naradowsky, Keisuke Sakaguchi, Keith Levin (who suggested I reach out to Ben), Kenton Murray, Lisa Bauer, Mahsa Yarmohammadi, Nathaniel Weir, Nicholas Andrews, Nils Holzenberger, Noah Weber, Patrick Martin, Patrick Xia, Pushpendre Rastogi, Rachel Rudinger, Rashmi Sankepally, Rebecca Knowles, Ryan Cotterell, Ryan Culkin, Seth Ebner, Sabrina Mielke, Sheng Zhang, Tim Viera, Ting Hua, Tongfei Chen, Travis Wolfe, Winston Wu, and Zhengping Jiang, among others. The CLSP administrative staff, Ruth

Scally, Carl Pupa, Steven Hart, Yamese Diggs, Lauren Bigham, Nicole Aaron, and Jennifer Linton, are all phenomenal and helped clear many administrative hurdles. I enjoyed discussing the newest podcasts, tv shows, and movies from Yamese and Ruth, and hearing about the latest adventures with the grandkids. I will miss the endless amount of candy the CLSP provided.

I was very fortunate to work and learn from collaborators outside of CLSP as well. In particular, I would like to thank Sam Bowman and Ellie Pavlick for leading and including me in the Sentence Representation Learning team at the CLSP's 2018 Summer JSALT Workshop. I would also like to thank Yonatan Belinkov. I'm very fortunate we both attended IJCNLP where a discussion over coffee led to many great collaborations, some of which are included in this thesis.

I am also very appreciative of the Computer Science deparment at Johns Hopkins University. The CS administrative staff, including Zachary Burwell, Kim Franklin, Laura Graham, and Debbie Deford, constantly made sure I was on track. Thank you Vladimir Braverman and Mike Dinitz for encouraging me when I started in CS research. Thank you Joanne Selinski and Sara Miner More for reviving the CS education reading group and sharing your teaching experiences. Thank you to the students in the Spring 2020 EN 601.466 Artificial Intelligence course - teaching and learning from you all was a blast! Finally, thank you Coach Bixby for including me in noon-time hoops. It made the commute from DC always worthwhile.

I would like to thank my parents and siblings for their continual support, encouragement, and honesty. Lastly and most importantly, I would like to thank Gracie for being by my side through out this crazy journey. You know exactly when I need you to be my biggest critic (especially as a copy-editor) and when to be my biggest cheerleader. I couldn't imagine doing this without you.

Contents

Abstract ii

Thesis Committee iv

Acknowledgments v

List of Tables xiv

List of Figures xvi

1 Introduction 1

1.1 Roadmap & Contributions 4

1.1.1 What this thesis is not 6

1.1.2 How to read this thesis: 7

1.2 Publications 8

2 Background 14

2.1 Evaluating Natural Language Processing Systems 15

2.1.1 General Purpose vs Task Specific Evaluations 16

2.1.2 Intrinsic vs Extrinsic Evaluations 22

2.1.3 Contemporary Probes for Linguistic Phenomena 24

2.2 Recognizing Textual Entailment 28

2.2.1 Entailment as an NLP Evaluation 29

2.2.2 Entailment as an Intermediate NLP Task 35

2.2.3 Entailment as a Downstream NLP Task 37

2.3 Revisiting RTE as an NLP Evaluation 38

2.4 Natural Language Inference or Recognizing Textual Entailment? . . . 40

3 The Diverse Natural Language Inference Collection 42

3.1 Overview 43

3.2 Motivation 46

3.3 Recasting existing data into RTE 49

3.3.1 Named Entity Recognition 49

3.3.2 Lexicosyntactic Inference 50

3.3.2.1 VerbNet 51

3.3.2.2 VerbCorner 53

3.3.3 Event Factuality 54

3.3.4 Subjectivity (Sentiment) 55

3.3.5 Figurative Language (Puns) 57

3.4 Experiments & Results 58

3.5 Discussion 68

4 Exploring Semantic Phenomena in neural models via recast-RTE 71

4.1 General Method 73

4.2 Machine Translation 75

4.2.1 Motivation 76

4.2.2 Experiments 79

4.2.3 Results 81

4.2.4 Further Analysis 87

4.2.5 Related Work 92

4.3 Evaluating Encoders Trained For Other Tasks 93

4.3.1 Tasks 93

4.3.2 Experiments 95

4.3.3 Results 96

4.4 Discussion 99

5 Hypothesis-only Biases in Recognizing Textual Entailment 101

5.1 Motivation 104

5.2 Methodology 105

5.3 Datasets 106

5.4 Results 112

5.5 Exploring Statistical Irregularities 116

5.6 Discussion 122

6 Revisiting Paraphrasing in RTE 127

6.1 Motivation 128

6.2 Creating $\hat{p}RTE$ 130

6.3 Models under investigation 133

6.4 Experiments & Results 135

6.5 Semantic Variability 139

6.6 Discussion 141

7 Overcoming Hypothesis-Only Biases 142

7.1 Motivation 143

7.2 Methods 146

7.2.1 General, baseline RTE model 146

7.2.2 Adversarial Learning 148

7.2.2.1 AdvCls: Adversarial Classifier 149

7.2.2.2 AdvDat: Adversarial Training Data 151

7.3 Intrinsic Evaluations 152

7.3.1 Synthetic Experiment 153

7.3.2 Transferring on Traditional RTE Datasets 156

7.4 Analysis 160

7.4.1 Interplay with known biases 160

7.4.2 Stronger hyper-parameters 162

7.4.3 Indicator Words 163

7.5 Discussion 164

8 Adversarial Learning for Emergency Need Discovery 167

8.1 DARPA LORELEI Challenge 168

8.2 Methods 171

8.3 Experiments 172

8.4 Discussion 176

9 Conclusion 177

9.1 Contributions 177

9.2 Future Work 178

Vita 255

List of Tables

2.1 Examples from Fracas 30
2.2 Examples from PASCAL RTE datasets 33
2.3 Examples from SNLI and Multi-NLI 34

3.1 Example RTE pairs for different semantic phenomena in the DNC. . . 45
3.2 Statistics summarizing DNC1.0 59
3.3 Accuracies on the DNC. 60
3.4 Templates used for hypothesis construction in the DNC. 64

4.1 White et al. (2017)'s RTE dataset statistics 81
4.2 Recast RTE accuracies from English NMT encoders 82
4.3 Accuracies on FN+ by sentence-length 83
4.4 FN+ fine-grained accuracies 84
4.5 DPR examples 86
4.6 Accuracies on SPR fine-grained 88
4.7 Accuracies on MNLI 90
4.8 Results on semantic phenomena with NMT encoders 91

5.1 RTE datasets we investigate hypothesis-only biases. 106
5.2 Hypothesis-only accuracies on each dataset. 113
5.3 Accuracies on FN+ for each class label. 121
5.4 Hypothesis-only accuracies on the SPR development data. 123

6.1 Ungrammatical or disfluent recast FN+ examples 129
6.2 Example in Paraphrased RTE 132
6.3 Results on $\hat{p}RTE$ 136
6.4 Qualitative examples of incorrect predictions on $\hat{p}RTE$ 140

7.1 Accuracies on the synthetic dataset 155
7.2 In-domain RTE adversarial results 157
7.3 Accuracies when transferring across RTE datasets 158

7.4 Stronger adversarial hyper-parameters 162
7.5 Indicator words after adversarial learning 165

8.1 Each disaster included in our dataset. 172
8.2 Example tweets sampled from the crowdsourced dataset. 173
8.3 Adversarial Learning for Emergency Need Discovery Results 174
8.4 Examples where the baseline and adversarial models' predictions differ 175

List of Figures

3.1 Results of fine-tuning learning curves for the DNC (1/2) 65
3.2 Results of fine-tuning learning curves for the DNC (2/2) 66

4.1 Evaluation Framework 72
4.2 Anaphora mistake in Google Translate 77
4.3 Pretrained Encoders Evaluated on the DNC 97
4.4 Pretrained Encoders Evaluated on VerbNet in the DNC 98

5.1 Typical NLI and Hypothesis-only model 103
5.2 Number of hypotheses for each label that contain indicator words. . . 116
5.3 Indicator Words Highly Correlated with NLI Labels 118

6.1 Instructions for semantic similarity and grammatically check. 133
6.2 Accuracies on $\hat{p}RTE$ 137
6.4 Accuracies on $\hat{p}RTE$ divided by original task 139

7.1 Illustration of adversarial learning methods 147
7.2 Difference in hypothesis-only biases between SNLI and other datasets 161

8.1 Situation Frame Examples 169
8.2 CADET Annotation Interface for Situation Frames 170

Chapter 1

Introduction

For decades, Artificial Intelligence (AI) researchers aimed to develop machines that humans can seamlessly interact with via language. The proliferation of real-world consumer products like Google Translate and Apple's Siri can be attributed to success in AI and Natural Language Processing (NLP) research. Nevertheless contemporary NLP systems are very brittle. Machine translation systems fail to translate noisy data and dialog systems like Amazon Alexa often fail to understand non-white American accents.[1] Mistakes made by NLP systems can cause global scandals. For example, during Chinese President Xi Jinpeng's recent visit to Burma, Facebook incorrectly translated his name from Burmese to English as "Mr. Shithole, President of China."

As NLP systems become more ubiquitous in our daily lives, it is important to understand where these models may fail, and the limits to their current reasoning

[1] https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/?noredirect=on

capabilities. Developing methods that provide insight into understanding the reasoning capabilities of advanced, contemporary NLP models might help prevent such mistakes.

Inspired by a tradition in linguistics, NLP researchers often rely on the task of Recognizing Textual Entailment (RTE), also known as Natural Language Inference (NLI), to evaluate the reasoning capabilities of NLP systems. Traditionally, RTE is a categorical sentence-pair classification task where a system must determine whether one sentence (*hypothesis*) could likely be inferred from another (*premise*). For example, the hypothesis *Adam received a drink* would likely be inferred by the premise that *Ruth gave Adam a can of La Croix seltzer*.

In recent years, with the introduction of large scale RTE datasets, researchers have become focused on developing models specifically for RTE. Researchers compete to develop more advanced models that predict whether one sentence can likely be inferred from another. In this thesis, I revisit the tradition of using RTE to provide insight into reasoning capabilities of NLP models. RTE is an ideal evaluation framework as coping with textual inferences is necessary for all NLP systems that deal with understanding language (Zaenen, Karttunen, and Crouch, 2005). Furthermore, since "evaluating a system requires the definition of an application task in terms of input/output pairs that are equally applicable to question-answering, text processing, or generation" (Palmer and Finin, 1990) and RTE is equally applicable to these and other downstream tasks, RTE is an ideal framework for evaluating the reasoning capabilities of NLP systems.

In this thesis, I demonstrate why existing resources and prior efforts in RTE fail to adequately accomplish the grand vision of being a unified framework to evaluate the reasoning capabilities of NLP systems. Most RTE datasets provide just a single metric in terms of how well a system accurately predicts whether one sentence likely follows another. Unfortunately, this metric does not provide any insight into the range of reasoning capabilities of our systems. Therefore, I argue for using RTE as a unified framework to test for a wide range of reasoning capabilities.

This thesis introduces methods for creating RTE datasets that provide insight into the range of reasoning capabilities of our NLP systems. The methods I introduce primarily rely on recasting annotations for different semantic phenomena into RTE. These semantic phenomena include humor, figurative speech, named entity recognition, and event factuality. As part of this thesis, I release a large scale dataset that tests for over fifteen types of reasoning phenomena. I introduce a method for using this data to evaluate the reasoning capabilities of NLP systems, and I evaluate NLP systems trained to translate, connect images with text, and parse sentences into syntactic chunks. Additionally, I discover biases and issues in popular RTE datasets that hinder their ability to test NLP models' reasoning capabilities. I use these discovered biases to refine the utility of and proper use cases for RTE. Sparck Jones (1994) argued against the idea of a "single correct way to evaluate an NLP system," and these discovered biases demonstrate the limits of RTE.

While RTE is primarily a tool to evaluate NLP systems, researchers are still interested in developing methods and systems that can determine whether one sentence likely entails another. Therefore, I introduce methods that allow models to overcome dataset specific biases and I apply these methods to successful identify and discover emergency needs during disastrous events.

1.1 Roadmap & Contributions

In **Chapter 2**, I begin this thesis by reviewing prior work in evaluating NLP systems. I will discuss how Recognizing Textual Entailment was introduced as a framework to evaluate NLP systems and I will highlight why most prior work in RTE cannot be used to adequately evaluate NLP systems.

In **Chapter 3**, I introduce the Diverse Natural Language Inference Collection (DNC), a collection of diverse semantic phenomena recast into RTE. The DNC includes phenomena that are necessary components of more general sentence-level semantic inference. The primary contribution of this chapter is the development and incremental release of a large scale collection of datasets that can provide better insights into models' reasoning capabilities.

Using the DNC, in **Chapter 4**, I investigate the ability of NLP models trained on different tasks to capture specific and focused types of sentence-level semantic inference. Some of the contributions of this chapter include a general purpose framework for using RTE to analysis models' reasoning capabilities, and the discovery that the choice of target language in a neural machine translation system can change how well the system captures a diverse range of semantic phenomena. I apply this framework to different models trained on a large array of NLP tasks.

In **Chapter 5**, with a critical eye, I introduce a simple approach that discovers evidence of biases in prior RTE datasets that may limit their usefulness in terms of understanding reasoning capabilities in NLP models. The biases discovered also call into question which types of phenomena are appropriate to convert to RTE. Additionally, this work sparked a renewed interested in the community to develop RTE models that overcome dataset specific biases and perform well across multiple NLI dataset.

Next, in **Chapter 6**, I present a new RTE dataset focused on paraphrases. Biases discovered in the previous chapter demonstrate blunders in prior efforts to recast an RTE dataset focused on paraphrases. This chapter presents a new approach and test set focused specifically on paraphrases.

In **Chapter 7**, I then discuss potential solutions to mitigate these biases when developing models to perform RTE and demonstrate how these solutions might enable models to ignore these biases.

Turning towards an applied setting, in **Chapter 8**, I demonstrate how these bias mitigation techniques can be applied to identifying and discovering emergency needs during disastrous events. This method resulted in top performance in the DARPA

Low Resource Languages for Emergent Incidents (LORELEI) challenge in 2019.

Finally, in **Chapter 9**, I summarize the contributions of this thesis and discuss open research problems and future research directions.

Software Contributions

Most of the work for this thesis has been released across multiple open-sourced software repositories. These include:

- `https://github.com/azpoliak/hypothesis-only-NLI`
- `https://github.com/azpoliak/robust-nli`
- `https://github.com/azpoliak/nmt-repr-analysis`
- `https://github.com/decompositional-semantics-initiative/DNC`

1.1.1 What this thesis is not

This thesis does not introduce state-of-the-art models for NLP tasks, including RTE. Instead, this thesis focuses on revisiting RTE as a method for evaluating how well NLP systems capture different semantic phenomena related to understanding human language. This thesis is not a treatise on what counts as *understanding* language, or more broadly defining the full scope of an intelligent agent or artificial general intelligence (AGI). I leave such discussions to philosophers and sci-fi writers. While the semantic phenomena covered in this thesis are important for general natural language understanding, I do not believe they are the be-all-and-end-all of understanding human language. Many phenomena important for natural language understanding, e.g. pragmatic inference, are not included in this thesis.

This thesis is not a comprehensive nor complete evaluation of the phenomena or types of reasoning captured in the models tested. Instead, the thesis argues for using RTE as a method that may shed light into reasoning capabilities of NLP models. This thesis introduces methods to efficiently develop RTE datasets that each probe for distinct types of reasoning. These datasets and methods serve as tools to discover shortcomings of NLP systems that can hopefully prevent diplomatic blunders as the one discussed earlier in the introduction.

1.1.2 How to read this thesis:

Inspired by Xuchen Yao's thesis (Yao, 2014), I provided a guide for how to read this thesis under different constraints/settings:

If you only have 20 minutes: Read Chapter 1 which provides an overview of the important ideas in this thesis.

If you have 40 minutes: Read Chapter 1, the first and discussion sections of each chapter, starting with Chapter 3.

If you are new to Recognizing Textual Entailment: Section 2.2 provides a good reference for fundamental datasets and work in RTE.

If you are interested in probing for semantic phenomena: Start with Section 2.1.3 for a brief background, then read Chapter 3 and Chapter 4.

If you have read all my papers and are looking for something new: This list contains previously unpublished material in this thesis:

- Experiments in the DNC about learning curves in Table 3.3.
- Section 4.3 describes experiments using the DNC to evaluate encoders pretrained on different tasks. This work was discussed at the 2018 JSALT closing presentations.
- Chapter 6 introduces a new RTE dataset where examples have been paraphrased. The chapter includes experiments demonstrate whether different types of models are robust to paraphrases.

1.2 Publications

This thesis is based on many peer-reviewed publications that I have co-authored.[2] These publications include:

- **On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference**. Adam Poliak, Yonatan Belinkov, James Glass, Benjamin Van Durme. *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. Association for

[2] Research is a team sport and I have been fortunate to learn from and publish work with great collaborators. In subsequent chapters, this thesis will use the first person plural instead of the singular. This is inspired by tradition (Napoles, 2018; Knowles, 2019; Rudinger, 2019) and a desire to recognize the contributions of my collaborators in this work.

Computational Linguistics. New Orleans, Louisiana, USA. June 2018, pages 513-523. `http://aclweb.org/anthology/N18-2082`

- **Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation** Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, Benjamin Van Durme. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. Brussels, Belgium. November 2018. `http://aclweb.org/anthology/D18-1007`
- **Hypothesis Only Baselines in Natural Language Inference** Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger and Benjamin Van Durme. *The Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*. Association for Computational Linguistics. New Orleans, Louisiana, USA. June 2018, pages 180-191. *Best Paper Award* `http://www.aclweb.org/anthology/S18-2023`
- **Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference** Yonatan Belinkov*, Adam Poliak*, Stuart M. Shieber, Benjamin Van Durme, Alexandar Rush. *57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics. Florence, Italy. July 2019
- **On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference** Yonatan Belinkov*, Adam Poliak*, Stuart M. Shieber, Ben-

jamin Van Durme, Alexandar Rush. *The Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*. Association for Computational Linguistics. Minneapolis, Minnesota, USA. June 2019, pages 256–262. https://www.aclweb.org/anthology/S19-1028

As a graduate student, I was fortunate to co-author additional publications that are not included in this thesis. These publications can be grouped into the following topics:

- Word Embeddings
  - **Frame-Based Continuous Lexical Semantics through Exponential Family Tensor Factorization and Semantic Proto-Roles**. Frank Ferraro, Adam Poliak, Ryan Cotterell, Benjamin Van Durme. In *Proceedings of the Sixth Joint Conference on Lexical and Computational Semantics (⋆SEM)*. Association for Computational Linguistics, Vancouver, Canada, Augsut 2017, pages 97–103. http://www.aclweb.org/anthology/S17-1011
  - **Efficient, Compositional, Order-Sensitive n-gram Embeddings**. Adam Poliak*, Pushpendre Rastogi*, M. Patrick Martin, Benjamin Van Durme. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics, Valencia, Spain, April 2017, pages 503–508. https://aclweb.org/anthology/E/E17/E17-2081.pdf.
  - **Explaining and Generalizing Skip-Gram through Exponential Fam-**

ily Principal Component Analysis. Ryan Cotterell, Adam Poliak, Benjamin Van Durme, Jason Eisner. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics, Valencia, Spain, April 2017, pages 175–181. `http://www.aclweb.org/anthology/E17-2028`.

- Analysis/Probing
  - **Probing what different NLP tasks teach machines about function word comprehension** Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, R. Thomas Mccoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel Bowman and Ellie Pavlick. *The Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*. Association for Computational Linguistics. Minneapolis, Minnesota, USA. June 2019, pages 235–249. *Best Paper Award*
    `https://www.aclweb.org/anthology/S19-1026`
  - **What do you learn from context? Probing for sentence structure in contextualized word representations** Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, Ellie Pavlick. *Seventh International Conference on Learning Representations (ICLR)*. 2019.
    `https://openreview.net/forum?id=SJzSgnRcKX`
- Semantics

  - **Semantic Proto-Role Labeling**. Adam Teichert, Adam Poliak, Benjamin Van Durme, Matt Gormley. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence.* Association for the Advancement of Artificial Intelligence (AAAI), San Francisco, California, February 2017 `https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14997`
  - **Generating Automatic Pseudo-entailments from AMR Parses**. Adam Poliak and Benjamin Van Durme. *6th Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL)* Washington D.C, USA, May 2017. Extended Abstract `http://www.cs.jhu.edu/~apoliak1/papers/Poliak-et-al-MASC-SLL_2017.pdf`

- Other
  - **Neural Variational Entity Set Expansion for Automatically Populated Knowledge Graphs** Pushpendre Rastogi, Adam Poliak, Vince Lyzinski, and Benjamin Van Durme. *Information Retrieval Journal* September 2018. `https://rdcu.be/98BY`
  - **CADET: Computer Assisted Discovery Extraction and Translation**. Benjamin Van Durme, Tom Lippincott, Kevin Duh, Deana Burchfield, Adam Poliak, Cash Costello, Tim Finin, Scott Miller, James Mayfield, Philipp Koehn, Craig Harman, Dawn Lawrie, Chandler May, Annabelle Carrell, Julianne Chaloux, Tongfei Chen, Alex Comerford, Mark Dredze, Benjamin Glass, Shudong Hao, Patrick Martin, Pushpendre Rastogi Rashmi

Sankepally, Travis Wolfe, Ying-Ying Tran, Ted Zhang. *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP), System Demonstrations.* The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei, Taiwan, November 2017, pages 5–8
`http://www.aclweb.org/anthology/I17-3002`

– **Training Relation Embeddings under Logical Constraints**. Pushpedre Rastogi, Adam Poliak, Benjamin Van Durme. In *The First Workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis (KG4IR)*. The 40th International ACM SIGIR Conference

Chapter 2

Background

> Research is not so much going round in circles as ascending a spiral, if only a rather flat one... NLP has returned to some of its early themes, and by a path on an ascending spiral rather than in a closed circle, even if the ascent is slow and uneven.
>
> (Sparck Jones, 1994)

How do we evaluate machines developed for humans to seamlessly interact with via language? How do we determine that one NLP system understands language or generates text better than another? As NLP-based technologies are more widely adopted, these questions are more relevant now than ever.

We begin this chapter by discussing different approaches to NLP evaluations over the past thirty years. We will explore different ways the community has evaluated and compared systems developed for understanding and generating language. Next, we will discuss how Recognizing Textual Entailment (RTE) was introduced as a specific answer to this broad question of how to best evaluate NLP systems. This will include a broad discussion of efforts in the past three decades to build RTE datasets and use RTE to evaluate NLP models. We will conclude with a discussion on why many existing RTE datasets fall short of providing a method for evaluating how well systems understand language.

2.1 Evaluating NLP Systems

The question of how best to evaluate NLP systems is an open problem that has intrigued the community for decades. Martha Palmer and Tim Finin's 1988 workshop on the evaluation of NLP systems explored key questions for evaluation. These included questions related to valid measures of "black-box" performance, linguistic theories that are relevant to developing test suites, reasonable expectations for robustness, and measuring progress in the field (Palmer and Finin, 1990). The large number of ACL workshops focused on evaluations in NLP demonstrate the lack of consensus on how to properly evaluate NLP systems, despite the constant interest in evaluation methods. These workshops include those focused on:

1. Evaluations in general (Pastra, 2003), including this year's Evaluation and Comparison of NLP Systems (Eval4NLP)[1];

2. Different NLP tasks, e.g. machine translation (*Workshop on MT Evaluation: Hands-On Evaluation* 2001; Goldstein et al., 2005) and summarization (Conroy

[1] https://nlpevaluation2020.github.io/index.html

et al., 2012; Giannakopoulos et al., 2017);

3. Contemporary NLP approaches that rely on vector space representations (Levy et al., 2016; Bowman et al., 2017; Rogers et al., 2019).

Evaluation Dichotomies

In the quest to develop an ideal evaluation framework for NLP systems, researchers proposed multiple evaluation methods. These approaches included EAGLES (King et al., 1995), TSNLP (Oepen and Netter, 1995; Lehmann et al., 1996), *FraCas* (Cooper et al., 1996), CLEF (Agosti et al., 2007), SENSEVAL (Kilgarriff, 1998), SEMEVAL (Agirre, Màrquez, and Wicentowski, 2007), and others. These approaches are often categorized into multiple dichotomies. Here, we will survey approaches along two dichotomies. The first is the distinction between general purpose compared to task specific evaluations and the second we will discuss is intrinsic versus extrinsic evaluations. Resnik and Lin (2010) summarize other evaluation dichotomies and Paroubek, Chaudiron, and Hirschman (2007) present a history and evolution of NLP evaluation methods.

2.1.1 General Purpose vs Task Specific Evaluations

General purpose evaluations determine how well NLP systems capture different linguistic phenomena. These evaluations often rely on the development of test cases that systematically cover a wide range of phenomena. Additionally, these evaluations generally do not consider how well a system under investigation performs on held out data for the task that the NLP system was trained on. In general purpose evaluations, specific linguistic phenomena should be isolated such that each test or example evaluates one specific linguistic phenomenon, as tests ideally "are controlled and exhaustive databases of linguistic utterances classified by linguistic features" (Lloberes, Castellón, and Padró, 2015).

In task specific evaluations, the goal is to determine how well a model performs on a held out test corpus. How well systems generalize on text classification problems is determined with a combination of metrics like accuracy, precision, and recall. For generation tasks like machine translation and summarization, NLP systems are often compared based on metrics like *BiLingual Evaluation Understudy* (BLEU) (Papineni et al., 2002) and *Recall Oriented Understudy for Gisting Evaluation* (Rouge) (Lin, 2004). Task specific evaluations, where "the majority of benchmark datasets … are drawn from text corpora, reflecting a natural frequency distribution of language phenomena" (Belinkov and Glass, 2019), is the common paradigm in NLP research today. Researchers often begin their research with provided training and held-out test corpora, as their research agenda is to develop systems that outperform other researchers' systems on a held-out test set based on a wide range of metrics. The majority of the work presented in this thesis deviates from this popular trend in NLP research. We are not focused on developing NLP systems that generalize better than other researchers' systems. Rather, we present a test suite covering a wide range of linguistic phenomena, and we introduce a general purpose method to evaluate the reasoning capabilities of NLP systems using the introduced test suite.

The dichotomy between general purpose and task specific evaluations is sometimes blurred. For example, while general purpose evaluations are ideally task agnostic, researchers develop evaluations that test for a wide range of linguistic phenomena captured by NLP systems trained to perform specific tasks. These include linguistic tests targeted for systems that focus on parsing (Lloberes, Castellón, and Padró, 2015), machine translation (King and Falkedal, 1990; Koh et al., 2001; Isabelle, Cherry, and Foster, 2017; Choshen and Abend, 2019; Popović and Castilho, 2019; Avramidis et al., 2019), summarization (Pitler, Louis, and Nenkova, 2010), and others (Chinchor, 1991; Chinchor, Hirschman, and Lewis, 1993).

Test Suites vs. Test Corpora

In turn, we can better classify the dichotomy between general purpose and task specific evaluations in terms of the data used to evaluate systems. Oepen and Netter (1995) refer to this distinction as test suites versus test corpora. They define a test suite as a "systematic collection of linguistic expressions (test items, e.g. sentences or phrases) and often includes associated annotations or descriptions." They lament the state of test suites in their time since

> most of the existing test suites have been written for specific systems or simply enumerate a set of 'interesting' examples; this clearly does not

> meet the demand for large, systematic, well-documented and annotated collections of linguistic material required by a growing number of NLP applications.
>
> (Oepen and Netter, 1995)

Oepen and Netter further delineate the difference between test corpora and test suites. Unlike "test corpora drawn from naturally occurring texts," test suites allow for 1) more control over the data, 2) systematic coverage, 3) non-redundant representation, 4) inclusion of negative data, and 5) coherent annotation. Thus, test suites "allow for a fine-grained diagnosis of system performance" (Oepen and Netter, 1995). Oepen and Netter argue that both should be used in tandem - "test suites and corpora should stand in a complementary relation, with the former building on the latter wherever possible and necessary." Hence, both test suites and test corpora are important for evaluating how well NLP systems capture linguistic phenomena and perform in practice on real world data.

Categorizing Approaches as General Purpose or Task Specific Evaluations

> Any actual test suite to be used for some given test or evaluation will have to be more or less specific in order to yield optimally informative and interpretable results. Therefore, the notion of a monolithic and fixed general-purpose test suite seems neither feasible nor desirable. On the other hand, there will obviously be a rather large amount of linguistic phenomena which any test suite might want to include.
>
> (Balkan et al., 1994)

When introducing Test Suite for NLP (TSNLP) (Oepen and Netter, 1995; Lehmann et al., 1996), a multi-year project funded by Linguistic Research Engineering program of the European Commission, Balkan et al. (1994) distinguish general-purpose diagnostics from task- and domain-specific evaluations. Balkan et al. idealize applying a generic test suite to different NLP systems in order to explore how well the systems capture linguistic phenomena, but acknowledge that this might not be feasible.

We divide different lines of research in evaluation methods based upon the dichotomy of General Purpose (Test Corpora) or Task Specific (Test Suite) Evaluations. At the conclusion of the 1988 Workshop of Evaluating NLP systems, Palmer and Finin (1990) argued for using an evaluation framework that is task and domain agnostic. Sparck Jones and Galliers (1996)'s textbook devoted to analyzing different techniques to evaluate NLP systems disagreed with this idea. Sparck Jones and Galliers argued for domain and task specific evaluations, as they claimed that it is infeasible, impractical, and not meaningful to evaluate an NLP system outside of an applied task. In her review of the textbook, Sharon Walter disagreed with the idea that a evaluation criteria for generic NLP systems cannot be adequately defined. Walter writes that

> The evaluation methodology does not, however, appear to strike at the heart of the evaluation problem of defining specific criteria by which to describe and compare system capabilities, evading the issue in fact by proposing that general criteria cannot be defined due to the necessity of case-by-case specification of evaluation criteria.
>
> (Walter, 1998)

Walter's idea of a generic evaluation methodology was evident by the Neal-Montgomery NLP System Evaluation Methodology, a methodology that "produces descriptive, objective profiles of system linguistic capabilities without a requirement for system adaptation to a new domain" (Walter, 1992).

The majority of contemporary NLP research relies on task-specific (test corpora) based evaluations. As pointed out in a recent survey of analysis methods in NLP, currently "the majority of benchmark datasets in NLP are drawn from text corpora, reflecting a natural frequency distribution of language phenomena" (Belinkov and Glass, 2019).

Gold Labeled Data

For both test corpora and test suite evaluations, assumptions are made that test data (and at least some of the training data) conform to a "gold standard," i.e. a "data set of natural language texts annotated by humans for correct solutions of that particular task" (Kováź, Jakubíźek, and Horák, 2016). Read et al. (1988) earlier referred to such annotated data points as "exemplars of representative problems in natural language understanding." An exemplar "includes a piece of text (sentence dialog fragment, etc.), a description of the conceptual issue represented, a detailed discussion of the problem in understanding the text and a reference to a more extensive discussion in the literature." Read et al. (1988) refer to a collection of exemplars plus "a conceptual taxonomy of the types of issues represented in the" exemplars as a Sourcebook.

Relying on standard collections of gold data is common in most shared tasks or (D)ARPA sponsored programs. It has been noted that by "bringing research communities together" (Kilgarriff and Palmer, 2000) and developing common resources (Gaizauskas, 1998), ARPA funded programs are responsible for the culture of rigorous evaluation in NLP and AI research (King, 1996). Sparck Jones (1994) similarly noted that

> the (D)ARPA speech recognition and message understanding conferences were important not only for the tasks they addressed but for their emphasis on rigorous evaluation, initiating a trend that became a major feature of the 1990s.

2.1.2 Intrinsic vs Extrinsic Evaluations

> Intrinsic evaluations test the system in of itself and extrinsic evaluation test the system in relation to some other task.
>
> (Farzindar and Lapalme, 2004)

The second dichotomy we explore is intrinsic versus extrinsic evaluations. When reviewing Sparck Jones and Galliers's textbook, Estival (1997) comment that "one of the most important distinctions that must be drawn when performing an evaluation of a system is that between *intrinsic criteria*, i.e. those concerned with the system's own objectives, and *extrinsic criteria*, i.e. those concerned with the function of the system in relation to its set-up." Resnik et al. (2006) similarly noted that "intrinsic evaluations measure the performance of an NLP component on its defined subtask, usually against a defined standard in a reproducible laboratory setting" while "extrinsic evaluations focus on the component's contribution to the performance of a complete application, which often involves the participation of a human in the loop."

Sparck Jones (1994) refers to the distinction of intrinsic vs extrinsic evaluations as the *orientation* of an evaluation.

Under these definitions, “an intrinsic evaluation of a parser would analyze the accuracy of the results returned by the parser as a stand-alone system, whereas an extrinsic evaluation would analyze the impact of the parser within the context of a broader NLP application” like answer extraction (Mollá and Hutchinson, 2003). When evaluating a document summarization system, an intrinsic evaluation might ask questions related to the fluency or coverage of key ideas in the summary while an extrinsic evaluation might explore whether a generated summary was useful in a search engine (Resnik and Lin, 2010). This distinction has also been referred to as application-free vs. application-driven evaluations (Kováź, Jakubíźek, and Horák, 2016).

In the case of evaluating different methods for training word vectors, intrinsic evaluations might consider how well similarities between word vectors correlate with human evaluated word similarities.[2] This is the basis of evaluation benchmarks like SimLex (Hill, Reichart, and Korhonen, 2015), Verb (Baker, Reichart, and Korhonen, 2014), RW (Luong, Socher, and Manning, 2013), MEN (Bruni et al., 2012), WordSim-353 (Finkelstein et al., 2001), and others. Extrinsic evaluations might consider how well different word vectors help models for tasks like sentiment analysis (Petrolito, 2018; Mishev et al., 2019), machine translation (Wang et al., 2019c), or named entity

[2] Word vectors, also known as word embeddings, are low dimensional vector representations (often $50-300$ dimensions in length) for words that are supposed to capture the meaning of a word. For an overview on word embeddings, see Sections 10.4 and 10.5 in Goldberg (2017)’s excellent textbook.

recognition (Wu et al., 2015; Nayak, Angeli, and Manning, 2016).

Proper extrinsic evaluations are often infeasible in an academic lab setting. Therefore, researchers often rely on intrinsic evaluations to approximate extrinsic evaluations, even though intrinsic and extrinsic evaluations serve different goals and many common intrinsic evaluations for word vectors (Tsvetkov et al., 2015; Chiu, Korhonen, and Pyysalo, 2016; Faruqui et al., 2016), generating natural language text (Belz and Gatt, 2008; Reiter, 2018), or text mining (Caporaso et al., 2008) might not correlate with extrinsic evaluations.[3] Developing intrinsic evaluations that correlate with extrinsic evaluations remains an open problem in NLP.

2.1.3 Contemporary Probes for Linguistic Phenomena

Similar to test suites discussed earlier, recent lines of research focus on probing how well NLP systems capture a wide range of linguistic phenomena. While some of this work has recently been referred to as intrinsic (Eichler, Şahin, and Gurevych, 2019), this does not follow the common definition of an intrinsic evaluation discussed above. Rather, we view probing for linguistic phenomena as an example of test suite based evaluation.

[3] Although recent work suggest that some intrinsic evaluations for word vectors do indeed correlate with extrinsic evaluations (Qiu et al., 2018; Thawani, Srivastava, and Singh, 2019).

Meaning Representation Formalisms as an Evaluation Metric

Some have argued for using meaning representations as an evaluation metric to probe how well NLP systems capture different semantic phenomena. In computational semantics, meaning representations are formalisms that map the meaning of natural language text to structured representations. Some examples of such meaning representations include Episodic Logic (Schubert and Hwang, 2000), Abstract Meaning Representation (AMR) (Banarescu et al., 2013), Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013), and Universal Decompositional Semantics (White et al., 2016). One motivation for developing meaning representation formalisms is to be able to evaluate NLP systems. Indeed, the MEANT (Lo and Wu, 2011a) metric, and its extensions XMEANT (Lo et al., 2014) and MEANT2.0 (Lo, 2017), use a semantic formalism to evaluate machine translation systems. The metric is used by automatically parsing a system's output and a reference translation into graphs of PropBank semantic roles (Palmer, Gildea, and Kingsbury, 2005) and then comparing the resulting graphs. In HMEANT (Lo and Wu, 2011b), human annotators are used to parse the texts into semantic roles. The HUME metric (Birch et al., 2016) works similarly as HMEANT but relies on parsing the texts into UCCA's meaning formalism instead.

Relying on semantic formalisms is an intuitive method to evaluate how well NLP systems capture semantics, i.e. linguistic phenomena related to understanding language. However, using semantic formalisms in this way is currently not a feasible or scalable approach to attain insightful evaluation metrics. As noted in the summary of the 1988 workshop on evaluating NLP systems,

> A fundamental underlying snag is the difficulty in arriving at a consensus on the nature of semantic representation. If the community was in agreement on what the representation of a sentence is supposed to be – whether it was a sentence from a dialog with an expert system, a sentence fragment from a tactical message, or a database query – then the task of assessing a system's performance would be much more straightforward. Given input X, does the system produce Y as an internal data structure? Unfortunately, there are now as many Y's for X as there are systems, so finding a reliable method of assessing a system in isolation, or of comparing two systems, becomes much more difficult.
>
> (Palmer and Finin, 1990)

Even if there is an agreement on the "characterization of phenomena, mappings from one style of semantic representation to another, [and] on content of representations for a common domain" (Palmer and Finin, 1990), automatically parsing text into these different formalism is still an unsolved problem, and relying on humans to manually parse each sentence, like in HMEAT or HUME, is not scalable.

Auxiliary Diagnostic classifiers

Recent popular approaches for evaluating how well NLP systems capture semantic, and other linguistic, phenomenon leverage auxiliary or diagnostic classifiers, which are often agnostic to specific meaning representation formulations. With the rise of deep learning in NLP, contemporary NLP systems often leverage pre-trained encoders to represent the meaning of a sentence in a fixed-length vector representation. Adi et al. (2017) introduced the notion of using auxiliary classifiers as a general purpose methodology to diagnose what language information is encoded and captured by contemporary sentence representations. They argued for using “auxiliary prediction tasks” where, like in Dai and Le (2015), pre-trained sentence encodings are “used as input for other prediction tasks.” The “auxiliary prediction tasks” can serve as diagnostics, and Adi et al. (2017)’s auxiliary, diagnostic tasks focused on how word order, word content, and sentence length are captured in pre-trained sentence representations.

As Adi et al.’s methodology is general “and can be applied to any sentence representation model,” researchers develop other diagnostic tasks that explore different linguistic phenomenon (Ettinger et al., 2018; Conneau et al., 2018; Hupkes, Veldhoen, and Zuidema, 2018). Belinkov (2018)’s thesis relied on and popularized this methodology when exploring how well speech recognition and machine translation systems capture phenomena related to phonetics (Belinkov and Glass, 2017), morphology (Belinkov et al., 2017b), and syntax (Belinkov et al., 2017a).

The general purpose methodology of auxiliary diagnostic classifiers is also used to explore how well different pre-trained sentence representation methods perform on a broad range of NLP tasks. For example, SentEval (Conneau and Kiela, 2018) and GLUE (Wang et al., 2018; Wang et al., 2019a) are used to evaluate how different sentence representations perform on paraphrase detection, semantic textual similarity, and a wide range of binary and multi-class classification problems. We categorize these methods of probing for linguistic phenomena as extrinsic evaluations since they often treat learned sentence-representations as features to train a classifier for an external task. However, most of these could be categorized as test corpora, rather than test suites as defined in Section 2.1.1, since the data is not tightly controlled to evaluate specific linguistic phenomena. Rather, they package existing test corpora for different tasks and provide an easy platform for researchers to compete on developing systems that perform well on the suite of pre-existing, and re-packaged test corpora.

This thesis leverages the general methodology introduced by Adi et al. (2017). However, we advocate for using a single framework, Recognizing Textual Entailment, to evaluate different linguistic phenomena. As we will discuss later, this allows us to use one consistent format and framework for testing how well contemporary, deep learning NLP systems capture a wide-range of linguistic phenomena.

2.2 Recognizing Textual Entailment

Recognizing Textual Entailment (RTE) emerged as a framework to evaluate how well NLP systems can perform semantic inferences that are necessary for multiple downstream NLP tasks. Rooted in linguistics, RTE is the task of determining whether the meaning of one sentence can likely be inferred from another. Unlike the strict definition of entailment in linguistics that "sentence A entails sentence B if in all models in which the interpretation of A is true, also the interpretation of B is true" (Janssen, 2011), RTE relies on a fuzzier notion of entailment. The original annotation guidelines for RTE stated that

> in principle, the hypothesis must be fully entailed by the text. Judgment would be False if the hypothesis includes parts that cannot be inferred from the text. However, cases in which inference is very probable (but not completely certain) are still judged as True.
>
> (Dagan, Glickman, and Magnini, 2006)

We will begin by discussing how RTE was introduced as an evaluation framework focused on the semantic inference capabilities of NLP systems. This will include a survey on prior RTE datasets. Next, we will discuss how researchers use RTE datasets as an intermediate step to improve NLP systems for downstream tasks. We then will discuss how researchers have move away from RTE's diagnostic and evaluation goals since researchers often compete to develop the best performing RTE models. We will conclude by critiquing prior RTE datasets, discussing why they cannot be used as sufficient test suites to explore the reasoning capabilities of NLP systems.

2.2.1 Entailment as an NLP Evaluation

> NLP systems cannot be held responsible for knowledge of what goes on in the world but no NLP system can claim to "understand" language if it can't cope with textual inferences.
>
> (Zaenen, Karttunen, and Crouch, 2005)

Recognizing and coping with inferences is key to understanding human language. While NLP systems might be trained to perform different tasks, such as translating, answering questions, or extracting information from text, most NLP systems require understanding and making inferences from text. Therefore, RTE was introduced as

	QUANTIFIERS (14)
P	Neither leading tenor comes cheap. One of the leading tenors is Pavarotti.
Q	Is Pavarotti a leading tenor who comes cheap?
H	Pavarotti is a leading tenor who comes cheap.
A	No
	PLURALS (94)
P	The inhabitants of Cambridge voted for a Labour MP.
Q	Did every inhabitant of Cambridge vote for a Labour MP?
H	Every inhabitant of Cambridge voted for a Labour MP.
A	Unknown
	COMPARATIVES (243)
P	ITEL sold 3000 more computers than APCOM. APCOM sold exactly 2500 computers.
Q	Did ITEL sell 5500 computers?
H	ITEL sold 5500 computers.
A	Yes

Table 2.1: Examples from Fracas: **P** represents the premise(s), **Q** represents the question from *FraCas*, **H** represents the declarative statement MacCartney (2009) created and, **A** represents the label. The number in the parenthesis indicates the example ID from *FraCas*.

a framework to evaluate NLP systems. Starting with *FraCas*, we will discuss early work that introduced and argued for RTE as an evaluation framework.

FraCas

Over a span of two years (December 1993 - January 1996), Cooper et al. (1996) developed *FraCas* as "an inference test suite for evaluating the inferential competence of different NLP systems and semantic theories". Created manually by many linguists and funded by FP3-LRE,[4] *FraCas* is a "semantic test suite" that covers a range of semantic phenomena categorized into 9 classes: generalized quantifiers, plurals, anaphora, ellipsis, adjectives, comparatives, temporal reference, verbs, and attitudes.

[4] https://cordis.europa.eu/programme/id/FP3-LRE

Examples in *FraCas* contain a premise paired with a hypothesis. Premises are at least one sentence, though sometimes they contain multiple sentences, and most hypotheses are written in the form of a question and the answers are either *Yes*, *No*, or *Don't know*. MacCartney (2009) (specifically Chapter 7.8.1) converted the hypotheses from questions into declarative statements.[5] Table 2.1 contains examples from *FraCas*.

In total, *FraCas* only contains about 350 labeled examples, potentially limiting the ability to generalize how well models capture these phenomena. Additionally, this limited number of examples in *FraCas* prevents its use as a dataset to train data hungry deep learning models.

Pascal Recognizing Textual Entailment Challenges

> A fundamental phenomenon of natural language is the variability of semantic expression, where the same meaning can be expressed by, or inferred from, different texts. This phenomenon may be considered the dual problem of language ambiguity, together forming the many-to-many mapping between language expressions and meanings. Many natural language processing applications, such as Question Answering, Information Extraction, summarization, and machine translation evaluation, need a model for this variability phenomenon in order to recognize that a particular target meaning can be inferred from different text variants ...
>
> It seems that major inferences, as needed by multiple applications, can indeed be cast in terms of textual entailment. For example, a QA system has to identify texts that entail a hypothesized answer. Given the question "What does Peugeot manufacture?", the text "Chrétien visited Peugeot's newly renovated car factor" entails the hypothesized answer form "Peugeot manufactures cars". Similarly, for certain Information Retrieval queries the combination of semantic concepts and relations denoted by the query should be entailed from relevant retrieved documents.

[5] urlhttps://nlp.stanford.edu/ wcmac/downloads/fracas.xml

(Dagan, Glickman, and Magnini, 2006)

With a similar broad goal as *FraCas*, the Pascal Recognizing Textual Entailment challenges (RTE) began as a "generic evaluation framework" to compare the inference capabilities of models designed to perform different tasks. Unlike *FraCas*'s goal of determining whether a model performs distinct types of reasoning, the Pascal RTE Challenges was primarily focused on using this framework to evaluate models for distinct, real-world downstream tasks. Thus, the examples in the Pascal RTE datasets were extracted from downstream tasks. The process was referred to as *recasting* in the thesis by Glickman (2006).

NLU problems were reframed under the RTE framework and candidate sentence pairs were extracted from existing NLP datasets and then labeled under variations of the RTE definition described above (Dagan, Glickman, and Magnini, 2006). For example, the RTE1 data came from 7 tasks: comparable documents, reading comprehension, question answering, information extraction, machine translation, information retrieval, and paraphrase acquisition.[6] Starting with Dagan, Glickman, and Magnini (2006), there have been eight iterations of the PASCAL RTE challenge, with the most recent being Dzikovska et al. (2013). Technically, Bentivogli et al. (2011) was the last challenge under PASCAL's aegis, but Dzikovska et al. (2013) was branded as the 8th RTE challenge. Table 2.2 contains examples from RTE1-3.

Researchers analyzed the RTE challenge datasets. Marneffe, Rafferty, and Man-

[6] Chapter 3.2 of Glickman's thesis discusses how examples from these datasets were converted into RTE.

Kessler 's team conducted 60,643 interviews with adults in 14 countries ▶ Kessler 's team interviewed more than 60,000 adults in 14 countries	entailed
Capital punishment is a catalyst for more crime ▶ Capital punishment is a deterrent to crime	not-entailed
Boris Becker is a former professional tennis player for Germany ▶ Boris Becker is a Wimbledon champion	not-entailed

Table 2.2: Examples from the PASCAL RTE datasets (modified for space): The first line in each example is the premise and the line starting with ▶ is the corresponding hypothesis. The first, second, and third examples are from the RTE1, RTE2, and RTE3 development sets respectively. The second column indicates the label for the example.

ning (2008) argued that there exist different levels and types of contradictions. They focus on different types of phenomena, e.g. antonyms, negation, and world knowledge, that can explain why a premise contradicts a hypothesis. MacCartney (2009) used a simple bag-of-words model to evaluate early iterations of Recognizing Textual Entailment (RTE) challenge sets and noted[7] that "the RTE1 test suite is the hardest, while the RTE2 test suite is roughly 4% easier, and the RTE3 test suite is roughly 9% easier." Additionally, Vanderwende and Dolan (2006) and Blake (2007) demonstrate how sentence structure alone can provide a high signal for some RTE datasets.[8]

SNLI and MNLI

The most popular recent RTE datasets, Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and its successor Multi-NLI (Williams, Nangia, and Bowman, 2017), follow the line of RTE work developed at Stanford (MacCartney,

[7] In Chapter 2.2 of his thesis

[8] Vanderwende and Dolan (2006) explored RTE-1 and Blake (2007) analyzed RTE-2 and RTE-3.

P	A woman is talking on the phone while standing next to a dog	
H1	A woman is on the phone	entailment
H2	A woman is walking her dog	neutral
H3	A woman is sleeping	contradiction
P	Tax records show Waters earned around $65,000 in 2000	
H1	Waters' tax records show clearly that he earned a lovely $65k in 2000	entailment
H2	Tax records indicate Waters earned about $65K in 2000	entailment
H3	Waters' tax records show he earned a blue ribbon last year	contradiction

Table 2.3: Examples from the development sets of SNLI (top) and MultiNLI (bottom). Each example contains one premise that is paired with three hypotheses in the datasets.

2009; Bowman, 2016). These datasets each contain over half a million examples and enabled researchers to apply data-hungry deep learning methods to RTE.

Unlike the RTE datasets, these two datasets were created by eliciting hypotheses from humans. Crowd-source workers were tasked with writing one sentence each that is entailed, neutral, and contradicted by a caption extracted from the Flickr30k corpus (Young et al., 2014a). Next, the label for each premise-hypothesis pair in the development and test sets were verified by multiple crowd-source workers and the majority-vote label was assigned for each example. Table 2.3 provides such examples for both datasets. Rudinger, May, and Van Durme (2017) illustrated how eliciting textual data in this fashion creates stereotypical biases in SNLI. Some of the biases are gender-, age-, and race-based. In Chapter 5, we will discusses other issues caused by this elicitation method.

2.2.2 Entailment as an Intermediate NLP Task

> By separating out the general problem of textual entailment from these task-specific problems, progress on semantic inference for many application areas can be promoted. Hopefully, research on textual entailment will finally lead to the development of entailment "engines", which can be used as a standard module in many applications (similar to the role of part-of-speech taggers and syntactic parsers in current NLP applications).
>
> (Giampiccolo et al., 2007)

NLP researchers have additionally argued for the usefulness of RTE in aiding NLP systems developed for applied, downstream tasks. For example, Bill MacCartney began his thesis, specifically Chapter 1.2 (MacCartney, 2009), by discussing some applications that RTE can help, e.g. question answering, semantic search, automatic summarization, and machine translation evaluation. Many of Ido Dagan's students' theses included components focused on leveraging RTE to help with downstream tasks. For example, Shachar Mirkin's thesis focused on using RTE "to investigat[e] the impact of context and discourse phenomena on inference" (Mirkin, 2011). Mirkin also demonstrated how RTE can be helpful for machine translation. In Aziz et al. (2010), RTE is used to "generat[e] alternative texts to the source sentence for translation" by replacing out-of-vocabulary words in a source sentence when translating a text from one language to another. Jonathan Berant's thesis extended this work by learning entailment rules between words and phrases, specifically predicates. Berant demonstrates that these learned entailment rules can help text exploration systems and applied this to a health-care domain where a health-care provider could "explor[e] relevant information about a given medical issue" (Berant, 2012).

There is also a large amount of work by others that leveraged RTE to improve performance for a wide range of NLP tasks. Bentivogli, Dagan, and Magnini (2017) summarize these works well in their recent survey paper:

> RTE methods originally developed based on the RTE datasets were later incorporated in various semantic applications, which have cast different inference needs in terms of textual entailment, and then used entailment technology to improve end-application performance. Examples of such works are educational tasks, including multiple choice comprehension tests (Clark, Harrison, and Yao, 2012) and answering science questions (Clark, Harrison, and Balasubramanian, 2012); evaluating tests (Miyao et al., 2012); answer validation in question answering (Harabagiu and Hickl, 2006; Rodrigo, Peñas, and Verdejo, 2009); relation extraction (Romano et al., 2006; Roth, Sammons, and Vydiswaran, 2009); machine translation evaluation (Pado et al., 2009); machine translation (Mirkin et al., 2009); multi-document summarization (Harabagiu, Hickl, and Lacatusu, 2007); text exploration (Adler, Berant, and Dagan, 2012); redundancy detection in Twitter (Zanzotto, Pennacchiotti, and Tsioutsiouliklis, 2011).

Modern deep learning research has also focused on using RTE to improve downstream tasks. Recent work by Mohit Bansal's group at UNC exemplifies this approach. They used Multi-task Learning (Guo, Pasunuru, and Bansal, 2018a; Guo, Pasunuru, and Bansal, 2018b) and Reinforcement Learning (Pasunuru and Bansal, 2018) to improve summarization and sentence-simplification models by leveraging large RTE datasets. They argue that sharing parameters between models trained to perform RTE and the tasks at hand "teaches the model to generate outputs that are entailed by the full input" (Guo, Pasunuru, and Bansal, 2018a). Additionally, sentence representation pre-trained on large RTE datasets have been shown to aid in other NLP tasks (Conneau et al., 2017). Phang, Févry, and Bowman (2018) refer to this practice as supplementary training on intermediate labeled-data tasks.

With the current zeitgeist of NLP research, now is a prime time to revisit RTE as a method to evaluate the inference capabilities of NLP models. In the current era of deep-learning, NLP researchers have increasingly moved beyond the single-task model paradigm where distinct models are developed to perform individual tasks. Although multi-task learning was presented over two decades ago (Caruana, 1993; Caruana, 1997), researchers have taken advantage of recently developed, easy-to-use deep-learning toolkits to seamlessly build single models (or at least models with shared parameters) to perform a multitude of tasks at once. In addition to relying on a range of evaluation metrics for each task that a single model performs, RTE can be used as a single metric that evaluates the inner workings of such complex models.

2.2.3 Entailment as a Downstream NLP Task

Coinciding with the recent "deep learning wave" that has taken over NLP and Machine Learning (Manning, 2015), the introduction of large scale RTE datasets led to a resurgence of interest in RTE amongst NLP researchers. Large scale RTE datasets focusing on specific domains, like grade-school scientific knowledge (Khot, Sabharwal, and Clark, 2018) or medical information (Romanov and Shivade, 2018), emerged. However, as the research community is fully devoted (or some might say blindly devoted) to the research agenda of developing NLP models that outperform each other on test corpora, this resurgence did not primarily focus on using RTE as a means to evaluate NLP systems. Rather, researchers primarily used these datasets to compete with one another to achieve the top score on leaderboards[9] for new RTE datasets.

2.3 Revisiting RTE as an NLP Evaluation

> It is worth noting that while these architectures have demonstrated strong performance, evaluation has been carried out almost exclusively on the SICK and SNLI datasets, and there has been little evidence to suggest they capture the type of compositional or world knowledge tested by other datasets like the FraCas test suite or the PASCAL challenge sets.
>
> (Pavlick, 2017)

As these large scale RTE datasets rapidly surged in popularity, some researchers critiqued the datasets' ability to test the inferential capabilities of NLP models. A high accuracy on these datasets does not indicate which types of reasoning RTE models perform or capture. These datasets cannot be used to determine how well an RTE model captures many desired capabilities of language understanding systems, e.g. paraphrastic inference, complex anaphora resolution (White et al., 2017), or compositionality (Pavlick and Callison-Burch, 2016; Dasgupta et al., 2018). In turn, researchers have recently created test suites to evaluate specific semantic phenomena (Pavlick, 2017; Naik et al., 2018a).

While *FraCas* and the PASCAL challenge sets require models to capture compositional or world knowledge, neither of these are adequate test sets in the era of deep

[9] `https://nlp.stanford.edu/projects/snli/`, `https://www.kaggle.com/c/multinli-matched-open-evaluation/leaderboard`, `https://leaderboard.allenai.org/scitail/submissions/public`

learning. While *FraCas* does indeed systematically cover a wide range of linguistic phenomena, the small size of the dataset limits the ability to meaningfully extrapolate from the results. Although determining the necessary "amount of data required ... to produce relevant performance measures" remains an open problem (Paroubek, Chaudiron, and Hirschman, 2007), 350 examples in *FraCas* is very small.

Unlike *FraCas*, the Pascal RTE challenge sets do not attempt to provide insight into reasoning capabilities or linguistic phenomena captured by NLP models. The single accuracy metric on these challenges indicates how well a model can recognize whether one sentence likely follows from another, but it does not illuminate how well NLP models capture different semantic phenomena that are important for general NLU. This issue was pointed out in Amoia (2008)'s thesis that presented "a test suite for adjectival inference developed as a resource for the evaluation of computational systems handling natural language inference."

Chen Zhang's thesis similarly focused on linguistic phenomena related to RTE. The thesis dealt with *conversational entailment*, "a task that determines whether a given conversation discourse entails a hypothesis about the participants" (Zhang and Chai, 2009). Later, the problem of *conversational entailment* was described as the "automated inference of hypotheses from conversation scripts" (Zhang and Chai, 2010). Zhang's thesis discusses semantic, pragmatic phenomena and world knowledge related to the task (Zhang, 2010).

In this thesis, we advocate for revisiting RTE as a framework to evaluate how well NLP models capture a suite of linguistic phenomena that are integral to NLU. We propose *recasting* as our solution; by leveraging existing annotations developed by computational semantics researchers via converting prior annotation of a specific phenomenon into RTE examples, recasting allows us to create a diverse RTE benchmark that tests a model's ability to perform distinct types of reasoning.

2.4 Natural Language Inference or Recognizing Textual Entailment?

Those familiar with the field are aware that the terms Natural Language Inference (NLI) and RTE are often used interchangeably. Many papers in the field on RTE begin by explicitly mentioning that these terms are synonymous (Liu et al., 2016; Gong, Luo, and Zhang, 2018; Camburu et al., 2018). In fact, variants of the phrase "natural language inference (NLI), also known as recognizing textual entailment (RTE)" appear in many papers (Chen et al., 2017b; Williams, Nangia, and Bowman, 2017; Naik et al., 2018b; Chen et al., 2018a; Tay, Luu, and Hui, 2018), including my own.

This thesis refers to the NLP task of predicting whether the truth condition of one sentence likely follows another primarily as RTE and not NLI. The broad phrase natural language inference is more appropriate for a class of problems that require making inferences from natural language. Tasks like sentiment analysis, event factuality, or even question-answering can be viewed as forms of natural language inference without having to convert them into the sentence pair classification format in RTE.[10] Earlier works used the term *natural language inference* in this way (Schwarcz, Burger, and Simmons, 1970; Wilks, 1975; Punyakanok, Roth, and Yih, 2004).

The leading term *recognizing* in RTE is fitting as the task is to classify or predict whether the truth of one sentence likely follows the other. The second term *textual* is similarly appropriate since the domain is limited to textual data. Critics of the name RTE often argue that the term *entailment* is inappropriate since the definition of the NLP task strays too far from the technical definition from *entailment* in linguistics.[11] In turn, both Zaenen, Karttunen, and Crouch (2005) and Manning (2006) prefer the term *textual inference* to describe the task. Additionally, Zaenen, Karttunen, and Crouch (2005) prefer the term *textual inference* because examples in the PASCAL RTE datasets required a system to not only identify entailments but also conventional implicatures, conversational implicatures, and world knowledge.

Based on these arguments, we would advocate for the new phrase *Recognizing Textual Inference*. However, given the choice between RTE and NLI, we prefer RTE since it is more representative of the task at hand.

[10] Dan Roth has made this argument in multiple settings.

[11] In personal correspondence, Ido Dagan commented that the term *textual entailment* was a way to differentiate RTE from the traditional (and stricter) definition in linguistics.

Chapter 3

The Diverse Natural Language Inference Collection

> Proper evaluation is a complex and challenging business. It implies, in particular, that we need to make a very rigorous "deconstructive" analysis of all the factors that affect the system being tested
>
> (Spärck Jones, 2005)

> A deeper and detailed analysis of . . . performance can provide the keys to exceed the current accuracy. Tests suites are a linguistic resource which makes it possible this kind of analysis and which can contribute to highlight the key issues to improve decisively the Natural Language Processing (NLP) tools (Flickinger et al., 1987; Blasband et al., 1999; Lehmann et al., 1996)
>
> (Lloberes, Castellón, and Padró, 2015)

The primary goal of this chapter is to introduce a collection of RTE datasets that each target a specific type of reasoning. These RTE datasets can be used to deconstruct general semantic inference into a set of factors that are integral to NLU. We begin this chapter with a detailed discussion on how we create these datasets. We recast existing annotations for a diverse set of semantic phenomena into RTE. We present details for 7 of the datasets and briefly discuss other phenomena recast into RTE by collaborators. The bulk of this chapter is based on Poliak et al. (2018a).

Second, the experiments in Section 3.4 explore how well models trained on these datasets can capture these different phenomena. We include results that determine whether models trained on other large RTE datasets or other DNC datasets capture the different phenomenon. We also analyze whether fine-tuning models that have been pre-trained on different datasets helps a model capture these phenomena.

Finally, we present results that demonstrate how quickly a model can perform well on these datasets. These experiments result in learning curves that plot the accuracy of a model on these datasets as we increase the number of training examples. These results have not been previously published.

3.1 Overview

As previously discussed, a plethora of new RTE datasets has been created in recent years (Bowman et al., 2015; Williams, Nangia, and Bowman, 2017; Lai, Bisk, and Hockenmaier, 2017; Khot, Sabharwal, and Clark, 2018). However, as we just argued, these datasets do not provide clear insight into what type of reasoning or inference a model may be performing. For example, these datasets cannot be used to evaluate whether competitive RTE models can determine if an event occurred, correctly differentiate between figurative and literal language, or accurately identify and categorize named entities. Consequently, these datasets cannot answer how well sentence representation learning models capture distinct semantic phenomena necessary for general natural language understanding (NLU).

To answer these questions, we introduce the **D**iverse **N**LI **C**ollection (DNC), a large-scale RTE test suite that evaluates a model's ability to perform diverse types of reasoning. The DNC is a collection of RTE problems, each requiring a model to perform a unique type of reasoning. Each RTE dataset contains labeled context-hypothesis pairs that are recast from semantic annotations for specific structured prediction tasks. We define recasting as leveraging existing datasets to create RTE examples (Glickman, 2006; White et al., 2017). In the first release of the DNC (DNC1.0), annotations are recast from a total of 13 datasets across 7 NLP tasks into labeled RTE examples. The tasks include event factuality, named entity recognition, datasets, gendered anaphora resolution, sentiment analysis, relationship extraction, pun detection, and lexicosyntactic inference. When first released, the DNC contained over half a million labeled examples. Currently, there are over a million examples in extensions to the DNC. Table 3.1 includes RTE pairs that test specific types of reasoning. Additionally, the DNC answers a recent plea to the community to test "more kinds of inference" than in previous challenge sets (Chatzikyriakidis et al., 2017).

Semantic Phenomena	✓	✗
Event Factuality	I would like to learn how The learning did not happen	I'll not say anything The saying happened
Named Entity Recognition	Ms. Rice said the United States must work intensively Ms. is a person 's title	Afghan officials are welcoming the Netherlands' decision The Netherlands is an event
Gendered Anaphora	The student met with the architect to view her blueprints for inspiration The architect has blueprints	The appraiser told the buyer that he had paid too much for the painting The appraiser had purchased a painting
MegaVeridicality	Someone assumed that a particular thing happened That thing might or might not have happened	A particular person craved to do a particular thing That person did that thing
VerbNet	The Romans destroyed the city The Romans caused the destroying	Andre presented the plaque Andre was transferred
VerbCorner	Molly wheeled Lisa to Rachel Someone moved from their location	Kyle bewildered Mark Someone or something changed physically
Relation Extraction	At least 100,000 Chinese live in Lhasa, outnumbering Tibetans two to one Tibetans live in Lhasa	Tropical storm Humberto is expected to reach the Texas coast tonight Humberto hit Texas
Puns	Jorden heard that my skiing skills are really going downhill Jorden heared a pun	Caiden heard that fretting cares make grey hairs Caiden heared a pun
Sentiment Analysis	When asked about the product, Liam said, "Don't waste your money" Liam did not like the product	When asked about the movie, Angel said, "A bit predictable" Angel liked the movie

Table 3.1: Example sentence pairs for different semantic phenomena. The ✓ and ✗ columns respectively indicate that the context entails, or does not entail the hypothesis. Each cell's first and second line respectively represent a context and hypothesis.

3.2 Motivation

The broad goals of recasting existing annotation from a diverse set of semantic phenomena into RTE is to 1) help determine whether an NLU model performs distinct types of reasoning, and 2) generate examples cheaply and at large scales.

NLU Insights

Popular RTE datasets, e.g. Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and its successor Multi-NLI (Williams, Nangia, and Bowman, 2017), were created by eliciting hypotheses from humans. Crowd-source workers were tasked with writing one sentence each that is entailed, neutral, and contradicted by a caption extracted from the Flickr30k corpus (Young et al., 2014a). Although these datasets are widely used to train and evaluate sentence representations, a high accuracy is not indicative of what types of reasoning RTE models perform. Workers were free to create any type of hypothesis for each context and label. Such datasets cannot be used to determine how well a model captures many desired capabilities of language understanding systems, e.g. paraphrastic inference, complex anaphora resolution (White et al., 2017), or compositionality (Pavlick and Callison-Burch, 2016; Dasgupta et al., 2018). By converting prior annotation of a specific phenomenon into RTE examples, recasting allows us to create a diverse RTE benchmark that tests a model's ability to perform distinct types of reasoning.

RTE Examples at Large-scale

Generating RTE datasets from scratch is costly. Humans must be paid to generate or label natural language text. Costs linearly scale as the amount of generated RTE-pairs increases. Existing annotations for a wide array of semantic NLP tasks are freely available. Each year, organizers for shared tasks at the International Workshop on Semantic Evaluation (SemEval)[1] release thousands of annotated examples for a wide range of NLP tasks. By leveraging existing semantic annotations already invested in and created by the community, we can generate and label RTE pairs at little cost, and create large RTE datasets that are necessary to train data hungry models.

Why These Semantic Phenomena?

A long term goal of NLP and AI research is to develop NLU systems that can achieve human levels of understanding and reasoning. Investigating how different architectures and training corpora can help a system perform human-level general NLU is an important step in this direction. The DNC contains recast RTE pairs that are easily understandable by humans and can be used to evaluate different sentence encoders and NLU systems. These semantic phenomena cover distinct types of reasoning that an NLU system may often encounter in the wild. While higher performance on these benchmarks might not be conclusive proof of a system achieving human-level reasoning, a system that does poorly should not be viewed as performing

[1] `https://en.wikipedia.org/wiki/SemEval`

human-level NLU. The semantic phenomena included in the DNC play integral roles in NLU. There exist more semantic phenomena integral to NLU (Allen, 1995) and they may be included in future versions of the DNC.

Previous Recast RTE

Example sentences in RTE1 (Dagan, Glickman, and Magnini, 2006) were extracted from MT, IE, and QA datasets, with the process referred to as 'recasting' in the thesis by Glickman (2006). NLU problems were reframed under the RTE framework and candidate sentence pairs were extracted from existing NLP datasets and then labeled under RTE (Dagan, Glickman, and Magnini, 2006). Years later, this term was independently used by White et al. (2017), who proposed to "leverage existing large-scale semantic annotation collections as a source of targeted textual inference examples." The term 'recasting' was limited to automatically converting existing semantic annotations into labeled RTE examples without manual intervention. We adopt the broader definition of 'recasting' since the RTE examples in the DNC were automatically or manually generated from prior NLU datasets.

Applied Framework versus Inference Probing

Traditionally, RTE has not been viewed as a downstream, applied NLP task (as discussed in Section 2.3).[2] Instead, the community has often used it as "a generic

[2] This changed as large RTE datasets have recently been used to train, or pre-train, models to perform RTE, or other tasks (Conneau et al., 2017; Pasunuru and Bansal, 2017).

evaluation framework" to compare models for distinct downstream tasks (Dagan, Glickman, and Magnini, 2006) or to determine whether a model performs distinct types of reasoning (Cooper et al., 1996). These two different evaluation goals may affect which datasets are recast. The DNC targets both goals as examples and annotations from applied tasks and linguistically focused phenomena are recast into RTE.

3.3 Recasting existing data into RTE

The DNC is a collaborative effort with researchers spanning across multiple institutions. The DNC includes RTE datasets that focus on many more phenomena, e.g. relation extraction, temporal reasoning, gendered anaphora resolution, and other types of lexico-syntactic inference, among others. The work for these datasets were primarily done by Ellie Pavlick, Siddharth Vashishtha, Rachel Rudinger, and Aaron Steven White. In this section, we discuss some of the semantic phenomena that we recast into RTE. For each of these phenomena, we discuss why they are important for NLU and how we recast their corresponding annotations into RTE.

3.3.1 Named Entity Recognition

Named Entity Recognition is the NLP task of identifying and classifying entities in text. Distinct types of entities have different properties and relational objects (Prince,

1978) that can help infer facts from a given context. For example, if a system can detect that an entity is a name of a nation, then that entity likely has a leader, a language, and a culture (Prince, 1978; Van Durme, 2010). When classifying RTE pairs, a model can determine if an object mentioned in the hypothesis can be a relational object typically associated with the type of entity described in the context. NER tags can also be directly used to determine if a hypothesis is likely to not be entailed by a context, such as when entities in contexts and hypotheses do not share NER tags (Castillo and Alemany, 2008; Sammons et al., 2009; Pakray et al., 2010).

Given a sentence annotated with NER tags, we recast the annotations by preserving the original sentences as contexts. We create hypotheses using the template "*NP* is a *Label*."[3] For ENTAILED hypotheses, *Label* is replaced with the correct NER label of the annotated noun phrase, for NOT-ENTAILED hypotheses, an incorrect label is chosen from the prior distribution of NER tags for the given phrase. We applied this procedure on the two NER dataset: the Gronigen Meaning Bank (Bos et al., 2017) and the ConLL-2003 Shared Task (Tjong Kim Sang and De Meulder, 2003).

3.3.2 Lexicosyntactic Inference

While many inferences in natural language are triggered by lexical items alone, there exist pervasive inferences that arise from interactions between lexical items and their syntactic contexts. This is particularly apparent among propositional attitude

[3] We ensure grammatical hypotheses by appropriately conjugating "is a" when needed.

verbs – e.g. *think*, *want*, *know* – which display complex distributional profiles (White and Rawlins, 2016). For instance, the verb *remember* can take both finite clausal complements and infinitival clausal complements.

(1) . Jo didn't **remember** *that she ate*

(2) . Jo didn't **remember** *to eat*

This small change in the syntactic structure gives rise to large changes in the inferences that are licensed: (1) presupposes that *Jo ate* while (2) entails that *Jo didn't eat.* The DNC recast data from three datasets that are relevant to these sorts of lexicosyntactic interactions. Here, we discuss how we recast two datasets to test lexicosyntactic inference.

3.3.2.1 VerbNet

VerbNet (Schuler, 2005) is a dataset containing classes of verbs that each can have multiple frames. Each frame contains a mapping from syntactic arguments to thematic roles, which are used as arguments in Neo-Davidsonian first-order logical predicates (4) that describe the frame's semantics. Each frame additionally contains an example sentence (3) that we use as RTE contexts. To generate hypotheses (6), we create templates (5) from the most frequent semantic predicates, determined using count-induced tree substitution grammars (Ferraro, Post, and Van Durme, 2012;

Ferraro, Van Durme, and Post, 2012).

(3) . Michael swatted the fly

(4) . cause(E, *Agent*)

(5) . *Agent* caused the E

(6) . Michael caused the swatting

We use the Berkeley Parser (Petrov et al., 2006) to match tokens in an example sentence with the thematic roles and then fill in the templates with the matched tokens (6). Multi-argument predicates are decomposed into unary predicates to increase the number of hypotheses generated. On average, each context is paired with 4.5 hypotheses. NOT-ENTAILED hypotheses are generated by filling in templates with incorrect thematic roles. This is similar to Aharon, Szpektor, and Dagan (2010)'s template matching to generate entailment rules from FrameNet (Baker, Fillmore, and Lowe, 1998). We partition the recast RTE examples into train/development/test splits such that all example sentences from a VerbNet class (which we use a RTE context) appear in only one partition of our dataset. In turn, the recast VerbNet dataset's partition is not exactly 80:10:10.

3.3.2.2 VerbCorner

VerbCorner (VC) (Hartshorne, Bonial, and Palmer, 2013) decomposes VerbNet predicates into simple semantic properties and "elicit[s] reliable semantic judgments corresponding to VerbNet predicates" via crowd-sourcing. The semantic judgments focus on movement, physical contact, application of force, change of physical or mental state, and valence, all of which "may be central organizing principles for a human's . . . conceptualization of the world." (Hartshorne, Bonial, and Palmer, 2013).

Each sentence in VC is judged based on the decomposed semantic properties. Each semantic property is converted into declarative statements to create hypotheses and they are paired with the original sentences, which are preserved as contexts. The RTE pair is ENTAILED or NOT-ENTAILED depending on the given sentence's semantic judgment.

The following templates are used for hypotheses, assigning them as ENTAILED and NOT-ENTAILED based on the positive or negative answers to the annotation task questions about the context sentence.

(7) . Someone {moved/did not move} from their location

(8) . Something touched another thing / Nothing touched anything else

(9) . Someone or something {applied/did not apply} force onto something

(10) . Someone or something {changed/did not change} physically

(11) . Someone {changed/did not change} their thoughts, feelings, or beliefs

(12) . Something {good/neutral/bad} happened

3.3.3 Event Factuality

Event factuality prediction is the task of determining whether an event described in text occurred. Determining whether an event occurred enables accurate inferences, e.g. monotonic inferences, based on the event (Rudinger, White, and Van Durme, 2018). Consider the following sentences:

(13) . She walked a beagle

(14) . She walked a dog

(15) . She walked a brown beagle

If the *walking* occurred, (13) entails (14) but not (15). If we negate the action in sentences (13), (14), and (15) to respectively become:

(16) . She did not walk a beagle

(17) . She did not walk a dog

(18) . She did not walk a brown beagle

the new hypothesis (18) is now entailed by the context (16) while (17) is not. Incorporating factuality to models has been shown to improve RTE predictions (Sauri and Pustejovsky, 2007).

Event factuality annotations from UW (Lee et al., 2015), MEANTIME (Minard et al., 2016), and Decomp (Rudinger, White, and Van Durme, 2018) are recast into RTE. Sentences from the original datasets are used as contexts and templates (19) and (20) are used as hypotheses.[4]

(19) $\underline{.}$ The *Event* happened

(20) $\underline{.}$ The *Event* did not happen

If the predicate denoting the *Event* was annotated as having happened in the factuality dataset, the context paired with (19) is labeled as ENTAILED and the same context paired with (20) is labeled as NOT-ENTAILED. Otherwise, the RTE labels are swapped.

3.3.4 Subjectivity (Sentiment)

Some of the previously discussed semantic phenomena deal with objective information – did an event occur or what type of entities does a specific name represent. Subjective information is often expressed differently (Wiebe, Wilson, and Cardie,

[4] *Event* is replaced with the event described in the context.

2005), making it important to use other tests to probe whether an NLU system understands language that expresses subjective information. We are interested in determining whether general NLU models capture ‘subjective clues’ that can help identify and understand emotions, opinions, and sentiment within a subjective text (Wilson, Wiebe, and Hwa, 2006), as opposed to differentiating between subjective and objective information (Yu and Hatzivassiloglou, 2003; Riloff, Wiebe, and Wilson, 2003).

We recast a sentiment analysis dataset since the task is the “expression of subjectivity as either a positive or negative opinion” (Taboada, 2016). We extract sentences from product, movie, and restaurant reviews labeled as containing positive or negative sentiment (Kotzias et al., 2015). The examples in this sentiment analysis dataset were compiled from previous sources. The movie dataset came from Maas et al. (2011), the Amazon product reviews were released by McAuley and Leskovec (2013) add the restaurant reviews were sourced from the Yelp dataset challenge.[5]

When recasting this data into RTE, we generate contexts (21) and hypotheses (22), (23) using the following templates:

(21) $\underline{.}$ When asked about *Item*, *Name* said *Review*

(22) $\underline{.}$ *Name* liked the *Item*

(23) $\underline{.}$ *Name* did not like the *Item*

[5] http://www.yelp.com/dataset_challenge

Item is replaced with either "product", "movie", or "restaurant", and the *Name* is sampled as previously discussed. If the original sentence contained positive (negative) sentiment, the (21)-(22) pair is labeled as ENTAILED (NOT-ENTAILED) and (21)-(23) is labeled as NOT-ENTAILED (ENTAILED).

3.3.5 Figurative Language (Puns)

Figurative language demonstrates natural language's expressiveness and wide variations. Understanding and recognizing figurative language "entail[s] cognitive capabilities to abstract and meta-represent meanings beyond *physical* words" (Reyes, Rosso, and Buscaldi, 2012). Puns are prime examples of figurative language that may perplex general NLU systems as they are one of the more regular uses of linguistic ambiguity (Binsted, 1996) and rely on a wide-range of phonetic, morphological, syntactic, and semantic ambiguity (Pepicello and Green, 1984; Binsted, 1996; Bekinschtein et al., 2011).

We recast puns from Yang et al. (2015) and Miller, Hempelmann, and Gurevych (2017) using templates to generate contexts (24) and hypotheses (25), (26). We replace *Name* with names sampled from a distribution based on US census data,[6] and *Pun* with the original sentence. If the original sentence was labeled as containing a pun, the (24)-(25) pair is labeled as ENTAILED and (24)-(26) is labeled as NOT-ENTAILED, otherwise we swap the labels. In total, we generate roughly 15K labeled

[6] `http://www.ssa.gov/oact/babynames/names.zip`

pairs.

(24) . *Name* heard that *Pun*

(25) . *Name* heard a pun

(26) . *Name* did not hear a pun

Puns in Yang et al. (2015) were originally extracted from `punsoftheday.com`, and sentences without puns came from newswire and proverbs. The sentences are labeled as containing a pun or not. Puns in Miller, Hempelmann, and Gurevych (2017) were sampled from prior pun detection datasets (Miller and Gurevych, 2015; Miller and Turković, 2016) and includes new examples generated from scratch for the shared task; the original labels denote whether the sentences contain homographic, heterographic, or no pun at all. Here, we are only interested in whether a sentence contains a pun or not instead of discriminating between homographic and heterographic puns.

3.4 Experiments & Results

Here, we will first discuss results of RTE models trained on these datasets. We will then demonstrate how to use these recast RTE datasets to evaluate an RTE model trained on a prior popular dataset, MNLI. In the next chapter we will discuss how these datasets can be used to evaluate the reasoning the capabilities of NLP models

Sem. Phenomena	Dataset	# pairs	Automated
Event Factuality	Decomp (Rudinger, White, and Van Durme, 2018)	42K (41,888)	✓
	UW (Lee et al., 2015)	5K (5,094)	✓
	MeanTime (Minard et al., 2016)	.7K (738)	✓
Named Entity Recognition	Groningen (Bos et al., 2017)	260K (261,406)	✓
	CoNLL (Tjong Kim Sang and De Meulder, 2003)	60K (59,970)	✓
Gendered Anaphora	Winogender (Rudinger et al., 2018)	.4K (464)	✗
Lexicosyntactic Inference	VerbCorner (Hartshorne, Bonial, and Palmer, 2013)	135K (138, 648)	✓
	MegaVeridicality (White and Rawlins, 2018)	11K (11,814)	✓
	VerbNet (Schuler, 2005)	2K (1, 759)	✓✗
Puns	(Yang et al., 2015)	9K (9,492)	✓
	SemEval 2017 Task 7 (Miller, Hempelmann, and Gurevych, 2017)	8K (8, 054)	✓
Relationship Extraction	FACC1 (Gabrilovich, Ringgaard, and Subramanya, 2013)	25K (25,132)	✓✗
Sentiment Analysis	(Kotzias et al., 2015)	6K (6,000)	✓
Combined	Diverse NLI Collection (DNC)	570K (570,459)	
—	SNLI (Bowman et al., 2015)	570K	
—	Multi-NLI (Williams, Nangia, and Bowman, 2017)	433K	

Table 3.2: Statistics summarizing the recast datasets in the first release of the DNC. The first column refers to the original annotation that was recast, the ‘Combined‘ row refers to the combination of our recast datasets. The second column indicates the datasets that were recast, and the 3rd column reports how many labeled RTE pairs were extracted from the corresponding dataset. The last column indicates whether the recasting method was fully-automatic without human involvement (✓), manual (✗), or used a semi-automatic method that included human intervention (✓✗). The Multi-NLI and SNLI numbers contextualize the scale of our dataset.

trained for other tasks, like machine translation or syntactic parsing.

Models

For demonstrating how well an RTE model performs these fine-grained types of reasoning, we use `InferSent` (Conneau et al., 2017). `InferSent` independently encodes a context and hypothesis with a bi-directional LSTM and combines the sentence representations by concatenating the individual sentence representations, their element-wise subtraction and product. The combined representation is then fed into a MLP with a single hidden layer.

Recast Data / Model	NER	EF	RE	Puns	Sentiment	GAR	VC	MV	VN
Majority (MAJ)	50.00	50.00	59.53	50.00	50.00	50.00	50.00	66.67	53.66
No Pre-training									
InferSent	**92.50**	83.07	61.89	60.36	50.00	–	88.60	**85.96**	46.34
Pre-trained DNC									
InferSent *(update)*	92.47	**83.86**	74.38	**93.17**	81.00	–	**89.00**	85.62	76.83
InferSent *(fixed)*	92.20	81.07	74.11	87.76	77.33	**50.65**	88.59	83.84	67.68
Pre-trained Multi-NLI									
InferSent *(update)*	92.37	83.03	**76.08**	92.48	**83.50**	–	88.45	85.11	**78.05**
InferSent *(fixed)*	52.99	54.88	66.75	56.04	56.50	**50.65**	45.33	55.92	45.73

Table 3.3: RTE accuracies on test data. Columns correspond to each semantic phenomena and rows correspond to the model used. Columns are ordered from larger to smaller in size, but the last three (VC, MV, VN) are separated since they fall under lexico-syntactic inference. *(update)* refers to a model that was initialized with pre-trained parameters and then re-trained on the corresponding recast data. *(fixed)* refers to a model that was trained and then evaluated on these data sets. Bold numbers in each column indicate which settings were responsible for the highest accuracy on the specific recast dataset.

Experimental Details

In these experiments, we use pre-computed GloVe embeddings (Pennington, Socher, and Manning, 2014) and use the OOV vector for words that do not have a defined embedding. We follow Conneau et al. (2017)'s procedure to train these models. During training, models are optimized with stochastic gradient descent. The initial learning rate is 0.1 with a decay rate of 0.99. The models train for at most 20 epochs and can optionally terminate early when the learning rate is less than 10^{-5}. If the accuracy deceases on the development set in any epoch, the learning rate is divided by 5.

Results

Table 3.3 reports the models' accuracies across the recast RTE datasets.[7] Even though we categorize VerbNet, MegaVeridicality, and VerbCorner as lexicosyntatic inference, we train and evaluate models separately on these three datasets because different strategies were employed to individually recast them. When evaluating RTE models, the baseline is the the majority class label (MAJ). We do not train on the gendered anaphora resolution dataset because of its small size. It is used here just as a testset.

The results suggest that `InferSent`, when not pre-trained on any other data, might capture specific semantic phenomena better than other semantic phenomena. `InferSent` seems to learn the most about determining if an event occurred. The model seems to similarly learn to perform (or detect) the type of lexico-syntactic inference present in VC and MV.

Pre-training models on DNC

Does initializing models with pre-trained parameters improves scores? Notice that when models are pre-trained on DNC, for the larger datasets, a pre-trained model does not seem to significantly outperform randomly initializing the parameters. For the smaller datasets, specifically Puns, Sentiment and VN, a pre-trained model

[7] These results are on all the RTE datasets in the first release of the DNC (`https://github.com/decompositional-semantics-initiative/DNC/releases/tag/v0.1`), some of which were not described in this chapter. See Poliak et al. (2018a) for a description of the remaining datasets.

significantly outperforms random initialization by 32.81, 31.00, and 30.83 respectively.

We are also interested to know whether fine-tuning these pre-trained models on each category (*update*) improves a model's ability to perform well on the category compared to keeping the pre-trained models' parameters static (*fixed*). Across all of the recast datasets, updating the pre-trained model's parameters during training improves `InferSent`'s accuracies more than keeping the model's parameters fixed. When updating a model pre-trained on the entire DNC, we see the largest improvements on VN (+9.15).

Models trained on Multi-NLI

Williams, Nangia, and Bowman (2017) argue that Multi-NLI "[makes] it possible to evaluate systems on nearly the full complexity of the language." However, how well does Multi-NLI test a model's capability to understand the diverse semantic phenomena captured in DNC? We posit that if a model, trained on and performing well on Multi-NLI, does not perform well on our recast datasets, then Multi-NLI might not evaluate a model's ability to understand the "full complexity" of language as argued.[8]

When trained on Multi-NLI, the `InferSent` model achieves an accuracy of 70.22% on (matched) Multi-NLI.[9] When the models are tested on the recast datasets (with-

[8] We treat Multi-NLI's NEUTRAL and CONTRADICTION labels as equivalent to the DNC's NOT-ENTAILED label.

[9] Although this is about 10 points below SoTA, we believe that the pre-trained model performs well enough to evaluate whether Multi-NLI tests a model's capability to understand the diverse semantic phenomena in the DNC.

out updating the parameters), we see significant drops.[10] On the datasets testing a model's lexico-syntactic inference capabilities, the model performs below the majority class baseline. On the NER, EF, and Puns datasets its performs below the hypothesis-only baseline. We also notice that on three of the datasets (EF, Puns, and VN), the fixed hypothesis-only model outperforms the fixed `InferSent` model.

These results might suggest that Multi-NLI does not evaluate whether sentence representations capture these distinct semantic phenomena. This is a bit surprising for some of the recast phenomena. We would expect Multi-NLI's fiction section (especially its humor subset) in the training set to contain some figurative language that might be similar to puns, and the travel guides (and possibly telephone conversations) to contain text related to sentiment.

Pre-training on DNC or Multi-NLI?

Initializing a model with parameters pre-trained on DNC or Multi-NLI often outperforms random initialization.[11] Is it better to pre-train on DNC or Multi-NLI? On five of the recast datasets, using a model pre-trained on DNC outperforms a model pre-trained on Multi-NLI. The results are flipped on the two datasets focused on downstream tasks (Sentiment and RE) and MV. However, the differences between pre-training on the DNC or Multi-NLI are small. From this, it is unclear whether pre-training on DNC is better than Multi-NLI.

[10] `InferSent` (*pre-trained, fixed*) in Table 3.3.

[11] Pre-training does not improve accuracies on NER or MV.

Semantic Phenomena	Template
Event Factuality	The *event* happened
Named Entity Recognition	*Entity* is a *label*
VerbNet	The *Agent* caused *verb*
VerbCorner	Someone moved from their location
Puns	*Name* heard a pun
Sentiment	*Name* liked the *item*

Table 3.4: Example templates used for some semantic phenomena in the DNC. Words in italics represent slots that are filled in when converting existing annotations for the corresponding semantic phenomena into RTE.

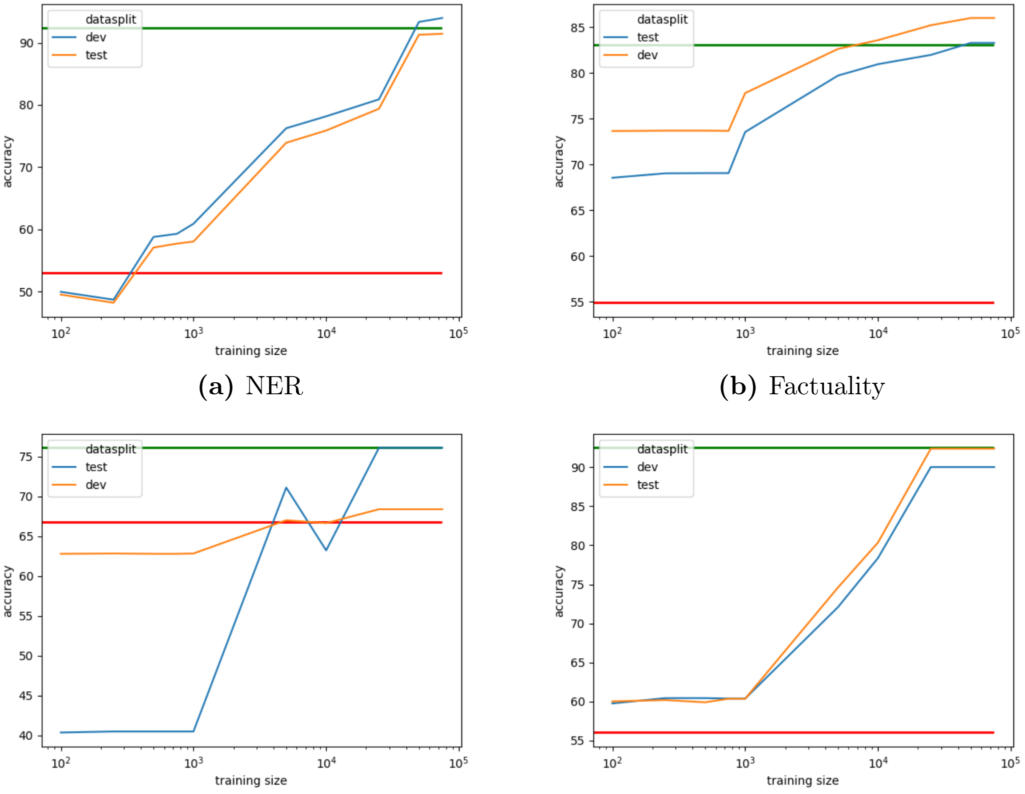
Learning curves/Incremental Training

Many of the recasting methods in the DNC rely on creating templates for hypotheses (Table 3.4). In the experiments just discussed, when the pre-trained model was allowed to update its parameters on each semantic phenomena's corresponding training set in the DNC, the model often greatly improved. Therefore, given enough training instances, a model either overcomes (or maybe learns) the templatic nature of many examples in the DNC or learns to perform the type of reasoning tested by that specific dataset. Here we focus on determining how many training examples is enough for a model pre-trained on an existing dataset, e.g. Multi-NLI, to learn either the templatic nature or the diverse types of reasoning tested in the DNC.

Experimental Setup

For each DNC dataset, we create subsets of each training dataset that differ in size. The training set sizes are 100, 250, 500, 750, $1K$, $5K$, $10K$, $25K$, $50K$ and $75K$.

For smaller DNC datasets, we cap the training set sizes. Additionally, we do not include the DNC's Gendered Anaphora Resolution dataset since it does not contain a training set for us to update the hyper-parameters.



(a) NER

(b) Factuality

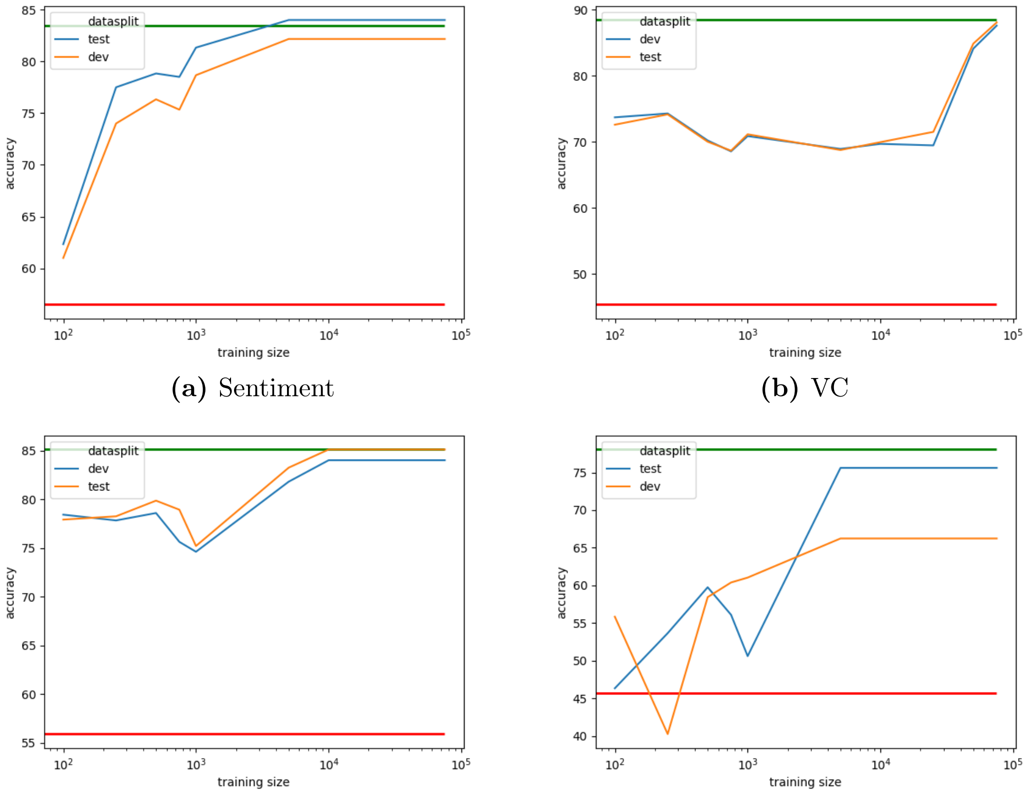
(c) RE

(d) Puns

Figure 3.1: Results of updating a model pre-trained on Multi-NLI on 4 of the DNC datasets with various training sizes. The horizontal lines represent the numbers reported in Poliak et al. (2018a) when they did not update a pre-trained model on each DNC dataset (red) or when they updated the pre-trained model on each corresponding DNC dataset (green). The blue/orange lines represent the accuracy (x-axes) on dev/test on each DNC dataset for the corresponding amount of training data (y-axes).

Results

Figure 3.1 and Figure 3.2 plot the results when we update the pre-trained models' parameters on each of the 8 DNC datasets under consideration. The red horizontal lines indicate our baseline, i.e. testing a model trained on MNLI, and the green horizontal line indicates the ceiling, i.e. fine-tuning the model on the entire DNC



(a) Sentiment

(b) VC

(c) MV

(d) VN

Figure 3.2: Results of updating a model pre-trained on Multi-NLI on 4 of the DNC datasets with various training sizes. The horizontal lines represent the numbers reported in Poliak et al. (2018a) when they did not update a pre-trained model on each DNC dataset (red) or when they updated the pre-trained model on each corresponding DNC dataset (green). The blue/orange lines represent the accuracy (x-axes) on dev/test on each DNC dataset for the corresponding amount of training data (y-axes).

dataset under consideration. Notice the gaps between red and the models when trained on just 10^2 examples in the DNC datasets focused on Sentiment (3.2a), Factuality (3.1b), and Puns (3.1d). By design, these datasets each have majority baselines of 50% since each sentence in the original corresponding datasets were paired with two new hypotheses, one ENTAILED and one NOT-ENTAILED. In the training set for Factuality over 71% of the events actually occurred, and in the training set for Puns 60% of the examples contained a pun. In turn, the same percentages of premises paired with the hypothesis that "The *event* happened" and "*Name* heard a pun" are be labeled as ENTAILED. In two of the three RTE datasets focused on lexico-syntactic inference, MegaVeridicality (3.2c) and VerbCorner (3.2b), we see even larger gaps between the baseline and the model that is trained on 100 examples. In the recast VerbCorner, almost 74% of the example where the hypothesis does not include "did not", "Nothing", or "bad" the label is True.

This explains the improvements on many of these recast RTE datasets when we allow a model to update its parameters on just 100 examples of each type of semantic phenomena in RTE form. In essence, the majority baseline in some DNC datasets might be a low estimate of a more indicative and true majority baseline. Including large recast datasets can be helpful for fine-tuning or pre-training, but for evaluation sets, these results suggest that we should re-consider the practice of duplicating contexts to ensure an artificially low 50% majority baseline.

There are other interesting trends in the plots to note. First, in the Sentiment

(3.2a) and VerbNet (3.2d) plots, we see that when the model is fine-tuned on the entire training dataset, it performs a bit below the results from earlier in the chapter (Table 3.3). Schluter and Varab (2018) demonstrate that permuting the order of a training set can change an RTE model's results on a test set, and additional experiments confirmed that this likely explains what is happening here.

In addition to differences with the ceiling for each model from Table 3.3, we note that sometimes fine-tuning a model on a small portion of a recast training set performs worse than not fine-tuning at all. This drastically occurs for the recast Relation Extraction (3.1c) and briefly for the recast NER datasets (3.1a).

3.5 Discussion

In the chapter, we described how we recast a wide range of semantic phenomena from many NLP datasets into labeled DNC sentence pairs. These examples serve as a diverse RTE suite that may help diagnose whether NLP models capture and perform distinct types of reasoning. The DNC is actively growing as we continue recasting more datasets into labeled RTE examples. We encourage dataset creators to recast their datasets in RTE and invite them to add their recast datasets into the DNC.

Since the introduction and the initial release of the DNC at EMNLP 2018, the DNC has grown. At the 2018 JSALT Summer Workshop, Najoung Kim and Ellie Pavlick led an effort to include more phenomena related to function words in the

DNC. The community recognized our work with a Best Paper Award at StarSem 2019 (Kim et al., 2019). Recently, Siddharth Vashishtha led an effort creating RTE datasets focused on temporal reasoning, specifically how long an event took place and the order of events (Vashishtha et al., 2020).

Additionally, recent efforts similarly create new RTE datasets that evaluate more phenomena. These include implicatures and presuppositions (Jeretic et al., 2020a), verb veridicality (Ross and Pavlick, 2019), monotonicity (Yanaka et al., 2019), and others. Staliūnaitė (2018)'s master's thesis improved our recasting method to create more natural hypotheses in the DNC dataset focused on factuality. There are also efforts to versions of the DNC in other languages. Rajiv Ratn Shah's group at IIITD released a Hindi RTE dataset which they created by recasting annotations from Hindi sentiment analysis and emotion detection datasets.[12]

Similar to the efforts described here to recast different NLU problems as RTE, others have recast NLU problems into a question answer format (McCann et al., 2018). Recasting problems into RTE, as opposed to question-answering, has deeper roots in linguistic theory, and continues a rich history within the NLP community.

[12] https://github.com/midas-research/hindi-nli-data

To leaderboard or not

The last few years have seen a rise in leaderboards in the academic community.[13] Leaderboards have been used in NLP related courses to encourage healthy competition amongst students (Lopez et al., 2013) and many government backed research programs used leaderboards to evaluate competitors. The long running SemEval competitions also have relied on leaderboards. However, a recent phenomena in our field has been to aggregate and host existing datasets on one platform to make it easy for researchers to develop and test a single model and compete across a large suite of benchmarks (Wang et al., 2018; Wang et al., 2019a). We intentionally did not create a leaderboard for the DNC datasets. RTE is primarily an evaluation framework and the goal of this work is not to create a dataset that researchers compete on. The DNC is a test suite to evaluate how well an NLP system captures specific phenomenon that are related to downstream NLP tasks.

[13] Allen Institute for AI's NLP highlights podcasts has an interesting episode with Siva Reddy about learderboards in the field - `https://soundcloud.com/nlp-highlights/80-leaderboards-and-science-with-siva-reddy`.

Chapter 4

Exploring Semantic Phenomena in neural models via recast-RTE

> Without a common classification of the problems in natural language understanding authors have no way to specify clearly what their systems do, potential users have no way to compare different systems and researchers have no way to judge the advantages or disadvantages of different approaches to developing systems.
>
> (Read et al., 1988)

Now that we have introduced the DNC, we will turn towards evaluating how well NLP models trained on different tasks capture the diverse semantic phenomena in the large collection of RTE datasets. We will use the DNC as *a common classification of the problems in natural language understanding* that can enable us to *specify clearly* what systems might do. We begin this chapter by describing a general modeling framework we will utilize to evaluate the reasoning capabilities of different NLP models. We will then demonstrate how to use this framework to evaluate the reasoning

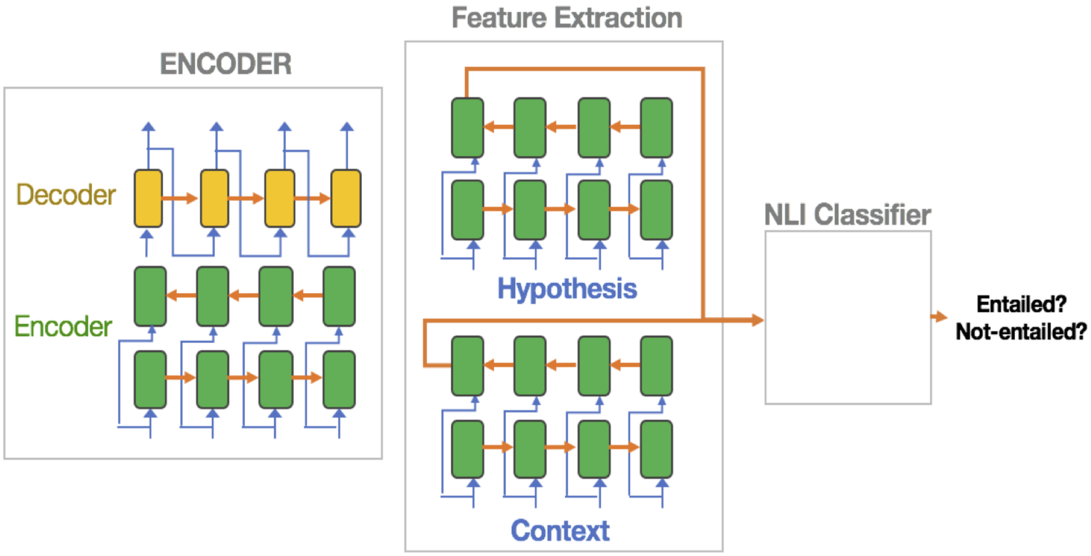


Figure 4.1: The general evaluation framework we propose when using RTE to evaluate the reasoning capabilities of contemporary neural NLP systems. The first step is to train an encoder as part of broader NLP system. Next, we freeze the encoder and use it to extract sentence representations. Finally, we use these representations as features to train an RTE classifier.

capabilities of NLP models trained to translate text from one language to another, match images with corresponding captions, or parse sentences into syntactic chunks.

The experiments evaluating how well a machine translation captures different semantic phenomena is based on Poliak et al. (2018c). The experiments evaluating the other NLP models are based on unpublished results from the 2018 Fred Jelinek Summer Workshop led by Sam Bowman and Ellie Pavlick.[1]

[1] https://jsalt18-sentence-repl.github.io/. A recording based on the second set of experiments is available online at https://www.youtube.com/watch?v=a-XhUdBWZDE&t=7625s

4.1 General Method

Figure 4.1 demonstrates the general framework we introduce for using RTE to evaluate how well different NLP models capture a diverse range of semantic phenomena. This general method has three main components: 1) training an NLP system; 2) extracting sentence representations from the trained system; and 3) training a classifier for different DNC datasets based on those sentence representations.

Training an NLP system

Contemporary NLP systems often rely on neural-network based encoders to convert input words and sentences into meaningful vector representations. The left portion of the figure, titled encoder, denotes a sequence to sequence model. This is an end-to-end neural-based system that encodes the input text (green nodes in the bottom figure) and generates output text from a decoder (yellow nodes in the top of the figure). Sequence to sequence models are often used for tasks like machine translation or summarization, where the decoder generates a translation or summary of the text encoded by the encoder.

This general framework is not limited to sequence to sequence models. This framework can be used to evaluate models that tag edges between tokens or predict a label for an individual sentence. For such settings, the decoder would be replaced with a classifier. The example in Figure 4.1 depicts a sequence to sequence model since we will begin the study in the chapter (Section 4.2) by evaluating how well a neural machine translation model captures different semantic phenomena.

Extracting sentence representations

After training a neural NLP system to translate sentences from one language to another, match pictures with corresponding captions, or parse sentences into syntactic chunks, we will use the trained encoders as feature extractors to create sentence representations of the input data for the different DNC datasets. During this process, we freeze the trained encoders, i.e. we do not update the parameters for encoders. This enables us to test how well these encoders capture the different specific phenomenon under consideration.

DNC specific classifier

Finally, after using the trained encoders to generate vector representations for the input data from the different DNC datasets, we train a classifier to predict whether the premises entail the hypotheses in each specific DNC dataset. The representations extracted from the encoders are fed as input to the classifier. If the sentence representations learned by the neural NLP systems capture distinct semantic phenomena that are integral to NLU, then the classifier should be able to perform well on the RTE datasets that test a model's ability to capture these different phenomenon.

This general framework is flexible and allows us to choose between the different methods for sentence encoders. The experiments here test sentence representations bidirectional RNN's, Bi-LSTM's in particular.

4.2 Machine Translation

We begin by demonstrating how recast RTE datasets can be used to study what do neural machine translation (NMT) models learn about semantics? We begin with machine translation since many researchers suggest that state-of-the-art NMT models learn representations that capture the meaning of sentences (Gu et al., 2016; Johnson et al., 2017; Zhou et al., 2017; Andreas and Klein, 2017; Neubig, 2017; Koehn, 2017). However, there is limited understanding of how specific semantic phenomena are captured in NMT representations beyond this broad notion. For instance, how well do these representations capture Dowty (1991)'s semantic proto-roles? Are these representations sufficient for understanding paraphrastic inference? Do the sentence representations encompass complex anaphora resolution? Existing semantic annotations recast as RTE can be leveraged to investigate whether sentence representations encoded by NMT models capture these semantic phenomena.

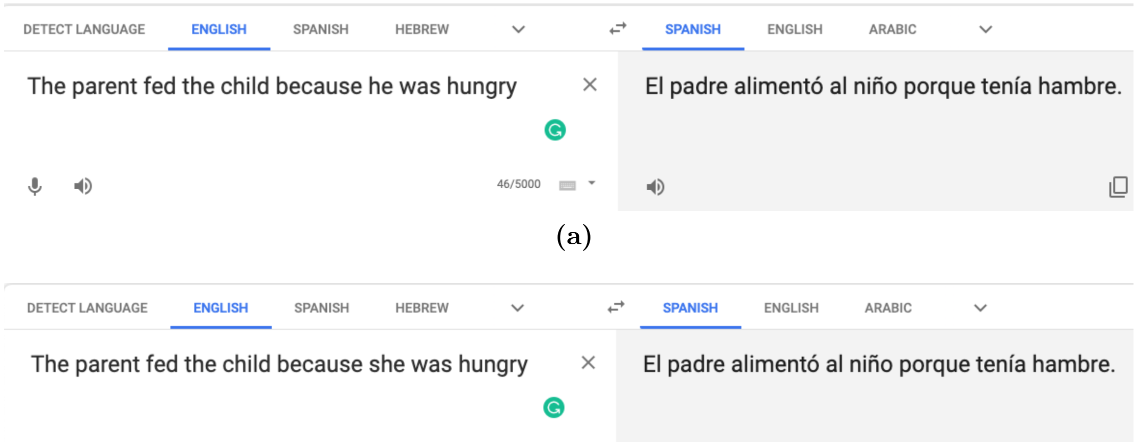
We use sentence representations from pre-trained NMT encoders as features to train classifiers for different recast RTE datasets. If the sentence representations learned by NMT models capture distinct semantic phenomena, we hypothesize that those representations should be sufficient to perform well on RTE datasets that test a model's ability to capture these phenomena.

We evaluate NMT sentence representations of 4 NMT models from 2 domains on 4 different RTE datasets to investigate how well they capture different semantic phenomena. In particular, we use White et al. (2017)'s *Unified Semantic Evaluation Framework* (USEF) that recasts three semantic phenomenon into RTE. These phenomena 1) semantic proto-roles, 2) paraphrastic inference, 3) and complex anaphora resolution. These three datasets awee a precursor to the DNC. Additionally, I evaluate the NMT sentence representations on 4) Multi-NLIs (Williams, Nangia, and Bowman, 2017). We contextualize the results with a standard neural encoder described in Bowman et al. (2015) and used in White et al. (2017).

Based on the recast RTE datasets, this investigation suggests that NMT encoders might learn more about semantic proto-roles than anaphora resolution or paraphrastic inference. Additionally, the experiments suggest that the target-side language affects how an NMT source-side encoder captures these semantic phenomena.

4.2.1 Motivation

Here we describe why it is appropriate to test how well NMT models capture anaphora resolution, semantic proto-role, and paraphrastic inference. We argue why we should expect high performing NMT models to capture these phenomena.



The parent fed the child because he was hungry

El padre alimentó al niño porque tenía hambre.

(a)

The parent fed the child because she was hungry

El padre alimentó al niño porque tenía hambre.

(b)

Figure 4.2: Screenshot from August 12th 2020 of correct (a) and miscorrect (b) translations in Google Translate based on correct/incorrect anaphora resolution.

Anaphora

Anaphora resolution connects tokens, typically pronouns, to their referents. Anaphora resolution should occur when translating from morphologically poor languages into some morphologically rich languages. For example, when translating "The parent fed the child because she was hungry," a Spanish translation should describe *the child* as *la niña (fem.)* and not *el niño (masc.)* since *she* refers to *the child*. Because world knowledge is often required to perform anaphora resolution (Rahman and Ng, 2012; Javadpour, 2013), this may enable evaluating whether an NMT encoder learns world knowledge. In this example, *she* refers to *the child* and not *the parent* since world knowledge dictates that parents often feed children when children are hungry.

When this work was originally published in 2018, Google Translate incorrectly translated this example. When checked in November 2019, Google Translate corrected translated this example. However, when checked again on on August 12th 2020 Google

Translate incorrectly translated this example, as demonstrated in Figure 4.2.

Proto-roles

Inspired by Dowty (1991)'s thematic role theory, Reisinger et al. (2015) introduced the Semantic Proto-Role (SPR) labeling task, which can be viewed as decomposing semantic roles into finer-grained properties, such as whether a predicate's argument was likely *aware* of the given predicated situation.

Dowty (1991)'s proto-roles may be expressed differently in different languages, and so correctly identifying them can be important for translation. For example, English does not usually explicitly mark *volition*, a proto-role, except by using adverbs like *intentionally* or *accidentally*. Other languages mark volitionality by using special affixes (e.g., Tibetan and Sesotho, a Bantu language), case marking (Hindi, Sinhalese), or auxiliaries (Japanese).[2] Correctly generating these markers may require the MT system to encode volitionality on the source side.

Paraphrases

Callison-Burch (2007) discusses how paraphrases help machine translation when alignments from source words to target-language words are unknown. If the alignment model can map a paraphrase of the source word to a word in the target language, then the machine translation model can translate the original word based on its paraphrase. Using paraphrases can also help NMT models generate text in the target language in

[2] For references and examples, see: `en.wikipedia.org/wiki/Volition_(linguistics)`.

some settings (Sekizawa, Kajiwara, and Komachi, 2017). Paraphrases are also used by professional translators to deal with non-equivalence of words in the source and target languages (Baker, 2018).

4.2.2 Experiments

To test how well NMT models capture these semantic phenomena, we use NMT models on four language pairs: English → {Arabic (ar), Spanish (es), Chinese (zh), and German (de)}. The first three pairs use the United Nations parallel corpus (Ziemski, Junczys-Dowmunt, and Pouliquen, 2016) and the English-German is trained on the WMT dataset (Bojar et al., 2014). Although the entailment classifier only uses representations extracted from the English-side encoders as features, using multiple language pairs enables us to explore whether different target languages affect what semantic phenomena are captured by an NMT encoder.

The neural machine translation models are based on bidirectional long short-term memory (Bi-LSTM) encoder-decoders with attention (Sutskever, Vinyals, and Le, 2014; Bahdanau, Cho, and Bengio, 2014). The encoders and decoders have 4-layers with 500-dimensional word embeddings and LSTM states (i.e., $d = 500$). The vocabulary size is 75K words. The NMT models are trained until convergence and the models that performed best on the machine translation datasets' development sets are used here. Following common NMT practice (Cho et al., 2014), long sentences ($>$ 50 words) are removed when training the NMT models. We train

English→Arabic/Spanish/Chinese NMT models on the first 2 million sentences of the United Nations parallel corpus training set (Ziemski, Junczys-Dowmunt, and Pouliquen, 2016), and the English→German model on the WMT dataset (Bojar et al., 2014). We use the official training/development/test splits.

After these NMT models are trained, we are ready to evaluate how well they capture these semantic phenomena. Given an RTE context-hypothesis pair, each sentence is encoded independently through a trained NMT encoder to extract their respective vector representations. We represent each sentence by concatenating the last hidden state from the forward and backward encoders, resulting in $\mathbf{v}$ and $\mathbf{u}$ (in $\mathbb{R}^{2d}$) for the context and hypothesis.[3] We follow the common practice of feeding the concatenation $(\mathbf{v}, \mathbf{u}) \in \mathbb{R}^{4d}$ to a classifier (Rocktäschel et al., 2015; Bowman et al., 2015; Mou et al., 2016; Liu et al., 2016; Cheng, Dong, and Lapata, 2016; Munkhdalai and Yu, 2017).

Sentence pair representations are fed into a classifier with a soft-max layer that maps onto the number of labels. Experiments with both linear and non-linear classifiers have not shown major differences, so we will report results with the linear classifier unless noted otherwise. In preliminary experiments, we also used a 3-layered MLP. Although the results slightly improved, we noted similar trends to the linear classifier.

[3] We experimented with other sentence representations and their combinations, and did not see differences in overall conclusions.

	DPR	SPR	FN+
Train	2K	123K	124K
Dev	.4K	15K	15K
Test	1K	15K	14K

Table 4.1: Number of sentences in RTE datasets under consideration.

Recast Recognizing Textual Entailment data

The RTE datasets are trained on previous recast RTE data that tests these semantic phenomena. Sentence-pairs and labels were recast from existing semantic annotations: FrameNet Plus (FN+) (Pavlick et al., 2015), Definite Pronoun Resolution (DPR) (Rahman and Ng, 2012), and Semantic Proto-Roles (SPR) (Reisinger et al., 2015). The FN+ portion contains sentence pairs based on paraphrastic inference, DPR's sentence pairs focus on identifying the correct antecedent for a definite pronoun, and SPR's sentence pairs test whether the semantic proto-roles from Reisinger et al. (2015) apply based on a given sentence.[4] Table 4.1 includes the datasets' statistics.

4.2.3 Results

Table 4.2 shows results of RTE classifiers trained on representations from different NMT encoders. We also report the majority baseline and the results of Bowman et al.'s 3-layer deep 200 dimensional neural network used by White et al. ("USEF"). We

[4] In Section 5.3, we provide a summary of how these datasets were recast, and in Section 5.5 we examine issues in these recasting methods. See White et al. (2017) for a further detailed discussion on how the existing datasets were recast into RTE.

Test	DPR: 50.0					SPR: 65.4					FN+: 57.5				
Train	ar	es	zh	de	USEF	ar	es	zh	de	USEF	ar	es	zh	de	USEF
DPR	49.8	**50.0**	**50.0**	**50.0**	49.5	45.4	57.1	47.0	43.9	**65.2**	48.0	**55.9**	51.0	46.8	19.2
SPR	50.1	50.3	50.1	49.9	**50.7**	72.1	74.2	73.6	73.1	**80.6**	56.3	57.0	56.9	56.1	**65.8**
FN+	50.0	50.0	**50.4**	50.0	49.5	57.3	**63.6**	54.5	60.7	60.0	56.2	56.1	54.3	55.5	**80.5**

Table 4.2: Accuracy on RTE with representations generated by encoders of English→{ar,es,zh,de} NMT models. Rows correspond to the training and validation sets and major columns correspond to the test set. The column labeled "USEF" refers to the test accuracies reported in White et al. (2017). The numbers on the top row represents each dataset's majority baseline. Bold numbers indicate the highest performing model for the given dataset.

will begin by discussing results across each of the three datasets.

Paraphrastic entailment (FN+)

The classifiers predict FN+ entailment worse than the majority baseline, and drastically worse than USEF when trained on FN+'s training set. Since FN+ tests paraphrastic inference and NMT models have been shown to be useful to generate sentential paraphrase pairs (Wieting and Gimpel, 2017; Wieting, Mallinson, and Gimpel, 2017), it is surprising that the classifiers using the representations from the NMT encoder perform poorly.

Although the sentences in FN+ are much longer than in the other datasets, sentence length does not seem to be responsible for the poor FN+ results. The classifiers do not noticeably perform better on shorter sentences than longer ones. The average sentence in the FN+ test dataset is 31 words and almost 10% of the test sentences are longer than 50 words. In SPR and DPR, each premise sentence has on average 21 and 15 words respectively and only 1% of sentences in SPR have more than 50

Sentence length	ar	es	zh	de	total
0-10	46.8	63.7	66.0	65.4	526
10-20	49.0	53.3	57.4	56.5	2739
20-30	48.4	54.0	53.2	54.9	4889
30-40	48.4	54.1	51.2	53.9	4057
40-50	47.7	59.0	55.0	58.7	2064
50-60	49.1	56.1	54.5	57.5	877
60-70	46.4	53.6	43.9	44.1	444
70-80	59.9	51.6	43.3	43.3	252

Table 4.3: Accuracies on FN+'s dev set based on sentence length. The first column represents the range of sentences length: first number is inclusive and second is exclusive. The last column represents how many context sentences have lengths that are in the given row's range.

words. No sentences in DPR have more than 50 words.

Table 4.3 reports accuracies for ranges of sentence lengths in FN+'s development set. When trained on sentence representations form an English→Chinese,German NMT encoder, the RTE accuracies steadily decrease. When using English→Arabic, the accuracies stay consistent until sentences have between 70–80 tokens while the results from English→Spanish quickly drops from 0–10 to 10–20 but then stays relatively consistent.

Upon manual inspection, we noticed that in many *not-entailed* examples, swapped paraphrases had different part-of-speech (POS) tags. This begs the question of whether different POS tags for swapped paraphrases affects the accuracies. Using Stanford CoreNLP (Manning et al., 2014), we partition our validation set based on whether the paraphrases share the same POS tag. Table 4.4 reports development set accuracies using classifiers trained on FN+. Classifiers using features from NMT en-

	ar	es	zh	de
Same Tag	52.9	52.6	52.6	50.2
Different Tag	55.8	59.1	53.4	46.0

Table 4.4: Accuracies on FN+'s dev set based on whether the swapped paraphrases share the same POS tag.

coders trained on the three languages from the UN corpus noticeably perform better on cases where paraphrases have different POS tags compared to paraphrases with the same POS tags. These differences might suggest that the recast FN+ might not be an ideal dataset to test how well NMT encoders capture paraphrastic inference. The sentence representations may be impacted more by ungrammaticality caused by different POS tags as opposed to poor paraphrases. We will discuss this issue, and its ramifications, further in Section 5.5.

We also notice that even though the classifiers perform poorly when predicting paraphrastic entailment, they surprisingly outperform USEF by a large margin (around 25–30%) when using a model trained on DPR.[5] This might suggest that an NMT encoder can pick up on how pronouns may be used as a type of lexical paraphrase (Bhagat and Hovy, 2013).

Anaphora entailment (DPR)

The low accuracies for predicting RTE targeting anaphora resolution are similar to White et al. (2017)'s findings. They suggest that the model has difficulty in capturing complex anaphora resolution. By using contrastive evaluation pairs, Bawden et al.

[5] This is seen in the last columns of the top row in Table 4.2.

(2017) recently suggested as well that NMT models are poorly suited for co-reference resolution. Our results are not surprising given that DPR tests whether a model contains common sense knowledge (Rahman and Ng, 2012). In DPR, syntactic cues for co-reference are purposefully balanced out as each pair of pro-nouns appears in at least two context-hypothesis pairs (Table 4.5). This forces the model's decision to be informed by semantics and world knowledge – a model cannot use syntactic cues to help perform anaphora resolution. When released, Rahman and Ng (2012)'s DPR dataset confounded the best co-reference models because "its difficulty stems in part from its reliance on sophisticated knowledge sources." Table 4.5 includes examples that demonstrate how world knowledge is needed to accurately predict these recast RTE sentence-pairs.

Although the poor performance of NMT representations may be explained by a variety of reasons, e.g. training data, architectures, etc., we would still like ideal MT systems to capture the semantics of co-reference, as evidenced in the example in section 4.2.1.

Proto-role entailment (SPR)

When predicting SPR entailments using a classifier trained on SPR data, the models noticeably outperform the majority baseline but are below USEF. Both ours and USEF's accuracies are lower than Teichert et al. (2017)'s best reported numbers. This is not surprising as Teichert et al. (2017)'s model predicts proto-role labels

Chris was running after John, because he stole his watch	
▶ Chris was running after John, because John stole his watch	✓
▶ Chris was running after John, because Chris stole his watch	✗
Chris was running after John, because he wanted to talk to him	
▶ Chris was running after John, because Chris wanted to talk to him	✓
▶ Chris was running after John, because John wanted to talk to him	✗
The plane shot the rocket at the target, then it hit the target	
▶ The plane shot the rocket at the target, then the rocket hit the target	✓
▶ The plane shot the rocket at the target, then the target hit the target	✗
Professors do a lot for students, but they are rarely thankful	
▶ Professors do a lot for students, but students are rarely thankful	✓
▶ Professors do a lot for students, but Professors are rarely thankful	✗
MIT accepted the students, because they had good grades	
▶ MIT accepted the students, because the students had good grades	✓
▶ MIT accepted the students, because MIT had good grades	✗
Obama beat John McCain, because he was the better candidate	
▶ Obama beat John McCain, because Obama was the better candidate	✓
▶ Obama beat John McCain, because John McCain was the better candidate	✗
Obama beat John McCain, because he failed to win the majority of the electoral votes	
▶ Obama beat John McCain, because John McCain failed to win the majority of the electoral votes	✓
▶ Obama beat John McCain, because Obama failed to win the majority of the electoral vote	✗

Table 4.5: Examples from DPR's dev set. The first line in each section is a context and lines with ▶ are corresponding hypotheses. ✓ (✗) in the last column indicates whether the hypothesis is entailed (or not) by the context.

conditioned on observed semantic role labels.

Table 4.6 reports accuracies for each proto-role. Whenever one of the classifiers outperforms the baseline for a proto-role, all the other classifiers do as well. The classifiers outperform the majority baseline for 6 of the reported 16 proto-roles. We observe these 6 properties are more associated with proto-agents than proto-patients.

The larger improvements over the majority baseline for SPR compared to FN+ and DPR is not surprising. Dowty (1991) posited that proto-agent and proto-patient should correlate with English syntactic subject, and object, respectively, and empirically the *necessity of [syntactic] parsing for predicate argument recognition* has been observed in practice (Gildea and Palmer, 2002; Punyakanok, Roth, and Yih, 2008). Further, recent work is suggestive that LSTM-based frameworks implicitly may encode syntax based on certain learning objectives (Linzen, Dupoux, and Goldberg, 2016; Shi, Padhi, and Knight, 2016; Belinkov et al., 2017a). It is unclear whether NMT encoders capture semantic proto-roles specifically or just underlying syntax that affects the proto-roles.

4.2.4 Further Analysis

Before using recast RTE data to evaluate the reasoning capabilities of other NLP models, we explore additional questions about the NMT models we test. We explore whether the target language in translation affects how well the encoders capture different phenomena, how well can these representations be used to predict RTE across multiple genres, as well as how do the results change when using other techniques for creating sentence representations.

Proto-Role	ar	es	zh	de	avg	MAJ
physically existed	70.6	70.8	**77.2**	70.8	72.4†	65.9
sentient	78.5	**82.2**	80.5	81.7	80.7†	75.5
aware	75.9	**77.0**	76.6	76.7	76.6†	60.9
volitional	74.3	**76.8**	74.7	73.7	74.9†	64.5
existed before	68.4	**70.5**	66.5	68.4	68.5†	64.8
caused	69.4	**74.1**	72.2	72.7	72.1†	63.4
changed	64.2	62.4	63.8	62.0	63.1	**65.1**
location	91.1	90.1	90.4	90.2	90.4	**91.7**
moved	90.6	88.8	90.1	90.3	89.9	**93.3**
used in	34.9	38.1	31.8	34.2	34.7	**55.2**
existed after	62.7	69.0	65.6	65.2	65.7	**69.7**
chang. state	61.8	60.7	60.9	60.7	61.0	**65.2**
chang. possession	89.6	88.6	89.9	88.3	89.1	**93.9**
stationary during	86.3	84.4	90.5	86.0	86.8	**96.3**
physical contact	85.0	82.0	84.5	84.4	84.0	**85.8**
existed during	59.3	71.8	60.8	64.4	64.1	**84.7**

Table 4.6: Accuracies on the SPR test set broken down by each proto-role. "avg" represents the score for the proto-role averaged across target languages. Bold and † respectively indicate the best results for each proto-role and whether all of our classifiers outperformed the proto-role's majority baseline.

NMT target language

Our experiments show differences based on which target language was used to train the NMT encoder, in capturing semantic proto-roles and paraphrastic inference. In Table 4.2, we notice a large improvement using sentence representations from an NMT encoder that was trained on en-es parallel text. The improvements are most profound when a classifier trained on DPR data predicts entailment focused on semantic proto-roles or paraphrastic inference. We also note that using the NMT encoder trained on en-es parallel text results in the highest results in 5 of the 6 proto-roles in the top portion of Table 4.6. Very recent work exploring how well syntax is captured in NMT models also explores the effect of the choice of target language (Chang and Rafferty, 2020). They found that the choice of target language did not noticeably alter how NMT encoder representations encode source syntax.

RTE across multiple domains

Though our main focus is exploring what NMT encoders learn about distinct semantic phenomena, we would like to know how useful NMT models are for general RTE across multiple domains. Therefore, we also evaluate the sentence representations with Multi-NLI. As indicated by Table 4.7, the representations perform noticeably better than a majority baseline. However, our results are not competitive with state-of-the-art systems trained specifically for Multi-NLI (Nangia et al., 2017).

	ar	es	zh	de	MAJ
MNLI-1	45.9	45.7	46.6	48.0	35.6
MNLI-2	46.6	46.7	48.2	48.9	36.5

Table 4.7: Accuracies for MNLI test sets. MNLI-1 refers to the matched case and MNLI-2 is the mismatched.

Evaluating Different Sentence Representations Techniques

In the experiments discussed so far, we used a simple sentence representation extracted from the Bi-LSTM encoders, the first and last hidden states of the forward and backward encoders. We concatenated them for both the context and the hypothesis and fed to a linear classifier. Here we compare the results of `InferSent` (Conneau et al., 2017), a more involved representation that was found to provide a good sentence representation based on RTE data. Specifically, we concatenate the forward and backward encodings for each sentence, and max-pool over the length of the sentence, resulting in $\mathbf{v}$ and $\mathbf{u}$ (in $\mathbb{R}^{2d}$) for the context and hypothesis. The `InferSent` representation is defined by

$$(\mathbf{u}, \mathbf{v}, |\mathbf{u} - \mathbf{v}|, \mathbf{u} * \mathbf{v}) \in \mathbb{R}^{8d}$$

where the product and subtraction are carried element-wise and commas denote vector-concatenation.

The pair representation is fed into a multi-layered perceptron (MLP) with one hidden layer and a ReLU non-linearity. We set the hidden layer size to 500 dimensions,

		FN+	DPR	SPRL	MNLI-1	MNLI-2
	en-ar	56.2	49.8	72.1	45.9	46.6
NMT	en-es	56.1	50.0	74.2	45.7	46.7
Concat	en-zh	54.3	50.0	73.1	46.6	48.2
	en-de	55.5	50.0	73.1	48.0	48.9
	en-ar	57.9	50.0	73.6	40.1	41.8
NMT	en-es	58.0	50.0	72.7	44.9	40.8
`InferSent`	en-zh	57.8	49.8	72.4	43.7	42.1
	en-de	58.3	50.1	73.7	41.3	41.1
Majority		57.5	50.0	65.4	35.6	36.5
SOTA		80.5	49.5	80.6	81.10	83.21

Table 4.8: RTE results on fine-grained semantic phenomena. FN+ = paraphrases; DPR = pronoun resolution; SPRL = proto-roles. NMT representations are combined with either a simple concatenation (results copied from Table 4.1) or the `InferSent` representation. State-of-the-art (SOTA) for the recast datasets is from White et al. (2017). The right half report results on language inference on MultiNLI (Williams, Nangia, and Bowman, 2017), matched/mismatched scenario (MNLI1/2).

similarly to Conneau et al. (2017). The soft-max layer maps onto the number of labels, which is 2 for the recast datasets and 3 for MNLI.

Table 4.8 shows the results of the classifier trained on NMT representations with the InferSent architecture. Here, the representations from NMT encoders trained on the English-German parallel corpus slightly and consistently outperform the other encoders. Since this data used a different corpus compared to the other language pairs, we cannot determine whether the improved results are due to the different target side language or corpus. The main difference with respects to the simpler sentence representation (Concat) is improved results on FN+. It is interesting to note that when using the sentence representations from NMT encoders, concatenating the sentence vectors outperformed the `InferSent` method on Multi-NLI.

4.2.5 Related Work

Prior work has focused on the relationship between semantics and machine translation. MEANT and its extension XMEANT evaluate MT systems based on semantics (Lo and Wu, 2011a; Lo et al., 2014). Others have focused on incorporating semantics directly in MT. Chan, Ng, and Chiang (2007) use word sense disambiguation to help statistical MT, Gao and Vogel (2011) add semantic-roles to improve phrase-based MT, and Carpuat, Vyas, and Niu (2017) demonstrate how filtering parallel sentences that are not parallel in meaning improves translation. Recent work explores how representations learned by NMT systems can improve semantic tasks. McCann et al. (2017) show improvements in many tasks by using contextualized word vectors extracted from a LSTM encoder trained for MT. Their goal is to use NMT to improve other tasks while we focus on using RTE to determine what NMT models learn about different semantic phenomena.

Researchers have explored what NMT models learn about other linguistic phenomena, such as morphology (Dalvi et al., 2017; Belinkov et al., 2017b), syntax (Shi, Padhi, and Knight, 2016), and lexical semantics (Belinkov et al., 2017a), including word senses (Marvin and Koehn, 2018; Liu, Lu, and Neubig, 2018).

4.3 Evaluating Encoders Trained For Other Tasks

As part of the 2018 JSALT Workshop focused on general sentence representation learning, we evaluated how well encoders trained for different NLP tasks capture some semantic phenomenon included in the DNC. The rest of this chapter discusses those results. These results have been presented at the end of workshop presentation in Summer 2018.

4.3.1 Tasks

Here, we explore whether pre-training encoders for different NLP tasks can help an RTE model trained on MNLI capture these different phenomena. The tasks include language modeling, syntactic parsing, discourse marking, and image-caption matching. We will discuss each of these in more detail. All of these encoders were trained by colleagues on the JSALT Workshop team.

Language Modeling

In NLP, a language model (LM) is a model that predicts the probability of a sentence or the probability of a word conditioned on previous words in a sentence. A unigram language model predicts the probability of a single token, a binary language model predicts the probability of a single token conditioned on the previous token, and an n-gram language model predicts the probability of a word conditioned on the previous $n-1$ words. The probabilities are learned in an unsupervised fashion from large amounts of text. LM probabilities are often used as strong baseline features to train Machine Learning models for different NLP tasks. In these experiments, we use a LM trained on the Billion Word Benchmark (BWB) (Chelba et al., 2013). In particular, separate forward and backward two-layer 1024-dimension encoders are trained and the hidden states are concatenated as token representations.

CCG Supertagging

The second task we explore is Combinatory Categorial Grammar (CCG) supertagging. CCG parsing is a "syntactic grammar formalism that pairs words with lexical categories, and a set of combinatory rules" (Clark, 2002). The encoder we explore is trained to predict each word's CCG supertag, a part-of-speech-like that includes broad syntactic context (Bangalore and Joshi, 1999). Data from Hockenmaier and Steedman (2007)'s CCGBank is used to train this model.

Discourse Marking

The third task we explore training an encoder on is Discourse Marking. This tais to predict the discourse marker Given two sentences in our curated corpus (which may have been full sentences in the original text or may have been subclauses), the model must predict which discourse marker was used by an author to connect two given texts (Nie, Bennett, and Goodman, 2019). The model is trained “on a dataset created from WikiText-103 following Nie, Bennett, and Goodman (2019)’s protocol, which involves extracting pairs of clauses with a specific dependency relation” (Kim et al., 2019).

Image-Caption Matching

The fourth task we explore training an encoder on is matching captions with images. Here, the model is trained to minimize the distance between features of an image and the sentence representations of its corresponding caption. The model is trained on the MSCOCO dataset (Lin et al., 2014) and follows the training setup of Kiela et al. (2018).

4.3.2 Experiments

Pre-training

We train one neural model for each of these tasks. Unlike the experiments discussed earlier in Section 4.2.2, we use ELMo word representations that have been trained using a character-level convolutional neural network (Peters et al., 2018). These word representations are passed to a 2-layered 1024 dimensional BiLSTM. A classifier receives as input the top-layer hidden states of BiLSTM and the original representation of each word (via a skip-connection).

The models are optimized with AMSGrad (Reddi et al., 2018) and a learning rate of 1e-4. The models are trained for at most 20 epochs. The learning rate is multiplied by 0.5 whenever validation performance does not improve after 4 epochs. If the learning rate falls below 1e-6, training is stopped.

RTE Classifier

Following the framework as depicted in Figure 4.1, we keep these pre-trained endocers fixed and we use them as as feature extractors to train a classifier. The sentences are extracted using max-pooling and are combined using Mou et al. (2016)'s popular heuristic matching technique. We use a MLP classifier with one hidden layer of 512 dimensions. When training the classifier, we use a dropout of 0.2, an initial learning rate of 0.0001, a learning rate decay factor of 0.5, and a minimum learning rate of $1e-06$. We do not train the RTE classifiers on the training sets from the DNC. Instead, following Kim et al. (2019), we train the classifiers on MNLI. In turn, here we use the DNC datasets solely as an evaluation. These experiments were implemented using the `jiant` toolkit (Wang et al., 2019b).[6]

4.3.3 Results

Figure 4.3 reports results across four of the DNC datasets: NER, Factuality, VerbNet, and Relation Extraction. In addition to the encoders discussed, we include three baselines that help contextualize these results. The baselines are the majority base-

[6] `https://jiant.info/`

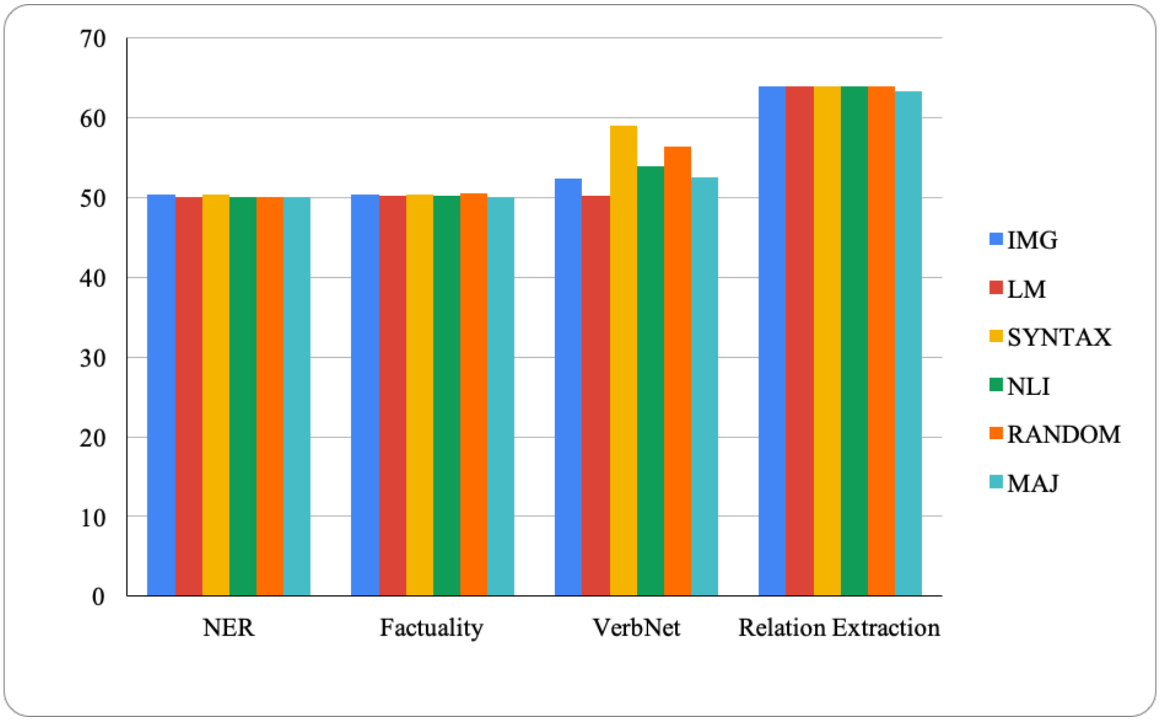


Figure 4.3: Results of the encoders that were trained on the different NLP tasks.

line (MAJ), a randomly initialized representation, and not pre-training the encoder of any of the discussed tasks (NLI). When testing the model on NER, Factuality, and Relation Extraction, we do not see a major difference in the results when pre-trained the model on any of the tasks. For NER, we see small improvements when the encoder is pre-trained on the image-caption matching (IMG) or CCG parsing (Syntax) tasks. These results are similar to `InferSent` model trained on just MNLI from Table 3.3. For the Factuality test set, it seems that pre-training a model on any of these tasks is worse than using random sentence representations. For the relation extraction test, we see that pre-training or just training on MNLI, performs slightly better than the majority baseline.

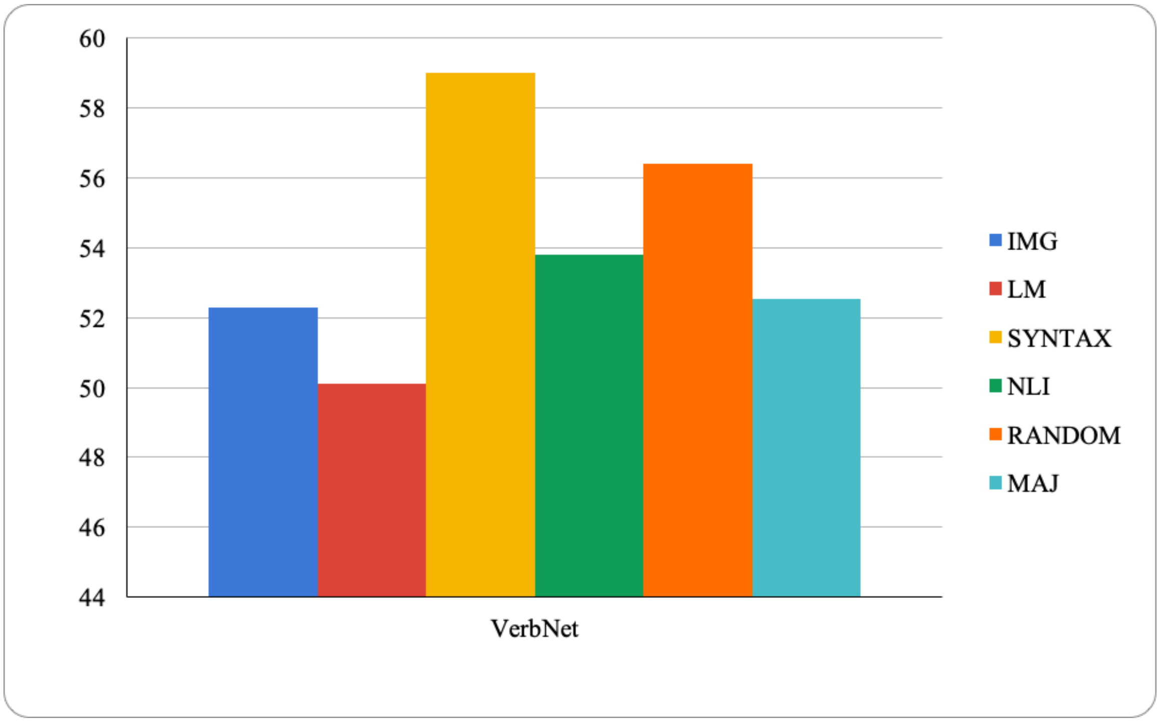


Figure 4.4: Results of the encoders that were trained on the tasks described here These results are tested on VerbNet.

The most interesting results are on the VerbNet test set (Figure 4.4). When the encoder is pre-trained on language modeling (LM) or image-caption matching, the model performs worse that the majority baseline. Pre-training the model on CCG parsing is the only pre-training task where the model outperforms a random sentence representation. This result might corroborate the commonly held belief that the semantic representations in PropBank, which are used in Verbnet, convey shallow semantics that are deeply connected to syntax (Teichert et al., 2017).

4.4 Discussion

In this chapter, we presented a general purpose framework for using RTE to evaluate the reasoning capabilities of NLP models. Researchers suggest that NMT models learn sentence representations that capture meaning. We delved deeply in discovering how well a neural machine translation encoder captures different semantic phenomena that are important for translation. Our experiments suggest that NMT encoders might learn the most about semantic proto-roles but do not focus on anaphora resolution. We conclude by suggesting that target-side language affects how well an NMT encoder captures these semantic phenomena.

The experiments focused on paraphrastic inference in MT might suggest that NMT models may poorly capture paraphrastic inference. However, work in back-translation, i.e. translating a translated text in a target language back into the source language, indicate that NMT systems indeed capture paraphrases. Furthermore, resources of large-scale sentence level paraphrases, like ParaBank (Hu et al., 2019) and ParaNMT-50M (Wieting and Gimpel, 2018), were developed using machine translation resources and methods. As mentioned in Section 4.2.3, and as we will discuss further in more detail later in the thesis, issues in the RTE dataset focused on paraphrastic inference might limit its utility. Therefore, these experiments do not contradict recent findings in the community that representations learned by NMT systems indeed capture paraphrases.

Additionally, we used this framework to survey encoders trained for different NLP tasks as well. We noticed that pre-training on syntactic parsing had the most benefit when evaluating the model on the recast RTE dataset focused on a shallow semantic representation.

Chapter 5

Hypothesis-only Biases in Recognizing Textual Entailment

When RTE datasets are constructed to facilitate the training and evaluation of natural language understanding (NLU) systems, it is tempting to claim that systems achieving high accuracy on such datasets have successfully "understood" natural language or at least a logical relationship between a premise and hypothesis. In this chapter, we explore whether issues or biases in datasets enable simple methods to achieve decent results without actually performing the reasoning supposedly required for these tasks. Specifically, we demonstrate that RTE datasets contain statistical irregularities that allow hypothesis-only models to outperform the datasets specific prior.

We do not attempt to prescribe the sufficient conditions of claiming that systems understand natural language. Rather, we argue for an obvious *necessary*, or at least *desired* condition: *that interesting textual entailment should depend on both premise and hypothesis.* In other words, a baseline system with access only to hypotheses (5.1b) can be said to perform RTE only in the sense that it is understanding language based on prior background knowledge. If this background knowledge is about the world, this may be justifiable as an aspect of natural language understanding, if not in keeping with the spirit of RTE. But if the “background knowledge" consists of learned statistical irregularities in the data, this may not be ideal. In such a case, the data constructed in a particular dataset may limit one’s ability to use the data as a test to evaluate the reasoning capabilities of a NLP model.

We present the results of a hypothesis-only baseline across eighteen RTE datasets and advocate for its inclusion in future dataset reports. We find that this baseline can perform above the majority-class prior across most of the eighteen examined datasets. We examine whether: (1) hypotheses contain statistical irregularities within each entailment class that are “giveaways” to a well-trained hypothesis-only model, (2) the way in which an RTE dataset is constructed is related to how prone it is to this particular weakness, and (3) the majority baselines might not be as indicative of “the difficulty of the task” (Bowman et al., 2015) as previously thought.

We will discuss what this means for RTE datasets and lessons that might be important for when creating new RTE datasets. This chapter is based on Poliak et al. (2018a) and Poliak et al. (2018b). The second paper received a best paper award

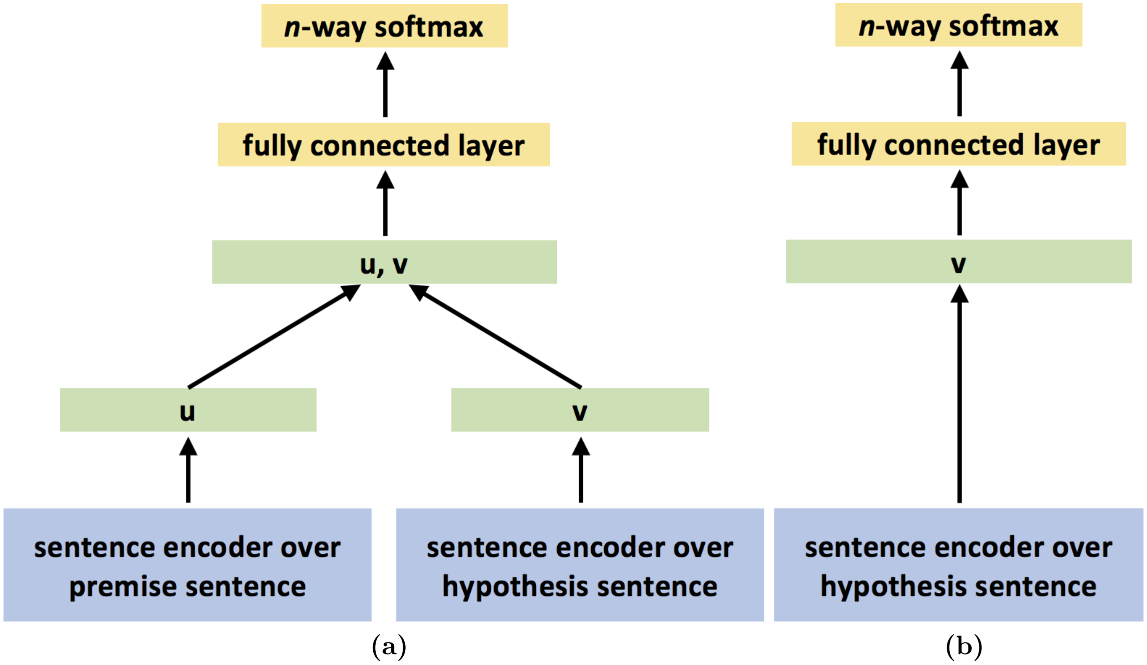

n-way softmax
fully connected layer
u, v
u
v
sentence encoder over premise sentence
sentence encoder over hypothesis sentence
(a)
n-way softmax
fully connected layer
v
sentence encoder over hypothesis sentence
(b)

Figure 5.1: (5.1a) shows a typical RTE model that encodes the premise and hypothesis sentences into a vector space to classify the sentence pair. (5.1b) shows our hypothesis-only baseline method that ignores the premise and only encodes the hypothesis sentence.

at StarSem 2018.

Related Studies

We are not the first to consider the inherent difficulty of RTE datasets. For example, MacCartney (2009) used a simple bag-of-words model to evaluate early iterations of Recognizing Textual Entailment (RTE) challenge sets.[1] Concerns have been raised previously about the hypotheses in the Stanford Natural Language Inference (SNLI) dataset specifically, such as by Rudinger, May, and Van Durme (2017) and in un-

[1] MacCartney (2009), Ch. 2.2: *"the RTE1 test suite is the hardest, while the RTE2 test suite is roughly 4% easier, and the RTE3 test suite is roughly 9% easier."*

published work.[2] Here, we survey of large number of existing RTE datasets under the lens of a hypothesis-only model.[3] Concurrently, Tsuchiya (2018) and Gururangan et al. (2018) similarly trained an NLI classifier with access limited to hypotheses and discovered similar results on three of the eighteen datasets that we study.

5.1 Motivation

Our approach is inspired by recent studies that show how biases in an NLU dataset allow models to perform well on the task without understanding the meaning of the text. In the Story Cloze task (Mostafazadeh et al., 2016; Mostafazadeh et al., 2017). a model is presented with a short four-sentence narrative and asked to complete it by choosing one of two suggested concluding sentences. While the task is presented as a new common-sense reasoning framework, Schwartz et al. (2017a) performed alarmingly well by ignoring the narrative and training a linear classifier with features related to the writing style of the two potential endings, rather than their content. It has also been shown that features focusing on sentence length, sentiment, and negation are sufficient for achieving high accuracy on this dataset (Schwartz et al., 2017b; Cai, Tu, and Gimpel, 2017; Bugert et al., 2017).

As discussed throughout this thesis, RTE is often viewed as an integral part of NLU. Condoravdi et al. (2003) argue that it is a necessary metric for evaluating

[2] A course project constituting independent discovery of our observations on SNLI: https://leonidk.com/pdfs/cs224u.pdf

[3] Our code and data can be found at https://github.com/azpoliak/hypothesis-only-NLI.

an NLU system as it forces a model to perform many distinct types of reasoning, Goldberg (2017) suggests that “solving [RTEssss] perfectly entails human level understanding of language”, and Nangia et al. (2017) argue that “in order for a system to perform well at natural language inference, it needs to handle nearly the full complexity of natural language understanding.” Thus, if biases in RTE datasets, especially those that do not reflect commonsense knowledge, allow models to achieve high levels of performance without needing to reason about hypotheses based on corresponding contexts, our current datasets may fall short of the broad goals of RTE.

5.2 Methodology

We modify Conneau et al. (2017)’s `InferSent` method to train a neural model to classify just the hypotheses. We choose `InferSent` because it performed competitively with the best-scoring systems on the Stanford Natural Language Inference (SNLI) dataset, while being representative of the types of neural architectures commonly used for RTE tasks. `InferSent` uses a BiLSTM encoder, and constructs a sentence representation by max-pooling over its hidden states. This sentence representation of a hypothesis is used as input to a MLP classifier to predict the RTE label.

Creation Protocol	Dataset	Size	Classes	Example Hypothesis
Recast	DPR	3K	2	*People raise dogs because dogs are afraid of thieves*
	SPR	150K	2	*The judge was aware of the dismissing*
	FN+	150K	2	*the irish are actually principling to come home*
	NER	325K	2	*Hong Kong is a location*
	EF	48K	2	*The proposing did not happen*
	RE	25K	2	*Obama served as President of the United States*
	Puns	17K	2	*Natalee heard a pun*
	Sentiment	6K	2	*Yaseen liked the restaurant*
	VC	125K	2	*Something bad happened*
	MV	11K	2	*that thing might or might not have happened*
	VN	2K	2	*The package moved*
Judged	ADD-1	5K	2	*A small child staring at a young horse and a pony*
	SCITAIL	25K	2	*Humans typically have 23 pairs of chromosomes*
	SICK	10K	3	*Pasta is being put into a dish by a woman*
	MPE	10K	3	*A man smoking a cigarette*
	JOCI	30K	3	*The flooring is a horizontal surface*
Elicited	SNLI	550K	3	*An animal is jumping to catch an object*
	MNLI	425K	3	*Kyoto has a kabuki troupe and so does Osaka*

Table 5.1: Basic statistics about the RTE datasets we consider. 'Size' refers to the total number of labeled premise-hypothesis pairs in each dataset (for datasets with $> 100K$ examples, numbers are rounded down to the nearest $25K$). The 'Creation Protocol' column indicates how the dataset was created. The 'Class' column reports the number of class labels/tags. The last column shows an example hypothesis from each dataset.

5.3 Datasets

We use a hypothesis-only model to study eighteen RTE datasets. We categorize them into three distinct groups based on the methods by which they were constructed. Table 5.1 summarizes the different RTE datasets that our investigation considers.

Human Elicited

In cases where humans were given a context and asked to generate a corresponding hypothesis and label, we consider these datasets to be **elicited**. Although we consider only two such datasets, they are the largest datasets included in our study and are currently popular amongst researchers. The elicited RTE datasets we look at are:

- **Stanford Natural Language Inference (SNLI)** To create SNLI, Bowman et al. (2015) showed crowdsource workers a premise sentence (sourced from Flickr image captions (Young et al., 2014b)), and asked them to generate a corresponding hypothesis sentence for each of the three labels (ENTAILMENT, NEUTRAL, CONTRADICTION). SNLI is known to contain stereotypical biases based on gender, race, and ethnic stereotypes (Rudinger, May, and Van Durme, 2017). Furthermore, Zhang et al. (2017) commented that this "elicitation protocol can lead to biased responses unlikely to contain a wide range of possible common-sense inferences."

- **Multi-NLI** Multi-NLI is a recent expansion of SNLI aimed to add greater diversity to the existing dataset (Williams, Nangia, and Bowman, 2017). Premises in Multi-NLI can originate from fictional stories, personal letters, telephone speech, and a 9/11 report.

Human Judged

Alternatively, if hypotheses and premises were automatically paired but *labeled* by a human, we consider the dataset to be **judged**. Our human-judged data sets are:

- **Sentences Involving Compositional Knowledge (SICK)** To evaluate how well compositional distributional semantic models handle "challenging phenomena", Marelli et al. (2014) introduced SICK, which used rules to expand or normalize existing premises to create more difficult examples. Workers were asked to label the relatedness of these resulting pairs, and these labels were then converted into the same three-way label space as SNLI and Multi-NLI.

- **Add-one RTE** This mixed-genre dataset tests whether RTE systems can understand adjective-noun compounds (Pavlick and Callison-Burch, 2016). Premise sentences were extracted from Annotated Gigaword (Napoles, Gormley, and Van Durme, 2012), image captions (Young et al., 2014a), the Internet Argument Corpus (Walker et al., 2012), and fictional stories from the GutenTag dataset (Mac Kim and Cassidy, 2015). To create hypotheses, adjectives were removed or inserted before nouns in a premise, and crowd-sourced workers were asked to provide reliable labels (ENTAILED, NOT-ENTAILED).

- **SciTail** Recently released, SciTail is an RTE dataset created from 4th grade science questions and multiple-choice answers (Khot, Sabharwal, and Clark, 2018). Hypotheses are assertions converted from question-answer pairs found in SciQ (Welbl, Liu, and Gardner, 2017).[4] Hypotheses are automatically paired with premise sentences from domain specific texts (Clark et al., 2016), and labeled (ENTAILMENT, NEUTRAL) by crowdsource workers. Notably, the construction method allows for the same sentence to appear as a hypothesis for more than one premise.

[4] `http://allenai.org/data/science-exam-questions.html`

- **Multiple Premise Entailment (MPE)** Unlike the other datasets we consider, the premises in MPE (Lai, Bisk, and Hockenmaier, 2017) are not single sentences, but four different captions that describe the same image in the FLICKR30K dataset (Plummer et al., 2015). Hypotheses were generated by simplifying either a fifth caption that describes the same image or a caption corresponding to a different image, and given the standard 3-way tags. Each hypothesis has at most a 50% overlap with the words in its corresponding premise. Since the hypotheses are still just one sentence, our hypothesis-only baseline can easily be applied to MPE.

- **Johns Hopkins Ordinal Common-Sense Inference (JOCI)** JOCI labels context-hypothesis instances on an ordinal scale from *impossible* (1) to *very likely* (5) (Zhang et al., 2017). In JOCI, context (premise) sentences were taken from existing NLU datasets: SNLI, ROC Stories (Mostafazadeh et al., 2016), and COPA (Roemmele, Bejan, and Gordon, 2011). Hypotheses were created automatically by systems trained to generate entailed facts from a premise.[5] Crowd-sourced workers labeled the likelihood of the hypothesis following from the premise on an *ordinal scale*. We convert these into 3-way RTE tags where 1 maps to CONTRADICTION, 2-4 maps to NEUTRAL, and 5 maps to ENTAILMENT. Converting the annotations into a 3-way classification problem allows us to limit the range of the number of RTE label classes in our investigation.

[5] We only consider the hypotheses generated by either a seq2seq model or from external world knowledge.

Recast

As a reminder from Chapter 3, recast RTE datasets are automatically, pragmatically generated from existing datasets for other NLP tasks. RTE sentence pairs are constructed and labeled with minimal human intervention. In addition to the datasets in the DNC, we consider the three datasets in White et al. (2017)'s pre-cursor to the DNC. The three other recast datasets we consider are:

- **Semantic Proto-Roles (SPR)** Inspired by Dowty (1991)'s thematic role theory, Reisinger et al. (2015) introduced the Semantic Proto-Role (SPR) labeling task, which can be viewed as decomposing semantic roles into finer-grained properties, such as whether a predicate's argument was likely *aware* of the given predicated situation. 2-way labeled RTE sentence pairs were generated from SPR annotations by creating general templates.

- **Definite Pronoun Resolution (DPR)** The DPR dataset targets an RTE model's ability to perform anaphora resolution (Rahman and Ng, 2012). In the original dataset, sentences contain two entities and one pronoun, and the task is to link the pronoun to its referent. In the recast version, the premises are the original sentences and the hypotheses are the same sentences with the pronoun replaced with its correct (ENTAILED) and incorrect (NOT-ENTAILED) referent. For example, *People raise dogs because they are obedient* and *People raise dogs because dogs are obedient* is such a context-hypothesis pair. We note that this mechanism would appear to maximally benefit a hypothesis-only approach, as the hypothesis semantically subsumes the context.

- **FrameNet Plus (FN+)** Using paraphrases from PPDB (Ganitkevitch, Van Durme, and Callison-Burch, 2013), Rastogi and Van Durme (2014) automatically replaced words with their paraphrases. Subsequently, Pavlick et al. (2015) asked crowd-source workers to judge how well a sentence with a paraphrase preserved the original sentence's meanings. In this RTE dataset that targets a model's ability to perform paraphrastic inference, premise sentences are the original sentences, the hypotheses are the edited versions, and the crowd-source judgments are converted to 2-way RTE-labels. For not-entailed examples, White et al. (2017) replaced a single token in a context sentence with a word that crowd-source workers labeled as not being a paraphrase of the token in the given context. In turn, we might suppose that positive entailments (2) are keeping in the spirit of RTE, but not-entailed examples might not because there are adequacy (3) and fluency (4) issues.[6]

(1) . That is the way the system works

(2) . That is the way the framework works

(3) ¸ That is the road the system works

[6] In these examples, (1) is the corresponding context.

(4) †. That is the way the system creations

Experimental details

We preprocess each recast dataset using the NLTK tokenizer (Bird and Loper, 2004). Following Conneau et al. (2017), we map the resulting tokens to 300-dimensional GloVe vectors (Pennington, Socher, and Manning, 2014) trained on 840 billion tokens from the Common Crawl, using the GloVe OOV vector for unknown words. For all experiments except for JOCI, we use each RTE dataset's standard train, dev, and test splits.[7] We optimize via SGD, with an initial learning rate of 0.1, and decay rate of 0.99. We allow at most 20 epochs of training with optional early stopping according to the following policy: when the accuracy on the development set decreases, we divide the learning rate by 5 and stop training when the learning rate is less than 10^{-5}.

5.4 Results

Our goal is to determine whether a hypothesis-only model outperforms the majority baseline and investigate what may cause significant gains. In such cases a hypothesis-only model should be used as a stronger baseline instead of the majority class baseline when evaluating an RTE model. Table 5.2 compares the hypothesis-only

[7] JOCI was not released with such splits so we randomly split the dataset into such a partition with 80:10:10 ratios.

	DEV				**TEST**					
Dataset	Hyp-Only	MAJ	$\lvert\Delta\rvert$	$\Delta\%$	Hyp-Only	MAJ	$\lvert\Delta\rvert$	$\Delta\%$	Baseline	SOTA
					Recast					
DPR	50.21	50.21	0.00	0.00	49.95	49.95	0.00	0.00	49.5	49.5
SPR	86.21	65.27	+20.94	+32.08	86.57	65.44	+21.13	+32.29	80.6	80.6
FN+	62.43	56.79	+5.64	+9.31	61.11	57.48	+3.63	+6.32	80.5	80.5
NER	93.50	50.00	+43.5	+87.0	91.48	50.00	+41.48	+82.96	92.50	92.50
EF	73.84	50.00	+23.84	+47.68	69.14	50.00	+19.14	+38.28	83.07	83.86
RE	66.06	62.85	+3.21	+5.11	64.78	59.53	+5.25	+8.82	61.89	76.08
Puns	60.39	50.00	+10.39	+20.78	60.36	50.00	+10.36	+20.72	60.36	93.17
Sentiment	50.00	50.00	0.0	-0.0	50.00	50.00	0.00	0.00	50.00	83.50
VC	77.82	50.00	+27.82	+55.64	76.82	50.00	+26.82	+53.64	88.60	89.00
MV	79.86	66.67	+13.19	+19.78	77.83	66.67	+11.16	+16.74	85.96	85.96
VN	59.74	59.74	0.0	-0.0	46.34	53.66	-7.32	-13.64	46.34	78.05
					Human Judged					
ADD-1	75.10	75.10	0.00	0.00	85.27	85.27	0.00	0.00	92.2	92.2
SciTail	66.56	50.38	+16.18	+32.12	66.56	60.04	+6.52	+10.86	70.6	77.3
SICK	56.76	56.76	0.00	0.00	56.87	56.87	0.00	0.00	56.87	84.6
MPE	40.20	40.20	0.00	0.00	42.40	42.40	0.00	0.00	41.7	56.3
JOCI	61.64	57.74	+3.90	+6.75	62.61	57.26	+5.35	+9.34	–	–
					Human Elicited					
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	–	35.6	– –		72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

Table 5.2: RTE accuracies on each dataset. Columns 'Hyp-Only' and 'MAJ' indicates the accuracy of the hypothesis-only model and the majority baseline. $|\Delta|$ and $\Delta\%$ indicate the absolute difference in percentage points and the percentage increase between the Hyp-Only and MAJ. Blue numbers indicate that the hypothesis-model outperforms MAJ. In the right-most section, 'Baseline' indicates the original baseline on the test when each dataset was released and 'SOTA' indicates current state-of-the-art results. MNLI-1 is the matched version and MNLI-2 is the mismatched for MNLI. The names of datasets are italicized if containing $\leq 10K$ labeled examples.

model's accuracy with the majority baseline on each dataset's dev and test set.[8]

Across most of the eighteen datasets, our hypothesis-only model ***significantly outperforms*** the majority-baseline, even outperforming the best reported results on one dataset, recast SPR. This indicates that there exists a significant degree of exploitable signal that may help RTE models perform well on their corresponding test set without considering RTE contexts. The largest relative gains are on human-elicited models where the hypothesis-only model more than doubles the majority baseline.

However, there are no obvious unifying trends across these datasets: Among the judged and recast datasets, where humans do not generate the RTE hypothesis, we observe lower performance margins between majority and hypothesis-only models compared to the human elicited data sets. In six of the eight DNC1.0 datasets that we test here, the hypothesis-only model outperforms MAJ. Interestingly, the hypothesis-only model outperforms `InferSent` on the recast relation extraction dataset.[9] The high hypothesis-only accuracy on the recast NER dataset is alarming and we will discuss this more in Section 5.6 when dealing with the recommendations for creating new NLI datasets based on these results.

In general the performances of these hypothesis-only models are noticeably much larger that MAJ on SNLI and Multi-NLI compared to the models' performances on

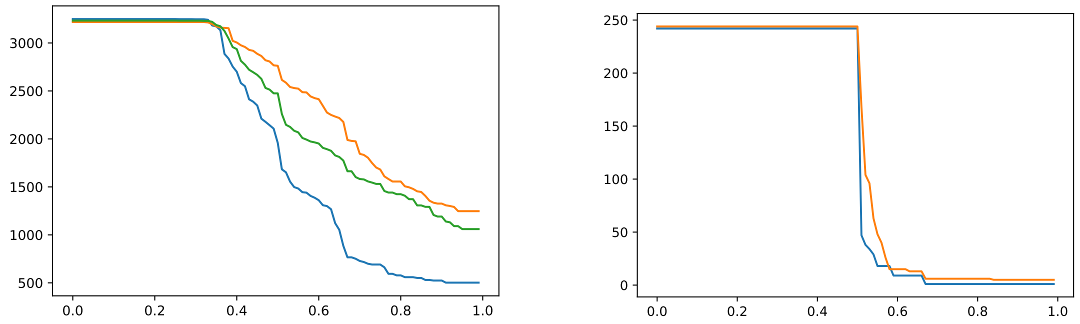
[8] We only report results on the Multi-NLI development set since the test labels are only accessible on Kaggle.

[9] As reported in Table 3.3, `InferSent` achieves a 61.89 accuracy and its hypothesis-only version achieves a 64.78 accuracy.

the other datasets. The drop between the models' performance on Mulit-NLI (56% increase) compared to SNLI (over 100% increase) suggests that by including multiple genres, an RTE dataset may contain less biases. However, adding additional genres might not be enough to mitigate biases as the hypothesis-only model still drastically outperforms the majority-baseline. Therefore, we believe that models tested on SNLI and Multi-NLI should include a baseline version of the model that only accesses hypotheses.

We notice that on the smaller RTE datasets under consideration, i.e. those with at most $10K$ examples, the hypothesis-only model always predicts the majority class label from the test set. On three of the five human judged datasets (ADD-1, SICK, and MPE), the hypothesis-only model defaults to labeling each instance with the majority class tag. We find the same behavior in the smaller recast datasets (DPR, Sentiment, and VN). This might be caused by the small size of each of these datasets. Data intensive neural networks seem to predict the majority class label for each test instance when trained on a small amount of data.

Across both these categories of dataset construction (recast and human-judged), we find smaller relative improvements than on SNLI and Multi-NLI. These results suggest the existence of exploitable signal in the datasets that is unrelated to contexts in RTE. Our focus now shifts to identifying precisely what these signals might be and understanding why they may appear in RTE hypotheses.



(a) SNLI

(b) DPR

Figure 5.2: Plots showing the number of sentences per each label (Y-axis) that contain at least one word w such that $p(l|w) >= x$ for at least one label l. Colors indicate different labels. Intuitively, for a sliding definition of what value of $p(l|w)$ might constitute a "give-away" the Y-axis shows the proportion of sentences that can be trivially answered for each class.

5.5 Exploring Statistical Irregularities

We are interested in determining what characteristics in the datasets may be responsible for the hypothesis-only model often outperforming the majority baseline. Here, we investigate the importance of specific words, grammaticality, and lexical semantics.

Can Labels be Inferred from Single Words?

Since words in hypotheses have a distribution over the class of labels, we can determine the conditional probability of a label l given the word w by

$$p(l|w) = \frac{count(w,l)}{count(w)} \tag{5.1}$$

If $p(l|w)$ is highly skewed across labels, there exists the potential for a predictive bias. Consequently, such words may be “give-aways” that allow the hypothesis model to correctly predict an RTE label without considering the context. Consequently, if hypotheses in an RTE dataset contain many such words, then the dataset may contain exploitable biases that do not require a RTE model to perform general NLU.

If a single occurrence of a highly label-specific word would allow a sentence to be deterministically classified, how many sentences in a dataset are prone to being trivially labeled? The plots in Figure 5.2 answer this question for SNLI and DPR. The Y-value where $X = 1.0$ captures the number of such sentences. Other values of $X < 1.0$ can also have strong correlative effects, but a priori the relationship between the value of X and the coverage of trivially answerable instances in the data is unclear. We illustrate this relationship for varying values of $p(l|w)$. When $X = 0$, all words are considered highly-correlated with a specific class label, and thus the entire data set would be treated as trivially answerable.

In DPR, which has two class labels, the uncertainty of a label is highest when $p(l|w) = 0.5$. The sharp drop as X deviates from this value indicates a weaker effect, where there are proportionally fewer sentences which contain highly label-specific words with respect to SNLI. As SNLI uses 3-way classification we see a gradual decline from 0.33.

				SNLI				
Word	**Score**	**Freq**	**Word**	**Score**	**Freq**	**Word**	**Score**	**Freq**
instrument	0.90	20	tall	0.93	44	sleeping	0.88	108
touching	0.83	12	competition	0.88	24	driving	0.81	53
least	0.90	10	because	0.83	23	Nobody	1.00	52
Humans	0.88	8	birthday	0.85	20	alone	0.90	50
transportation	0.86	7	mom	0.82	17	cat	0.84	49
speaking	0.86	7	win	0.88	16	asleep	0.91	43
screen	0.86	7	got	0.81	16	no	0.84	31
arts	0.86	7	trip	0.93	15	empty	0.93	28
activity	0.86	7	tries	0.87	15	eats	0.83	24
opposing	1.00	5	owner	0.87	15	sleeps	0.95	20

(a) entailment **(b)** neutral **(c)** contradiction

				MPE				
Word	**Score**	**Freq**	**Word**	**Score**	**Freq**	**Word**	**Score**	**Freq**
an	0.57	21	smiling	0.56	16	sitting	0.51	88
gathered	0.58	12	An	0.60	10	woman	0.55	80
girl	0.50	12	for	0.56	9	men	0.56	34
trick	0.55	11	front	0.75	8	Some	0.62	26
Dogs	0.55	11	camera	0.62	8	doing	0.59	22
watches	0.60	10	waiting	0.50	8	Children	0.50	22
field	0.60	10	posing	0.50	8	boy	0.67	21
singing	0.50	10	Kids	0.57	7	having	0.65	20
outside	0.67	9	smile	0.83	6	sit	0.60	15
something	0.62	8	wall	0.50	6	children	0.53	15

(d) entailment **(e)** neutral **(f)** contradiction

Figure 5.3: Lists of the most highly-correlated words in each dataset for given labels, thresholded to the top 10 and ranked according to frequency. These numbers are based on statistics of each datasets' development set.

What are "Give-away" Words?

Now that we analyzed the extent to which highly label-correlated words may exist across sentences in a given label, we would like to understand what these words are and why they exist.

Figure 5.3 reports some of the words with the highest $p(l|w)$ for SNLI, a human elicited dataset, and MPE, a human judged dataset, on which our hypothesis model performed identically to the majority baseline. Because many of the most discriminative words are low frequency, we report only words which occur at least five times. We rank the words according to their overall frequency, since this statistic is perhaps more indicative of a word w's effect on overall performance compared to $p(l|w)$ alone.

The score $p(l|w)$ of the words shown for SNLI deviate strongly, regardless of the label. In contrast, in MPE, scores are much closer to a uniform distribution of $p(l|w)$ across labels. Intuitively, the stronger the word's deviation, the stronger the potential for it to be a "give-away" word. A high word frequency indicates a greater potential of the word to affect the overall accuracy on RTE.

Qualitative Examples

Turning our attention to the qualities of the words themselves, we can easily identify trends among the words used in contradictory hypotheses in SNLI. In our top-10 list, for example, three words refer to the act of sleeping. Upon inspecting corresponding context sentences, we find that many contexts, which are sourced from Flickr, naturally deal with activities. This leads us to believe that as a common strategy, crowd-source workers often do not generate contradictory hypotheses that require fine-grained semantic reasoning, as a majority of such activities can be easily negated by removing an agent's agency, i.e. describing the agent as sleeping. Two of the other terms, *driving* and *eats*, support this theory since these are actions that someone in a picture of Flickr would not be doing, especially if they are busy with another activity. We refer to these terms as "sage-advice", actions that should not be performed while multi-tasking.

Another trend we notice is that universal negation constitutes four of the remaining seven terms in this list, and may also be used to similar effect.[10] The human-elicited protocol does not guide, nor incentivize crowd-source workers to come up with less obvious examples. If not properly controlled, elicited datasets may be prone to many label-specific terms. The existence of label-specific terms in human-elicited NLI datasets does not invalidate the datasets nor is surprising. Studies in eliciting norming data are prone to repeated responses across subjects (McRae et al., 2005) (see discussion in §2 of (Zhang et al., 2017)).

On the Role of Grammaticality

Like MPE, FN+ contains few high frequency words with high $p(l|w)$. However, unlike on MPE, our hypothesis-only model outperforms the majority-only baseline. If these gains do not arise from "give-away" words, then what is the statistical irregularity responsible for this discriminative power?

Upon further inspection, we notice an interesting imbalance in how our model performs for each of the two classes. The hypothesis-only model performs similarly

[10] These are "Nobody", "alone", "no", and "empty".

label	Hyp-Only	MAJ	Δ%
entailed	44.18	43.20	+2.27
not-entailed	76.31	56.79	+34.37

Table 5.3: Accuracies on FN+ for each class label.

to the majority baseline for entailed examples, while improving by over 34% those which are not entailed, as shown in Table 5.3.

As shown by White et al. (2017) and discussed in Section 4.2.4,FN+ contains more grammatical errors than the other recast datasets. We explore whether grammaticality could be the statistical irregularity exploited in this case. We manually sample a total of 200 FN+ sentences and categorize them based on their gold label and our model's prediction. Out of 50 sentences that the model correctly labeled as ENTAILED, 88% of them were grammatical. On the other-hand, of the 50 hypotheses incorrectly labeled as ENTAILED, only 38% of them were grammatical. Similarly, when the model correctly labeled 50 NOT-ENTAILED hypotheses, only 20% were grammatical, and 68% when labeled incorrectly. This suggests that a hypothesis-only model may be able to discover the correlation between grammaticality and RTE labels on this dataset.

Lexical Semantics

A survey of gains (Table 5.4) in the SPR dataset suggest a number of its property-driven hypotheses, such as *X was sentient in [the event]*, can be accurately guessed based on lexical semantics (background knowledge learned from training) of the argument. For example, the hypothesis-only baseline correctly predicts the truth of hypotheses in the dev set such as: *Experts were sentient ...* or *Mr. Falls was sentient ...*, and the falsity of *The campaign was sentient*, while failing on referring expressions like *Some* or *Each side*. A model exploiting regularities of the real world would seem to be a different category of dataset bias: while not strictly *wrong* from the perspective of NLU, one should be aware of what the hypothesis-only baseline is capable of, to recognize those cases where access to the context is required and therefore more interesting under RTE.

Open Questions

There may remain statistical irregularities, which we leave for future work to explore. For example, are there correlation between sentence length and label class in these data sets? Is there a particular construction method that minimizes the amount of "give-away" words present in the dataset? And lastly, our study is another in a line of research which looks for irregularities at the word level (MacCartney, Galley, and Manning, 2008; MacCartney, 2009). Beyond bag-of-words, are there multi-word expressions or syntactic phenomena that might encode label biases?

5.6 Discussion

We introduced a stronger baseline for eighteen RTE datasets. Our baseline reduces the task from labeling the relationship between two sentences to classifying a single

Proto-Role	H-model	MAJ	Δ%
aware	88.70	59.94	+47.99
used in	77.30	52.72	+46.63
volitional	87.45	64.96	+34.62
physically existed	87.97	65.38	+34.56
caused	82.11	63.08	+30.18
sentient	94.35	76.26	+23.73
existed before	80.23	65.90	+21.75
changed	72.18	64.85	+11.29
chang. state	71.76	64.85	+10.65
existed after	79.29	72.91	+8.75
existed during	90.06	85.67	+5.13
location	93.83	91.21	+2.87
physical contact	89.33	86.92	+2.77
chang. possession	94.87	94.46	+0.44
moved	93.51	93.20	+0.34
stationary during	96.44	96.34	+0.11

Table 5.4: RTE accuracies on the SPR development data; each property appears in 956 hypotheses.

hypothesis sentence. Our experiments demonstrated that in most of the eighteen datasets, always predicting the majority-class label is not a strong baseline, as it is significantly outperformed by the hypothesis-only model. Our analysis suggests that statistical irregularities, including word choice and grammaticality, may reduce the difficulty of the task on popular RTE datasets by not fully testing how well a model can determine whether the truth of a hypothesis follows from the truth of a corresponding premise.

Lessons for future recasting

The high hypothesis-only accuracy on the recast NER dataset may demonstrate that the hypothesis-only model is able to detect that the distribution of class labels for a given word may be peaky. For example, *Hong Kong* appears 130 times in the training set and is always labeled as a location. Based on this, considerations should be taken into account when recasting more datasets in the future. First, since *interesting natural language inference should depend on both premise and hypothesis*, datasets and tasks where context is often not necessary for predicting a label might not be suitable for recasting into RTE. Second, large datasets might make it difficult to evaluate NLP models as peaky distributions might be learned during training. Therefore, we advocate for small, high quality test sets. If a large dataset is indeed constructed, which has been show to be beneficial for pretraining, then one might want to ensure that terms with peaky distributions do not appear in more than one split of the data.

Impact in the field

When this work was originally published at StarSem2018, we hoped our findings would encourage the development of new RTE datasets which exhibit less exploitable irregularities, and that encourage the development of richer models of inference. We advocated for the inclusion of hypothesis-only baselines in future dataset and model reports. Since then, the community has answered our call as many have included hypothesis-only baselines for new RTE (or related) datasets (Welleck et al., 2019; Clark et al., 2019; Sileo et al., 2019; Yanaka et al., 2019; Víta and Klímek, 2019; Sakaguchi et al., 2020; Bhagavatula et al., 2020; Yu et al., 2020; Bisk et al., 2020;

Kim et al., 2020), models (Yin and Schütze, 2018), or similar studies (Schuster et al., 2019; Feng, Wallace, and Boyd-Graber, 2019; Bras et al., 2020; Sun, Guzmán, and Specia, 2020).

People in the community have taken liberty with this work by extrapolating conclusions that we are not fully comfortable with based on our results. For example, Li, Mou, and Keller (2019) state that we "argue that [RTE] as currently formulated is not a difficult task." We used these results to claim that "the majority baselines might not be as indicative of the difficulty of the task as previously thought" and that these biases might reduce the diffuclty of the task on popular datasets. We did not argue that RTE is a not difficult task. In follow-up work (Chen et al., 2020b) we argue that RTE should move away from categorical classes to scalar predictions representing the likelihood of a hypothesis given a premise. However, this was motivated by human subjective probability assessments, not by the seemingly lack of difficulty in the traditional categorical based RTE formation.

In his EMNLP 2018 keynote presentation titled "The Moment of Meaning and the Future of Computational Semantics",[11] John Bos acknowledged that "the Poliak people say we should use other baselines" since "'they show you can make pretty good predictions by looking at the hypothesis ... that's just ridiculous, right?" However, I disagree with his conclusion that these datasets are not good for checking textual entailment. The task and most datasets are still challenging and difficult

[11] https://vimeo.com/306143088

as state-of-the-art models achieve much higher accuracies than their hypothesis-only counterparts. These datasets are not tainted by the fact that hypothesis-only models have drastically improved a majority baselines, rather, these hypothesis-only models help us understand biases in the datasets that might be based on annotation artifacts or domain specific biases. Our analysis helps understand what domain specific biases might be present in a dataset and can illuminate what a specific RTE dataset might actually be testing for.

Chapter 6

Revisiting Paraphrasing in RTE

> A capacity for reliable, robust, open-domain natural language inference is arguably a necessary condition for full natural language understanding
>
> (MacCartney, 2009).

We now turn towards addressing issues related to hypothesis-only biases discovered in the previous chapter. In RTE, systems "must be able to deal with all manners of linguistic phenomena and broad variability of semantic expression" (MacCartney, 2009). We posit that if an RTE model has a sufficiently high *capacity for reliable, robust inference necessary for full NLU*, then the model's predictions should be consistent when an input example is paraphrased. In multiple times in this thesis,[1] we mentioned how the recast dataset focused on paraphrastic inference contained biases. Due to White et al. (2017)'s recasting method, *non-entailed* hypotheses were often ungrammtical or disfluent.

In this chapter, we introduce a RTE test set that evaluates how *reliable* and *robust*

[1] Section 5.5 and Section 4.2.3

models are to paraphrases. The test set that we present here consists of examples from the Pascal RTE1-3 challenges (Dagan, Glickman, and Magnini, 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007) that have be rewritten with a lexical rewriter and manually verified to be high quality paraphrases of the original premises and hypotheses. We use this dataset to determine whether different classes of models (bag of words, LSTMs, or transformers) are robust to paraphrases. If an RTE model's predictions are not consistent when premises and hypotheses are paraphrased, then the system has a far way to go towards understanding natural language.

While this may not be a sufficient test to determine whether RTE models *fully understand* language, as there are many semantic phenomena that RTE models should capture (Cooper et al., 1996; Naik et al., 2018a), it is *necessary* that any NLU system be able to handle our paraphrasing test.

6.1 Motivation

RTE and Paraphrasing

Paraphrasing and Textual Entailment are tightly connected (Androutsopoulos and Malakasiotis, 2010) phenomena that are key to NLU. Many researchers often define paraphrasing as a special case of entailment where both premise and hypothesis entail each other (Nevěřilová, 2014; Fonseca and Aluísio, 2015; Víta, 2015). Paraphrasing has been used to improve RTE systems (Bosma and Callison-Burch, 2006) and RTE

unemployment is at an all-time low ▶ unemployment is at an all-time poor
aeoi 's activities and facility have been tied to several universities ▶ aeoi 's activities and local have been tied to several universities
jerusalem fell to the ottomans in 1517 , remaining under their control for 400 years ▶ jerusalem fell to the ottomans in 1517 , remaining under their regulate for 400 years
usually such parking spots are on the side of the lot ▶ usually such parking spots are dated the side of the lot

Table 6.1: Not-entailed examples from FN+'s dev set where the hypotheses are ungrammatical. The first line in each section is a premise and the lines with ▶ are corresponding hypotheses. Underline words represent the swapped paraphrases.

has been used for paraphrase identification (Seethamol and Manju, 2017) and generation (Brad and Rebedea, 2017).

NLP systems are often sensitive to changes to input data. For example, neural machine translation systems are brittle to synthetic and natural noise (Belinkov and Bisk, 2018), machine reading models fail when distracting sentences are added (Jia and Liang, 2017), and sentiment analysis fails when tokens are swapped with synonyms (Iyyer et al., 2018). In general, the brittleness of NLP systems is undesirable, but in particular, models trained towards the goal of natural language understanding (NLU) *must not* be brittle to paraphrases.

White et al. (2017) recast FN+ dataset

As mentioned in Section 5.3, to create not-entailed hypotheses, White et al. (2017) replaced a single token in a context sentence with a word that crowd-source workers labeled as not being a paraphrase of the token in the given context. In FN+ (Pavlick et al., 2015), two words might be deemed to be incorrect paraphrases in context based on a difference in the words' part of speech tags. Table 6.1 demonstrates such examples, and in the last example, the words "on" and "dated" in the premise and hypothesis respectively have the `NN` and `VBN` POS tag. Therefore, the goal of this chapter is to create a new RTE dataset of grammatical and fluent examples that can be used to determine whether RTE systems are brittle to paraphrases.

6.2 Creating $\hat{p}RTE$

In this section, we describe how we create $\hat{p}RTE$, a vetted, paraphrased version of an existing RTE dataset to test whether models are robust to paraphrased input. Broadly, we used a sentence-level rewriter to create P' and H', paraphrases of premises (P) and hypotheses (H) in a existing RTE dataset. We then rely on crowd-source workers to determine whether P' and H' are fluent and grammatical and how well each P' and H' are paraphrases of corresponding P and H. For each P-H pair, we create three new RTE examples: P'-H, P-H', P'-H'.

We use the examples from the PASCAL RTE1-3 challenges as our test set, as opposed to other RTE datasets for multiple reasons. First, the examples and annotations in RTE1-3 are known to be of high quality. Secondly, the examples originated from multiple domains and different tasks, enabling us to incorporate aspects of *open domain inference* that MacCartney (2009) describes. Thirdly, although recent "probing" RTE test sets perturb MNLI examples (Kim et al., 2019; Ross and Pavlick, 2019), we do not generate paraphrases of MNLI because we want to avoid social (Rudinger, May, and Van Durme, 2017) and hypothesis-only (Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018b) biases that are present in human elicited datasets like MNLI (McRae et al., 2005). A paraphrased version of MNLI has been used to train an RTE model to be "more tolerant of minor lexical differences, better able to generalize, and less inclined to memorize" (Hu et al., 2019). Here, our focus is using this data set as a test metric and not to develop robust models.

Rewriting premises and hypotheses

We use ParaBank (Hu et al., 2019) to create three re-written sentences for each premise and hypothesis in the 3,277 pairs[2] in the PASCAL RTE1-3 datasets.[3] ParaBank is a state-of-the-art sentence-level rewriter that uses lexically-constrained sequence decoding with a vectorized dynamic beam allocation. Since we want to test paraphrastic understanding beyond simple lexical replacement, we discarded the re-written sentences that had at most a 80% lexical overlap with the corresponding original sentence.

[2] 277 pairs and 3k pairs in dev/test sets respectively.

[3] At each time step during decoding, we allowed the rewriter to sample 10 words.

P	The recent G8 summit, held June 8-10, brought together leaders of the world's major industrial democracies
P'	At a recent G8 summit, held in June 8-10, leaders of the world's largest industrial lords were brought together
H	The recent G8 summit took place on June 8-10
H'	The recent G8 summit was in June 8-10

Table 6.2: The top half represents an original premise (P) and its paraphrase (P'). The bottom half depicts an original hypothesis (H) and its paraphrase (H'). If a model is robust to paraphrases, it's prediction (entailed in this example) should be consistent for the following pairs: P-H, P'-H', P-H', and P'-H'.

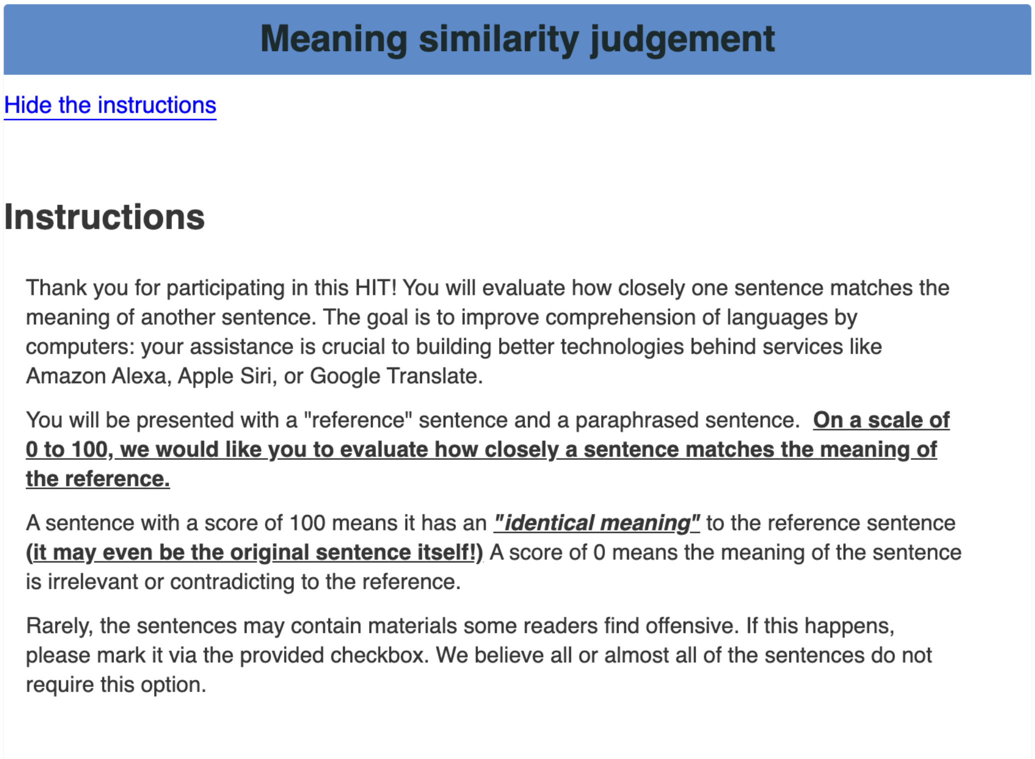
Evaluating paraphrase quality

To ensure that the re-written sentences are indeed sentence-level paraphrases for the original sentences, we relied on crowdsource workers to remove low quality paraphrases. Annotators were asked to assign a score between 0 and 100 representing the similarity of a sentence and 3 presented paraphrases and also determine whether the paraphrases were grammatical. Figure 6.1 include the instructions showen to crowdsource workers for judging similarity between sentences. We retained rewritten sentences that were deemed to be grammatical and had a paraphrase score between 90 (inclusive) and 100 (exclusive).[4] This resulted in a set of 812 new sentence-level paraphrases.[5]

Next, we map the re-written sentences to the original RTE examples to create paraphrased RTE-pairs. Any original RTE pairs that have an approved P' and H' are included in our robust-to-paraphrase collection. Out of the original 3,277 RTE

[4] We exclused rewritten sentences with a score of 100 as a simple method to deal with bad annotators.

[5] 95 and 717 sentences from the dev/test sets respectively.



Meaning similarity judgement

Hide the instructions

Instructions

Thank you for participating in this HIT! You will evaluate how closely one sentence matches the meaning of another sentence. The goal is to improve comprehension of languages by computers: your assistance is crucial to building better technologies behind services like Amazon Alexa, Apple Siri, or Google Translate.

You will be presented with a "reference" sentence and a paraphrased sentence. **On a scale of 0 to 100, we would like you to evaluate how closely a sentence matches the meaning of the reference.**

A sentence with a score of 100 means it has an ***"identical meaning"*** to the reference sentence **(it may even be the original sentence itself!)** A score of 0 means the meaning of the sentence is irrelevant or contradicting to the reference.

Rarely, the sentences may contain materials some readers find offensive. If this happens, please mark it via the provided checkbox. We believe all or almost all of the sentences do not require this option.

Figure 6.1: Instructions for semantic similarity and grammatically check.

examples, 379 RTE pairs remain,[6] resulting in a total of 1,516 premise-hypothesis pairs.[7] We use these examples to evaluate whether an RTE model changes its predictions when RTE examples are paraphrased. Table 6.2 provides an example of a premise-hypothesis pair with corresponding paraphrases.

6.3 Models under investigation

Our goal is to determine whether RTE models are robust to paraphrasing. In this study, we explore models built upon three different classes of sentence encoders: bag

[6] 40 dev and 339 test examples

[7] 160 dev and 1,356 test examples

of words (BOW), Recurrent Neural Networks (RNN), and Transformer's.

Bag of Words Representation

A linear classifier on top of a bag of words representation is a strong baseline for NLU tasks in general (Joachims, 1998; Joulin et al., 2017).[8] We use 300 dimensional GloVe embeddings (Pennington, Socher, and Manning, 2014) to represent each word and sentence representations of the premise and hypothesis are obtained by averaging their respective word vectors. The sentence representations are then concatenated together to form a single sentence-pair representation which is passed to a fully-connected hidden layer with 100 dimensions. The output from the hidden layer is fed to a logistic regression softmax classifier.

RNN representation

The second type of encoder we explore is a recurrent neural network (RNN). Specifically, we use a Bi-LSTM encoder to follow the methods of the InferSent (Conneau et al., 2017) model due to its popularity and high performance on a wide range of RTE datasets. In particular, the InferSent model uses a Bi-LSTM encoder to individually encode the premise and hypotheses. We use BiLSTM encoders with one layer in each direction. Sentence representations of length 2048 are extracted from the encoders by max-pooling and their concatenation is fed to a MLP with one hidden layer of 512 dimensions.

[8] Thank you to Nils Holzenberger for pointing me to these citations.

Transformer

We consider two recent transformer models: BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2020). The latter is a modification of the former where the next sentence prediction objective is removed during training and the masking pattern is dynamically changed on training data. Additionally, RoBERTa is trained on longer sentences, for a higher number of epochs and with bigger batches over an increased amount of training data. For the transformer models, use pre-built models released by huggingface.

6.4 Experiments & Results

We train each of these models on the MNLI dataset (Williams, Nangia, and Bowman, 2017). We use MNLI as a representative RTE dataset because it's large size (over $550k$ examples) enable us to train data-intensive neural models, its covers a wide range of genres, and others have recently used models trained on MNLI when investigating how well RTE models capture different linguistic phenomena (Richardson et al., 2020; Yanaka et al., 2020; Jeretic et al., 2020b). MNLI uses three labels and RTE uses only two labels. When we test the models on $\hat{p}RTE$, we map the models' 'contradiction' and 'neutral' predictions to the 'not-entailed' label in $\hat{p}RTE$.

We are interested in exploring whether these models are robust to paraphrases in RTE. Consequently, we compute how often the models' predictions are consistent

Model Metric	RoBERTa	BiLSTM	BOW
accuracy	73.30	55.09	55.60
robust-0	61.65	61.36	51.92
robust-1	10.03	12.39	20.65
robust-2	22.42	20.65	20.94
robust-3	5.90	5.60	6.49

Table 6.3: Results on $\hat{p}RTE$. First line represents the models accuracy in general. Each subsequent row represents different robust-to-paraphrase measures.

across each each P-H, P'-H, P-H', P'-H' grouping. We refer to this metric as robust-0 since it indicates in what percentages of groupings did the model's predictions not change. The results in Table 6.3 demonstrate that RoBERTa performs well on $\hat{p}RTE$ as it achieves a 73.30% accuracy and is more robust to paraphrases than the other models. Table 6.3 also reports the percentage of times the models' predictions change from its prediction on the unparaphrased example for 1, 2, or 3 instances in each grouping.

Figure 6.2 further breaks down these results by reporting how well the models perform on the test set examples from $\hat{p}RTE$. Here, accuracies are reported based on whether the premise/hypothesis was paraphrased. The RoBERTa based model drastically outperforms the other models, but we see the performance drop on paraphrased examples. We see a similar drop in performance for the BOW model for paraphrased examples. Interestingly, the BLSTM-based model performs the best when both the premises and hypotheses are paraphrased.

We also use $\hat{p}RTE$ to evaluate BERT and RoBERTa sentence-level classification

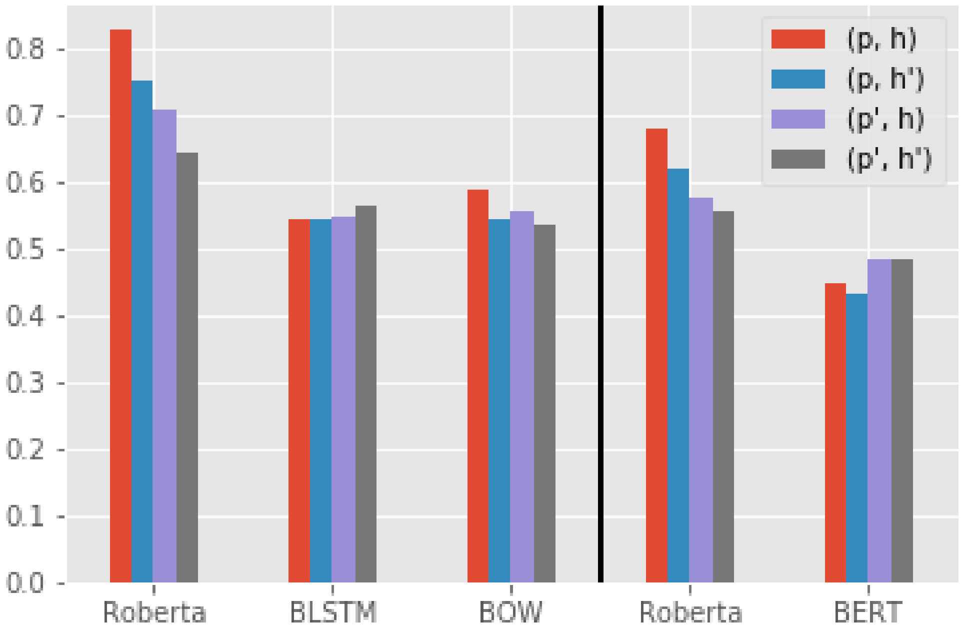
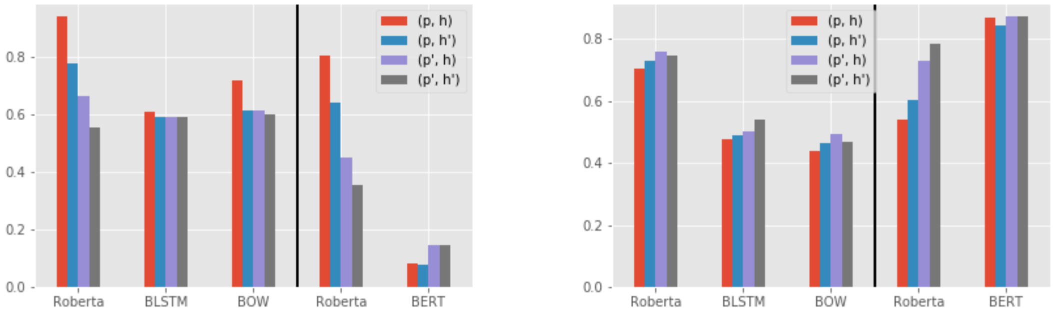


Figure 6.2: Accuracies across all 1,356 test examples where RTE labels from each P-H are propagated to the corresponding P'-H, P-H', and P'-H' examples. The vertical line seperates the models trained on MNLI (left) and the models that were not (right).

models that have not been fine tuned on MNLI (right of the vertical line in Figure 6.2). Interestingly, the RoBERTa model that was not fine-tuned on MNLI outperforms the BLSTM and BOW models that were trained on MNLI. Both of these models outperform the BERT model that was not fine tuned on MNLI, regardless of whether the premises or hypotheses were paraphrased.

Figure 6.3 break down the results from Figure 6.2 based on when the gold label is entailed (6.3a) and not entailed (6.3b). For the examples where the hypothesis is entailed by the premise, the models trained on MNLI consistently perform better when the examples are not paraphrased. Interestingly, the opposite is true when the



(a)

(b)

Figure 6.3: Accuracies across the examples where the gold label is entailed (a) or not-entailed (b).

hypothesis is not entailed by the premise. In these examples, the models trained on MNLI perform better on the paraphrased.

When the BERT model is not fine-tuned on MNLI, it more often predicts that the hypothesis is not entailed by the premise. For the RoBERTa model not fine-tuned on MNLI, we see the model often predicts entailed for the non-paraphrased examples and not-entailed when the sentences are paraphrased.

The examples in RTE1-3 originate from datasets for downstream tasks, e.g. question-answering, information extraction, and summarization. Figure 6.4 reports accuracy for each model broken down by the original task. In general, these models perform the best on the examples originating from the Comparable Document application. This is similar to the finding from the first PASCAL RTE challenge that

> The Comparable Documents (CD) task stands out when observing the performance of the various systems broken down by tasks. Generally the results on the this task are significantly higher than the other tasks with results as high as 87% accuracy and cws of 0.95. This behavior might indicate that in comparable documents there is a high prior probability

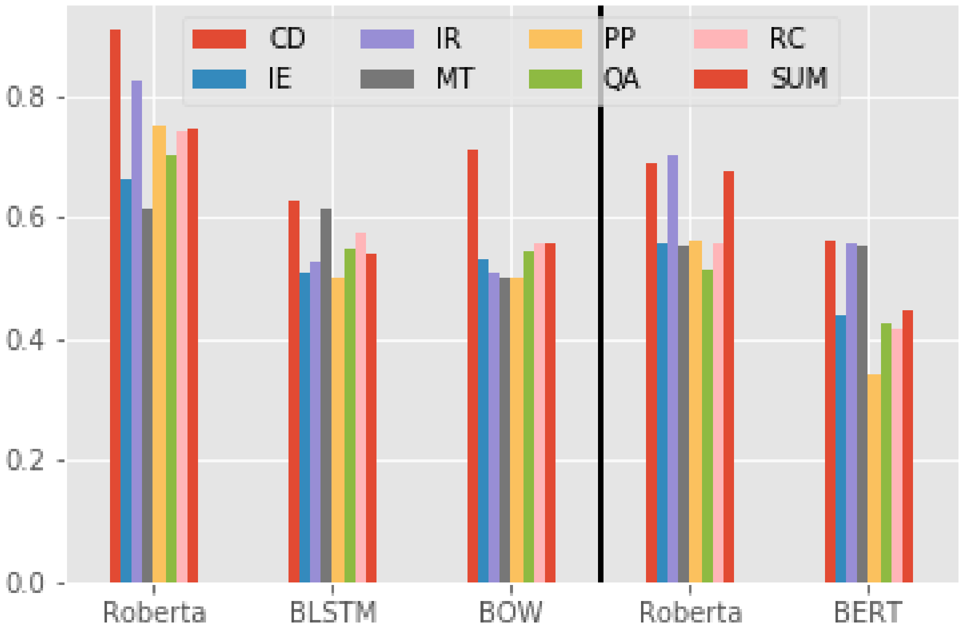


Figure 6.4: Accuracies across $\hat{p}RTE$ broken down by each task where the RTE data originated from.

> that seemingly matching sentences indeed convey the same meanings. We also note that that for some systems it is the success on this task which pulled the figures up from the insignificance baselines.
>
> (Dagan, Glickman, and Magnini, 2006)

6.5 Semantic Variability

MacCartney (2009) argues that in addition to being *reliable* and *robust*, RTE models must deal with the *broad variability of semantic expression*. In other words, though two sentences may be semantically congruent, it is possible that small variations in a paraphrased sentence contain enough semantic variability to change what

ID	Premise	Hypothesis
104-RTE2	Hands Across the Divide was formed in March 2001, and one of its immediate aims was to press for more freedom of contact and communication right away between the two parts of Cyprus, and for early progress towards a solution to 'the Cyprus problem'.	Cyprus was divided into two parts in March 2001.
412-RTE2	Philadelphia is considered the birthplace of the United States of America, where the Declaration of Independence and Constitution were written and signed in the city's Independence Hall.	The US Declaration of Independence is located in Philadelphia.
729-RTE3	Mice given a substance found in red wine lived longer despite a fatty diet, a study shows.	Mice fed with red wine lived longer despite a fatty diet.

Table 6.4: Examples in the development set that RoBERTa incorrectly predicted entailed for all paraphrased versions.

would likely, or not likely be inferred from the sentence. In other words, even though all P' and H' in $\hat{p}RTE$ have been deemed to be semantically congruent with their corresponding original sentences, the semantic variability of paraphrases might change whether H or H' can be inferred from P' or P. Consequently, projecting the RTE label from an original P-H example to the corresponding P-H', P-H', and P'-H' examples might introduces incorrect annotations. We explore the paraphrased examples in the development set where the model disagreed with the propogated RTE label to understand whether the model correctly predicted whether the label should change.

The RoBERTa model trained on Multi-NLI disagrees with the original label in just 31 instances out of the 160 dev examples. For three P-H, P'-H, P-H', P'-H' sets (shown in Table 6.4), the RoBERTa model incorrectly predicted entailed while the original label was not entailed. While the RoBERTa model incorrectly predicts the labels in these examples, these specific examples might demonstrate that RoBERTa is robust to paraphrases since the model's predictions remains the same as these sentences are paraphrased.

6.6 Discussion

In this chapter, we introduced $\hat{p}RTE$, a test suite of human vetted paraphrased RTE examples. This test suite was created by rewriting RTE examples using a state-of-the-art sentence level rewriter and then relying on crowdsource workers to determine the grammaticality and fluency of the rewritten sentences. We retained the sentences that passed this step to created paraphrased RTE examples.

Our experiments demonstrated that contemporary a state-of-the-art transformer model is more robust to paraphrases than the models based on a bag of words or BiLSTM representations. In the experiments here, we propagated the labels from the original RTE examples to their corresponding paraphrased examples. We leave the question of whether these new paraphrased examples should be annotated instead of relying on the original RTE labels an an open question for future work.

This chapter concludes the focus on RTE as an evaluation framework to explore the reasoning capabilities of NLP models. In the next chapter, we will explore modeling approaches for overcoming biases in RTE datasets.

Chapter 7

Overcoming Hypothesis-Only Biases

This thesis primarily advocates for Recognizing Textual Entailment (RTE) as an evaluation framework. Since many tasks require common inference capabilities, first pre-training a model to perform RTE before applying and updating its parameters for a specific task is productive. As discussed in Section 2.2.2, many researchers successfully leverage RTE datasets to improve models for downstream tasks (Bentivogli, Dagan, and Magnini, 2017; Guo, Pasunuru, and Bansal, 2018a; Guo, Pasunuru, and Bansal, 2018b).

However, biases in RTE datasets, like hypothesis-only biases discussed in Chapter 5 or stereotypical biases discovered by Rudinger, May, and Van Durme (2017), might limit or negatively impact RTE's usefulness as an intermediate step in the process of building NLP systems. Therefore, it is important to develop methods that overcome biases in RTE datasets. In this chapter, we demonstrate how adversarial learning can help limit a model's ability to capture hypothesis-only biases and spurious correlations in RTE datasets. This section of the chapter is based on our published work (Belinkov et al., 2019a; Belinkov et al., 2019b).

In the last section of this chapter, we will demonstrate how to similarly apply adversarial learning to a real world NLP task. Since RTE encompasses semantic inference that is necessary for many NLP tasks, adversarial learning can help overcome domain-specific biases in an applied setting. The real world task we explore is discovering emergency needs during disaster events. We successfully employed this method in an assemble approach for the DARPA LORELEI challenges. Initial aspects of this work has been presented at a refereed regional conference (Poliak and Van Durme, 2019).

7.1 Motivation

There are multiple approaches for dealing with the issues presented with biases in RTE datasets. One common approach is to create new unbiased data or to remove biased examples. In fact, recent studies have tried to create new RTE datasets that do not contain such biases, but many such approaches remain unsatisfactory. Constructing new datasets (Sharma et al., 2018) is costly and may still result in other artifacts. Filtering "easy" examples and defining a harder subset is useful for evaluation purposes (Gururangan et al., 2018), but difficult to do on a large scale that enables training. Compiling adversarial examples (Szegedy et al., 2014; Goodfellow, Shlens, and Szegedy, 2015; Glockner, Shwartz, and Goldberg, 2018) is informative but again limited by scale or diversity.

Rather than modifying the biased data itself, another common approach in NLP advocates for removing discovered biases from already trained models. For example, Bolukbasi et al. (2016) present a method, `HARD-DEBIAS`, for removing gender biases in pre-trained word embeddings. Bolukbasi et al. (2016) first identify a gendered specific subspace based on pre-computed word embeddings for gender specific words. Next, they either neutralize or soften the gender aspects of the word embeddings. Neutralizing "ensures that gender neutral words are zero in the gender subspace" while softening "reduces the differences between" sets of gendered terms "while maintaining as much similarity to the original embedding as possible." Liang et al. (2020) introduce an extension called `SENT-DEBIAS` that removes gender and religion based biases from contemporary pre-trained sentence representations. They similarly compute a biased subspace, but then subtract that subspace from learned sentence representations. Instead of substracting a biased subspace, Ravfogel et al. (2020) project sentence representations to their null-space to mitigate biases. Each of these approaches require a known set of biased terms or phrases that are used to discover these biased subspaces.

The third common approach is to encourage models to ignore or overcome biases during training. Such contemporary approaches often leverage domain-adversarial neural networks, which aim to increase robustness to domain change, by learning to be oblivious to the domain using gradient reversals (Ganin et al., 2016). NLP researchers primarily employ gradient reversal based adversarial learning to build models that generalize across domains (Zhang, Barzilay, and Jaakkola, 2017; Chen et al., 2018b; Lample et al., 2018). Recently, researchers similarly leverage this method to encourage models to ignore specific biases. For example, Li, Baldwin, and Cohn (2018) use adversarial networks to discourage models from learning sensitive information like sex and age when performing part of speech tagging or sentiment analysis. Similarly, Elazar and Goldberg (2018) use adversarial learning to ignore protected attributes like race, age, and gender. However, they suggest that fully removing such attributes from text representations may be difficult. In particular, sentence representations might still contain aspects related to the protected attributes even though the adversarially trained model might not rely on that information when making predictions. A similar approach has also been used to mitigate biases in Visual Question Answering (Ramakrishnan, Agrawal, and Lee, 2018; Grand and Belinkov, 2019). Here, we will use adversarial learning to develop RTE models that are robust to hypothesis-only biases.

We will begin by discussing two adversarial learning methods that we apply to a common RTE modeling approach. We will then demonstrate how these approaches help a model overcome hypothesis-only biases and we will discuss related work that followed our results.

7.2 Methods

We consider two types of adversarial methods when training models for RTE. In the first method, we incorporate an external classifier to force the hypothesis-encoder to ignore hypothesis-only biases. In the second method, we randomly swap premises in the training set to create noisy examples.[1] First, we review a general RTE model and introduce notation we will use throughout this chapter.

7.2.1 General, baseline RTE model

Let (P, H) denote a premise-hypothesis pair, and let g denote an encoder that maps a sentence S to a vector representation $\boldsymbol{v}$, and c a classifier that maps $\boldsymbol{v}$ to an output label y. A general RTE framework (Figure 7.1a) contains the following components:

- A **premise encoder** g_P that maps the premise P to a vector representation $\boldsymbol{p}$.
- A **hypothesis encoder** g_H that maps the hypothesis H to a vector representation $\boldsymbol{h}$.
- A **classifier** c_{RTE} that combines and maps $\boldsymbol{p}$ and $\boldsymbol{h}$ to an output y.

In this model, the premise and hypothesis are each encoded with separate encoders. This is common to many RTE models (Rocktäschel et al., 2015; Mou et al., 2016;

[1] In Belinkov et al. (2019a), we present a probabilistic interpretation of these methods as well.

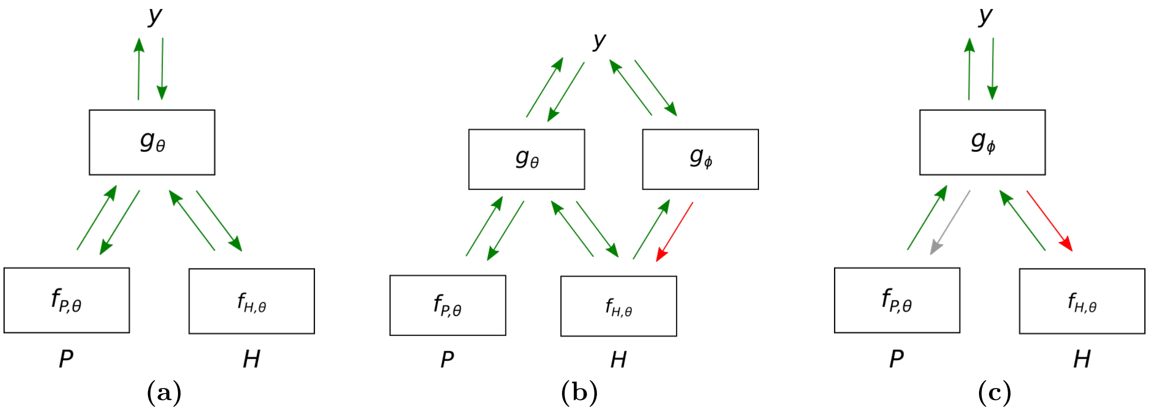


Figure 7.1: Illustration of (a) the baseline RTE architecture, and our two proposed methods to remove hypothesis only-biases from an RTE model: (b) uses a hypothesis-only classifier, and (c) samples a random premise. Arrows correspond to the direction of propagation. Green or red arrows respectively mean that the gradient sign is kept as is or reversed. Gray arrow indicates that the gradient is not back-propagated - this only occurs in (c) when we randomly sample a premise, otherwise the gradient is back-propagated. f and g represent encoders and classifiers.

Cheng, Dong, and Lapata, 2016; Nie and Bansal, 2017; Chen et al., 2017a), although some share information between the encoders via attention (Parikh et al., 2016; Duan et al., 2018).

The RTE classifier is usually trained to minimize the following objective:

$$L_{\text{RTE}} = L(c_{\text{RTE}}([g_P(P); g_H(H)], y)) \tag{7.1}$$

where $L(\tilde{y}, y)$ is the cross-entropy loss. If g_P is not used, a model should not be able to successfully perform RTE. However, models without g_P may achieve non-trivial results, indicating the existence of biases in hypotheses (Chapter 5).

7.2.2 Adversarial Learning

When developing models for RTE, we want to encourage models to overcome or ignore biases in hypotheses. There are different potential Machine Learning approaches that might help this goal. Like the approaches discussed earlier in this chapter (Bolukbasi et al., 2016; Liang et al., 2020; Ravfogel et al., 2020), we could first train a RTE model, then identify a hypothesis-only biased subspace in the learned sentence representations and remove the biased subspace from the model. However, those approaches require a set of pre-identified biased terms or phrases and unlike gender, racial, or religion based biases, hypothesis-only biases are unclear and it may difficult to identify hypothesis-only biased examples apriori. We would like to enable a model to discover hypothesis-only biases on its own and simultaneously ignore and overcome such biases.

Domain-adversarial training, which we will refer to in this thesis as adversarial learning,[2] was introduced as a representation learning approach to help a model generalize to data sampled from different distributions (Ganin et al., 2016). In this form of adversarial learning, an additional domain specific classifier is added to maximize the loss of the domain classifier while the main classifier is optimized to minimize the error on the training set. Ganin et al. (2016) argue that maximizing the loss of the domain classifier "encourages domain-invariant features to emerge in the course of the optimization." Adversarial learning derives from the work of Ben-David et al. (2007)

[2] The term adversarial learning has also been used to describe "the task of learning sufficient information about a classifier to construct adversarial attacks" (Lowd and Meek, 2005).

and Ben-David et al. (2010) and Kifer, Ben-David, and Gehrke (2004) focused on theoretical bounds for domain divergences. In particular, that “the divergence relies on the capacity of the hypothesis class $\mathcal{H}$ to distinguish between examples generated by” different distributions (Ganin et al., 2016). Ganin et al. (2016) use an domain classifier to discriminate between the different distributions, or domains. During training, the domain classifier is trained to maximize its loss. This “adversarial” aspect is implemented with a gradient reversal layer in the network “that leaves the input unchanged during forward propagation and reverses the gradient by multiplying it by a negative scalar during the backpropagation” (Ganin et al., 2016).

Training an adversarial classifier this way allows the model to figure out the domain-invariant features on its own, without any human supervision or prior knowledge. In our setting, we would like the model to determine the hypothesis-only biases on its and consequently learn sentence representations that are invariant to such biases. Therefore, adversarial learning is an appropriate and promising technique for modeling RTE.

7.2.2.1 AdvCls: Adversarial Classifier

Our first approach, referred to as AdvCls, follows the common adversarial training method (Goodfellow, Shlens, and Szegedy, 2015; Ganin and Lempitsky, 2015; Xie et al., 2017; Beutel et al., 2017; Zhang, Lemoine, and Mitchell, 2018). We add an additional adversarial classifier c_{Hypoth} to our general RTE model. c_{Hypoth} maps the hypothesis representation $\boldsymbol{h}$ to an output y. When adversarial learning is applied to NLP problems, the adversarial classifier is typically used to predict unwanted features, e.g., protected attributes like race, age, or gender (Elazar and Goldberg, 2018). Here, we do not have explicit protected attributes but rather *latent* hypothesis-only biases discovered during training. Therefore, we use c_{Hypoth} to predict the RTE label given only the hypothesis. To successfully perform this prediction, c_{Hypoth} needs to exploit latent biases in $\boldsymbol{h}$.

We modify the objective function in Equation 7.1 to become

$$\begin{aligned} L =& L_{\text{RTE}} + \lambda_{\text{Loss}} L_{\text{Adv}} \\ L_{\text{Adv}} =& L(c_{\text{Hypoth}}(\lambda_{\text{Enc}} \text{GRL}_{\lambda}(g_H(H)), y)) \end{aligned}$$

To control the interplay between c_{RTE} and c_{Hypoth} we set two hyper-parameters: λ_{Loss}, the importance of the adversarial loss function, and λ_{Enc}, a scaling factor that multiplies the gradients after reversing them. This is implemented by the scaled gradient reversal layer, GRL_{λ} (Ganin and Lempitsky, 2015). The goal here is modify the representation $g_H(H)$ so that it is maximally informative for RTE while simultaneously minimizes the ability of c_{Hypoth} to accurately predict the RTE label. Figure 7.1b depicts the AdvCls approach.

7.2.2.2 AdvDat: Adversarial Training Data

For our second approach, which we call AdvDat, instead of adding an external component to the general RTE model, we use the general model as is, but rather train it with perturbed training data. For a fraction of example (P, H) pairs in the training data, we replace P with P', a premise from another training example, chosen uniformly at random. For these instances, during back-propagation, we similarly reverse the gradient but only back-propagate through g_H. The adversarial loss function L_{RandAdv} is defined as:

$$L_{\text{RandAdv}} = L(c_{\text{RTE}}([\text{GRL}_0(g_P(P')); \lambda_{\text{Enc}}\text{GRL}_\lambda(g_H(H))], y))$$

where GRL_0 implements gradient blocking on g_P by using the identity function in the forward step and a zero gradient during the backward step. At the same time, GRL_λ reverses the gradient going into g_H and scales it by λ_{Enc}, as before.

We set a hyper-parameter $\lambda_{\text{Rand}} \in [0, 1]$ that controls what fraction P's are swapped at random. In turn, the final loss function combines the two losses based on λ_{Rand} as

$$L = (1 - \lambda_{\text{Rand}})L_{\text{RTE}} + \lambda_{\text{Rand}}L_{\text{RandAdv}}$$

In essence, this method penalizes the model for correctly predicting y in perturbed examples where the premise is uninformative. This implicitly assumes that the label for (P, H) should be different than the label for (P', H), which in practice does not always hold true.[3]

Implementation Details

As in earlier experiments in this thesis, we use `InferSent` (Conneau et al., 2017) as our baseline model because it has been shown to work well on popular RTE datasets and is representative of many RTE models. We use separate BiLSTM encoders to learn vector representations of P and H. The vector representations are combined following Mou et al. (2016),[4] and passed to an MLP classifier with one hidden layer. Our proposed methods for mitigating biases use the same technique for representing and combining sentences. The classifiers are single-layer MLPs of size 20 dimensions. We train these models with SGD until convergence. The sentence representations learned by the BiLSTM encoders and the MLP classifier's hidden layer have a dimensionality of 2048 and 512 respectively. We follow `InferSent`'s training regime, using SGD with an initial learning rate of 0.1 and optional early stopping. For both methods, we sweep the hyper-parameters over values $\{0.05, 0.1, 0.2, 0.4, 0.8, 1.0\}$.

7.3 Intrinsic Evaluations

We propose these two adversarial learning based methods to develop RTE models that overcome hypothesis only biases. While the main motivation is to use RTE as

[3] As pointed out by a reviewer, a pair labeled as neutral might end up remaining neutral after randomly sampling the premise, so adversarially training in this case might weaken the model. Instead, one could limit adversarial training to cases of entailment or contradiction.

[4] Specifically, representations are concatenated, subtracted, and multiplied element-wise.

a pre-training setup for downstream NLP systems, exhaustive extrinsic evaluations (applying the adverserially- and pre-trained models to real world downstream tasks) is beyond the scope of this thesis. Instead, we rely on intrinsic evaluations to explore whether these models are indeed robust to hypothesis-only biases. As a refresher from Section 2.1.2, extrinsic evaluations test how well a specific model improves a broader downstream system and intrinsic evaluations test a model on the specific task it was trained to perform.

If a model is robust to hypothesis-only biases, then it should perform better than an un-robust model when tested on other datasets that are different than the data used to train them, especially if datasets contain different biases. Our experiments on synthetic and more traditional RTE datasets demonstrate how well our proposed methods improve a model's robustness.

7.3.1 Synthetic Experiment

Consider an example where P and H are strings containing the letters $\{a, b, c\}$, and an environment where P *entails* H if and only if the first letters are the same, as in synthetic dataset A. In such a setting, a model should be able to learn the correct condition for P to entail H. In fact, this is equivalent to XOR and hence is learnable by a multi-layer perceptron (MLP).

Synthetic dataset A

$$(a, a) \rightarrow \textsc{True} \qquad (a, b) \rightarrow \textsc{False}$$

$$(b, b) \rightarrow \textsc{True} \qquad (b, a) \rightarrow \textsc{False}$$

Now, imagine synthetic dataset B that contains a hypothesis-only bias. Let us imagine the biases is that c is appended to every entailed H. A model of y with access only to the hypotheses can fit the data perfectly by detecting the presence or absence of c in H, ignoring the desired definition for entailment in this setting. Therefore, we hypothesize that a baseline model trained on such data would be misled by the bias c, and in turn would fail to learn the desired relationship between the premise and the hypothesis. Consequently, the model would not perform well on the unbiased synthetic dataset A.

Synthetic dataset B (with artifact)

$$(a, ac) \rightarrow \textsc{True} \qquad (a, b) \rightarrow \textsc{False}$$

$$(b, bc) \rightarrow \textsc{True} \qquad (b, a) \rightarrow \textsc{False}$$

We create these synthetic datasets A and B, where P entails H if and only if their first letters are the same. The training and test sets have $1K$ examples each, uniformly distributed among the possible entailment relations. In the test set (dataset A), each premise or hypothesis is a single symbol: $P, H \in \{a, b\}$, where P entails H iff $P = H$. In the training set (dataset B), a letter c is appended to the hypothesis side in the True examples, but not in the False examples. In order to transfer well to the test set, a model that is trained on this training set needs to learn the underlying relationship—that P entails H if and only if their first letter is identical—rather than relying on the presence of c in the hypothesis side.

	λ_{Enc}					
λ_{Loss}	0.1	0.25	0.5	1	2.5	5
0.1	50	50	50	50	50	50
0.5	50	50	50	50	50	50
1	50	50	50	50	50	50
1.5	50	50	50	50	50	100
2	50	50	50	50	100	100
2.5	50	50	100	75	100	100
3	50	100	100	100	100	100
3.5	100	100	100	100	100	100
4	100	100	100	100	100	100
5	100	100	100	100	100	100
10	100	100	100	100	100	100
20	100	100	100	100	100	100

(a) AdvCls

	λ_{Rand}				
λ_{Enc}	0.1	0.25	0.5	0.75	1
0.1	50	50	50	50	50
0.5	50	50	50	50	50
1	50	50	50	50	50
1.5	50	50	50	50	50
2	50	50	50	50	50
2.5	50	50	50	50	50
3	50	50	100	50	50
3.5	50	50	100	50	50
4	50	100	100	50	50
5	50	50	100	100	50*
10	75	100	100	100	50*
20	100	100	100	50*	50*

(b) AdvDat

Table 7.1: Accuracies on the synthetic dataset, when training on the biased training set and evaluating on the unbiased test set. Darker boxes represent higher accuracies. * indicates failure to learn the biased training set; all other configurations learned the training set perfectly.

For our experiments on the synthetic dataset, the characters are embedded with 10-dimensional vectors. Input strings are represented as a sum of character embeddings, and the premise-hypothesis pair is represented by a concatenation of these embeddings.

Synthetic Results

As expected, without a method to remove hypothesis-only biases, the baseline method fails to generalize to the test set. Examining the model's predictions, we found that the baseline model learned to rely on the presence/absence of the bias term c, always predicting TRUE/FALSE respectively.

Table 7.1 shows the results of our two proposed methods. As we increase the hyper-parameters, we find that our methods initially behave like the baseline, learning the training set but failing on the test set. However, with strong enough hyper-parameters (moving towards the bottom in the tables), they perform perfectly on both the biased training set and the unbiased test set. For AdvCls, stronger hyper-parameters work better. AdvDat, in particular, breaks down with too many random samples (increasing λ_{Rand}), as expected. We also found that AdvCls did not require as strong λ_{Enc} as AdvDat. From the synthetic experiments, it seems that AdvCls learns to ignore the bias c and learn the desired relationship between P and H across many configurations, while AdvDat requires much stronger λ_{Enc} in this synthetic setup.

7.3.2 Transferring on Traditional RTE Datasets

For the second set of intrinsic experiments, we train a model on SNLI using our proposed methods, test them on a large suite of other RTE datasets, and compare the difference in accuracy with the baseline model s accuracies on the datasets. We train the models on SNLI since it contains significant annotation artifacts and biases. In these experiments, we use pre-computed 300-dimensional GloVe embeddings (Pennington, Socher, and Manning, 2014) trained on CommonCrawl.[5]

[5] Specifically, glove.840B.300d.zip.

Model	Val	Test
Baseline	84.25	84.22
AdvCls	84.58	83.56
AdvDat	78.45	78.30

Table 7.2: Accuracies for the approaches on SNLI. Baseline refers to the unmodified, non-adversarial InferSent.

In-domain Results

Table 7.2 reports the results on SNLI, with the configurations that performed best on the validation set for each of the adversarial methods. As expected, both training methods perform worse than our unmodified, non-adversarial InferSent baseline on SNLI's test set. The difference for AdvCls is minimal, and it even slightly outperforms InferSent on the validation set. While AdvDat's results are noticeably lower than the non-adversarial InferSent, the drops are still less than 6% points.

RTE test sets

As target datasets, we use the 10 datasets investigated in Chapter 5. These are SCITAIL (Khot, Sabharwal, and Clark, 2018), ADD-ONE-RTE (Pavlick and Callison-Burch, 2016), JOCI (Zhang et al., 2017), MPE (Lai, Bisk, and Hockenmaier, 2017), SICK (Marelli et al., 2014), and MNLI (Williams, Nangia, and Bowman, 2017);[6] as well as the three datasets White et al. (2017) recast into RTE, namely FN+ (Pavlick et al., 2015), DPR (Rahman and Ng, 2012), and SPR (Reisinger et

[6] MNLI comes with two dev/test sets: domains that match or mismatched with the training set.

	Test On Target Dataset			Test On SNLI	
Target Test Dataset	Baseline	Δ AdvCls	Δ AdvDat	Δ AdvCls	Δ AdvDat
SCITAIL	58.14	-0.47	-7.06	-0.18	-9.06
ADD-ONE-RTE	66.15	0.00	17.31	-2.29	-49.63
JOCI	41.50	0.24	-1.87	-0.44	-5.92
MPE	57.65	0.45	-5.30	-0.57	-0.54
DPR	49.86	1.10	-0.45	-0.73	-7.81
MNLI matched	45.86	1.38	-2.10	-1.25	-8.93
FN+	50.87	1.61	6.16	-1.94	-0.44
MNLI mismatched	47.57	1.67	-3.91	-1.25	-8.93
SICK	25.64	1.80	31.11	-0.57	-8.93
GLUE	38.50	1.99	4.71	-1.25	-8.93
SPR	52.48	6.51	12.94	-1.76	-14.01
SNLI-hard	68.02	-1.75	-12.42		

Table 7.3: Accuracy results of transferring representations to new datasets. In all cases the models are trained on SNLI. Left: baseline results on target test sets and differences between the proposed methods and the baseline. Right: test results on SNLI with the models that performed best on each target dataset's dev set. Δ are absolute differences between the method and the baseline on each target test set (left) or between the method and the baseline performance (84.22) on SNLI test (right). Black rectangles show relative changes in each column.

al., 2015).[7] Many of these target datasets have different label spaces than SNLI. Therefore, like in Section 3.4, we map the models' NEUTRAL and CONTRADICTION predictions to the NOT-ENTAILED label in the datasets that annotate the task as a binary classification problem. We also use two other datasets: GLUE's diagnostic test set, which was carefully constructed to not contain hypothesis-biases (Wang et al., 2018), and SNLI-hard, a subset of the SNLI test set that is thought to contain fewer biases (Gururangan et al., 2018). Finally, we also test on the Multi-genre RTE dataset (MNLI; Williams, Nangia, and Bowman, 2017), a successor to SNLI.

[7] See Section 5.3 for details about those datasets.

RTE Transfer Results

Table 7.3 (left block) reports the results of our proposed methods compared to the baseline when tested on the other RTE datasets. For each target dataset, we choose the best-performing model on its development set and report results on the test set.[8] The method using the hypothesis-only classifier to remove hypothesis-only biases from the model (AdvCls) outperforms the baseline in 9 out of 12 target datasets ($\Delta > 0$), though most improvements are small. The training method using negative sampling (AdvDat) only outperforms the baseline in 5 datasets, 4 of which are cases where the other method also outperformed the baseline. These gains are much larger than those of AdvCls.

We also report results of the proposed methods on the SNLI test set (right block). These results are based on the same hyper-parameters used in the left block for each row. As our results improve on the target datasets, we note that AdvCls's performance on SNLI does not drastically decrease (small Δ), even when the improvement on the target dataset is large (for example, in SPR). For this method, the performance on SNLI drops by just an average of 1.11 (0.65 STDV). For AdvDat, there is a large decrease on SNLI as results drop by an average of 11.19 (12.71 STDV). For these models, when we see large improvement on a target dataset, we often see a large drop on SNLI. For example, on ADD-ONE-RTE, AdvDat outperforms the baseline

[8] For MNLI, since the test sets are not available, we tune on the matched dev set and evaluate on the mismatched dev set, or vice versa. For GLUE, we tune on MNLI matched. The hyper-parameters for the best performing model for each dataset can be found online at `https://github.com/azpoliak/robust-nli#hyper-parameters-for-transfer-experiments`.

by roughly 17% but performs almost 50% lower on SNLI. Based on this, as well as the results on the synthetic dataset, AdvDat seems to be much more unstable and highly dependent on the right hyper-parameters.

7.4 Analysis

Our results demonstrate that our approaches may be robust to many datasets with different types of bias. We next analyze our results and explore modifications to the experimental setup that may improve model transferability across RTE datasets.

7.4.1 Interplay with known biases

A priori, we expect our methods to provide the most benefit when a target dataset has no hypothesis-only biases or such biases that differ from ones in the training data. To determine how different a dataset's hypothesis-only biases are from those in SNLI, we compare the performance of a hypothesis-only classifier trained on SNLI and evaluated on each target dataset, to a majority baseline of the most frequent class in the target dataset's training set (MAJ). We also compare to a hypothesis-only classifier trained and tested on each target dataset.[9]

Figure 7.2 shows the results. When the hypothesis-only model trained on SNLI is tested on the target datasets, the model performs below MAJ (except for MNLI),

[9] A reviewer noted that this method may miss similar bias "types" that are achieved through different lexical items. Using pre-trained word embeddings might mitigate this concern.

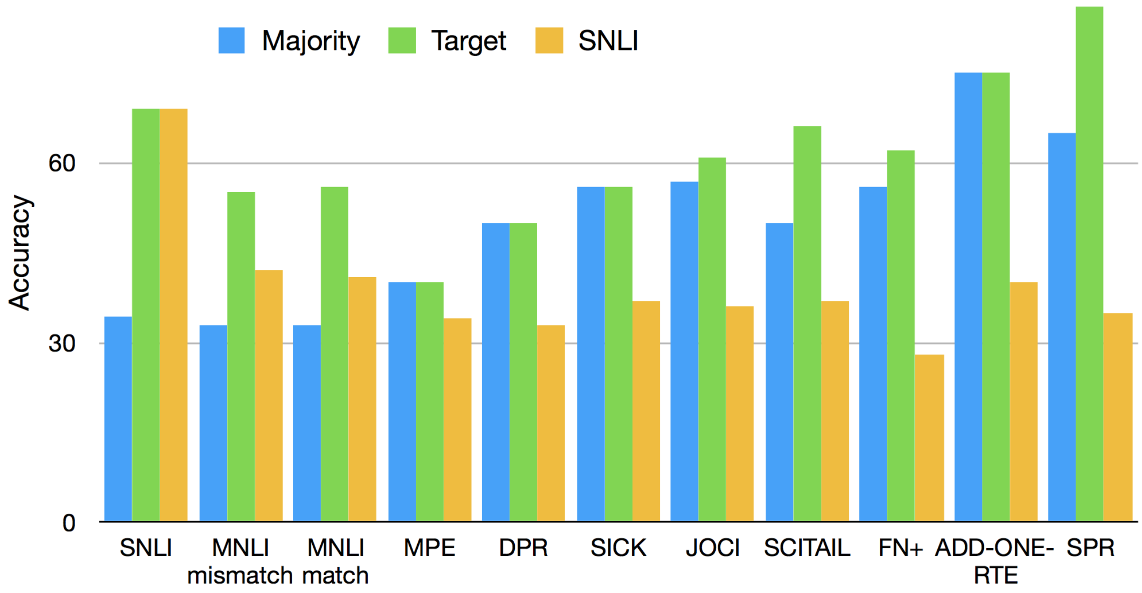


Figure 7.2: Accuracies of majority and hypothesis-only baselines on each dataset (x-axis). The datasets are generally ordered by increasing difference between a hypothesis-only model trained on the target dataset (green) compared to trained on SNLI (yellow).

indicating that these target datasets contain different biases than those in SNLI. The largest difference is on SPR: a hypothesis-only model trained on SNLI performs over 50% worse than one trained on SPR. Indeed, our methods lead to large improvements on SPR (Table 7.3), indicating that they are especially helpful when the target dataset contains different biases. On MNLI, this hypothesis-only model performs 10% above MAJ, and roughly 20% worse compared to when trained on MNLI, suggesting that MNLI and SNLI have similar biases. This should be unsurprising since they both were created by eliciting hypotheses from crowdsource workers. This results may explain why our methods only slightly outperform the baseline on MNLI (Table 7.3).

Dataset	Base	AdvCls	Δ
JOCI	41.50	39.29	-2.21
SNLI	84.22	82.40	-1.82
DPR	49.86	49.41	-0.45
MNLI matched	45.86	46.12	0.26
MNLI mismatched	47.57	48.19	0.62
MPE	57.65	58.60	0.95
SCITAIL	58.14	60.82	2.68
ADD-ONE-RTE	66.15	68.99	2.84
GLUE	38.50	41.58	3.08
FN+	50.87	56.31	5.44
SPR	52.48	58.68	6.20
SICK	25.64	36.59	10.95
SNLI-hard	68.02	63.81	-4.21

Table 7.4: Results with stronger hyper-parameters for AdvCls vs. the baseline. Δ's are absolute differences.

7.4.2 Stronger hyper-parameters

In the synthetic experiment, we found that increasing the hyper-parameters improves the models' ability to generalize to the unbiased dataset. Does the same apply to these RTE datasets? We expect that strengthening the auxiliary losses (L_2 in our methods) during training will hurt performance on the original data (where biases are useful), but improve on the target data, which may have different or no biases (Figure 7.2). To test this, we increase the hyper-parameter values during training; we consider the range $\{1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0\}$.[10] While there are other ways to strengthen our methods, e.g., increasing the number or size of hidden layers (Elazar and Goldberg, 2018), we are interested in the effect of the hyper-parameters as they control how much bias is subtracted from our baseline model.

[10] The synthetic setup required very strong hyper-parameters, possibly due to the clear-cut nature of the task. In the natural RTE setting, moderately strong values sufficed.

Table 7.4 shows the results of AdvCls with stronger hyper-parameters on the existing RTE datasets. Interestingly, performance on SNLI-hard in Table 7.4 noticeably decreases. This drop may indicate that SNLI-hard may still have biases, a pointed Gururangan et al. (2018) concede.[11] Many of the other datasets benefit from stronger hyper-parameters (compared with Table 7.3). We see the largest improvement on SICK, achieving over 10% increase compared to the 1.8% gain in Table 7.3. As for AdvDat, we found large drops in quality even in our basic configurations (appendix), so we do not increase the hyper-parameters further. This should not be too surprising, adding too many random premises will lead to a model's degradation.

7.4.3 Indicator Words

Certain words in SNLI are more correlated with specific entailment labels than others, e.g., negation words ("not", "nobody", "no") correlated with CONTRADICTION. In Section 5.5, we referred to these as "give-away" words. Do the adversarial methods encourage models to make predictions that are less affected by these biased indicator words?

For each of the most biased words in SNLI associated with the CONTRADICTION label, we computed the probability that a model predicts an example as a contradiction, given that the hypothesis contains the word. Table 7.5 shows the top 10 examples in the training set. For each word w, we give its frequency in SNLI, its em-

[11] As pointed out by a reviewer, an alternative explanation is a general loss of information in the encoded hypothesis.

pirical correlation with the label and with InferSent's prediction, and the percentage decrease in correlations with CONTRADICTION predictions by three configurations of our methods. Generally, the baseline correlations are more uniform than the empirical ones ($\hat{p}(l|w)$), suggesting that indicator words in SNLI might not greatly affect a NLI model, a possibility that both we and Gururangan et al. (2018) do concede. For example, Gururangan et al. (2018) explicitly mention that "it is important to note that even the most discriminative words are not very frequent."

However, we still observed small skews towards CONTRADICTION. Thus, we investigate whether our methods reduce the probability of predicting CONTRADICTION when a hypothesis contains an indicator word. The model trained with AdvDat (where $\lambda_{\text{Rand}} = 0.4$, $\lambda_{\text{Enc}} = 1$) predicts contradiction much less frequently than InferSent on examples with these words. This configuration was the strongest AdvDat model that still performed reasonably well on SNLI. Here, AdvDat appears to remove some of the biases learned by the baseline, unmodified `InferSent`. We also provide two other configurations that do not show such an effect, illustrating that this behavior highly depends on the hyper-parameters.

7.5 Discussion

Biases in annotations are a major source of concern for the quality of RTE datasets and systems, and these may limit the usefulness of pre-training models on RTE before

		Score		Percentage decrease from baseline		
Word	Count	$\hat{p}(l\|w)$	Baseline	AdvCls (1,1)	AdvDat (0.4,1)	AdvDat (1,1)
sleeping	108	0.88	0.24	15.63	53.13	-81.25
driving	53	0.81	0.32	-8.33	50	-66.67
Nobody	52	1	0.42	14.29	42.86	14.29
alone	50	0.9	0.32	0	83.33	0
cat	49	0.84	0.31	7.14	57.14	-85.71
asleep	43	0.91	0.39	-18.75	50	12.5
no	31	0.84	0.36	0	52.94	-52.94
empty	28	0.93	0.3	-16.67	83.33	-16.67
eats	24	0.83	0.3	37.5	87.5	-25
naked	20	0.95	0.46	0	83.33	-33.33

Table 7.5: Indicator words and how correlated they are with CONTRADICTION predictions. The parentheses indicate hyper-parameter values: $(\lambda_{\text{Loss}}, \lambda_{\text{Enc}})$ for AdvCls and $(\lambda_{\text{Rand}}, \lambda_{\text{Enc}})$ for AdvDat. Baseline refers to the unmodified InferSent.

applying them to downstream tasks. We presented a solution for combating annotation biases by proposing two adversarial learning based methods. When empirically evaluating our approaches, we found that in a synthetic setting, as well as on a wide-range of existing RTE datasets, our methods perform better than the traditional training method to predict a label given a premise-hypothesis pair. Furthermore, we performed several analyses into the interplay of our methods with known biases in RTE datasets, the interplay with known biases, the effects of stronger bias removal, and the empirical probability the models assign a label based on a single indicator word.

We were not the first to apply adversarial learning to RTE. Minervini and Riedel (2018) generate adversarial examples that do not conform to logical rules and regularize models based on those examples. Similarly, Kang et al. (2018) incorporate external linguistic resources and use a GAN-style framework to adversarially train robust RTE models. We similarly use adversarial learning to train RTE models that are robust to biases. However, in contrast, we do not use external resources and we are interested in mitigating hypothesis-only biases when training RTE models.

Concurrently to our work appearing at StarmSem 2019 and ACL 2019, Grand and Belinkov (2019) received a Best Paper Award at the Workshop on Shortcomings in Vision and Language (SiVL) for their work on exploring adversarial learning for overcoming biases in Visual Question Answering, the task of answering questions based on a given image. Subsequently, others used similar methods to overcome biases when training models for RTE or related tasks (He, Zha, and Wang, 2019; Clark, Yatskar, and Zettlemoyer, 2019; Chen et al., 2020a; Chang et al., 2020; Thorne and Vlachos, 2020). Inspired by our work, Stacey et al. (2020) use an *ensemble of adversaries* to overcome biases and Karimi Mahabadi, Belinkov, and Henderson (2020) propose training with a product of experts or debiased focal loss for developing robust RTE models.

Chapter 8

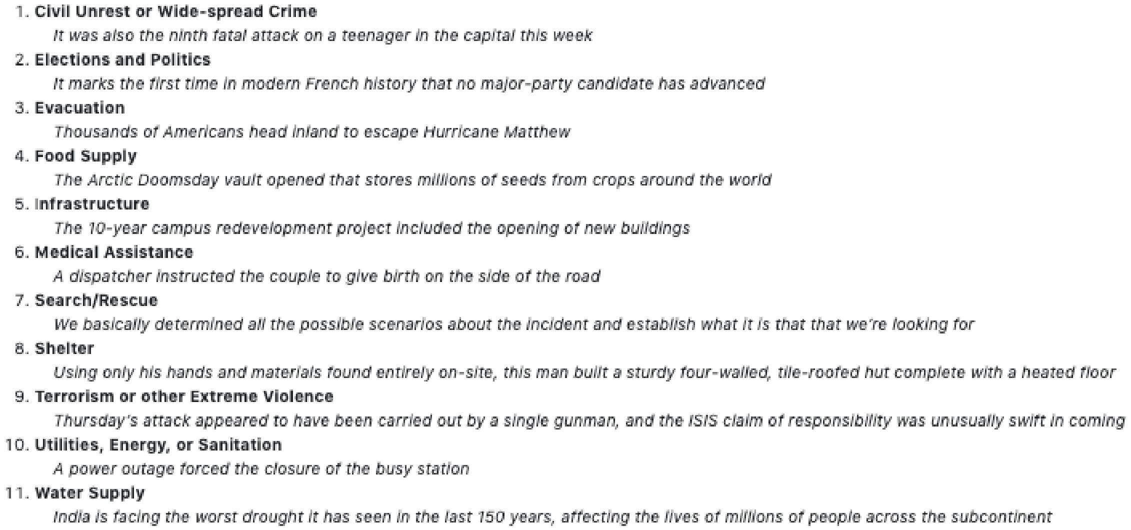
Adversarial Learning for Emergency Need Discovery

As RTE encompasses semantic inference that is necessary for many NLP tasks, we now turn our attention to applying adversarial learning to overcome domain-specific biases in an applied setting. The real world task we explore is discovering emergency needs during disaster events. Developing technologies to discover emergency needs in low resource settings is vital for effectively providing aid during disastrous events. In emergency scenarios with limited time and resources, humans may not be able to quickly scan incoming texts and SOSs. NLP models might help with identifying, classifying, and prioritizing distress signals. In low resource and time-sensitive settings, supervised data for training such models is sparse and human annotators might be hard to find. Furthermore, distributions of needs might not be consistent across different emergency scenarios, and populations in varying emergency scenarios may use distinct vocabulary or phrases to express the same need. In turn, applying models across multiple emergency scenarios might be disadvantageous.

Based on our experience using adversarial learning to overcome domain- and dataset-specific biases in RTE, we apply adversarial learning to the task of discovering emergency needs in low resource settings. When training a classifier to predict whether and which type of emergency need is expressed in a text, we force our model to predict which disaster occurred. Adversarial learning, implemented through a gradient reversal described in Section 7.2, penalizes our model when correctly predicting the disaster that occurred. We hypothesize that this may force our networks to generalize well across different disaster scenarios.

8.1 DARPA LORELEI Challenge

This work is motivated by the DARPA Low Resource Languages for Emergent Incidents (LORELEI) program (Christianson, Duncan, and Onyshkevych, 2018). Spanning over four years, research groups competed to develop technologies that could be deployed within short time frames after the emergence of an unknown disaster in a surprise language. One day, one week, and one month after the surprise language and disaster was announced, researchers' systems were evaluated on the DARPA-internal LORELEI tasks.

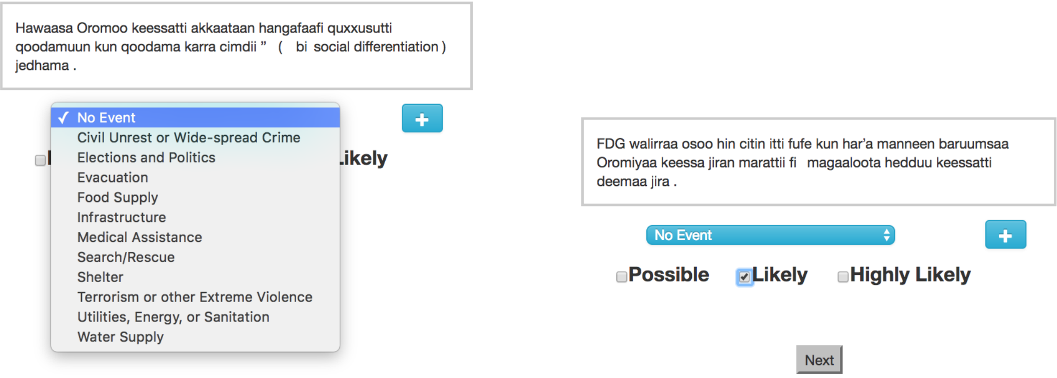


1. **Civil Unrest or Wide-spread Crime**
   It was also the ninth fatal attack on a teenager in the capital this week
2. **Elections and Politics**
   It marks the first time in modern French history that no major-party candidate has advanced
3. **Evacuation**
   Thousands of Americans head inland to escape Hurricane Matthew
4. **Food Supply**
   The Arctic Doomsday vault opened that stores millions of seeds from crops around the world
5. **Infrastructure**
   The 10-year campus redevelopment project included the opening of new buildings
6. **Medical Assistance**
   A dispatcher instructed the couple to give birth on the side of the road
7. **Search/Rescue**
   We basically determined all the possible scenarios about the incident and establish what it is that that we're looking for
8. **Shelter**
   Using only his hands and materials found entirely on-site, this man built a sturdy four-walled, tile-roofed hut complete with a heated floor
9. **Terrorism or other Extreme Violence**
   Thursday's attack appeared to have been carried out by a single gunman, and the ISIS claim of responsibility was unusually swift in coming
10. **Utilities, Energy, or Sanitation**
    A power outage forced the closure of the busy station
11. **Water Supply**
    India is facing the worst drought it has seen in the last 150 years, affecting the lives of millions of people across the subcontinent

Figure 8.1: Examples for each of the Situation Frame labels

Here, we focus on the task of detecting emergency needs in text. These needs are referred to as situation frames, which "are structured representations of key elements gleaned from English or foreign language text and speech" (Christianson, Duncan, and Onyshkevych, 2018). The types of situation frames we consider are Civil Unrest or Wide-spread Crime, Elections and Politics, Evacuation, Food Supply, Infrastructure, Medical Assistance, Search and Rescue, Shelter, Terrorism or other Extreme Violence, Utilities Energy or Sanitation, and Water Supply. Figure 8.1 includes examples that express each of these emergency needs.

Within each short time frame of the evaluation, teams were able to interact with a Native Informant (NI). These are lay people who speak the surprise language. During limited hour-long sessions, we relied on a NI to annotate which situation frame occurred in select examples. Throughout the years, we used multiple annotation frameworks, including the Computer Assisted Discovery Extraction and Translation



Hawaasa Oromoo keessatti akkaataan hangafaafi quxxusutti qoodamuun kun qoodama karra cimdii " (bi social differentiation) jedhama .

(a)

FDG walirraa osoo hin citin itti fufe kun har'a manneen baruumsaa Oromiyaa keessa jiran marattii fi magaaloota hedduu keessatti deemaa jira .

No Event

□ Possible ☑ Likely □ Highly Likely

(b)

Figure 8.2: Annotation interface in CADET for Native Informants to annotated examples with correct situation frame labels. 8.2a shows the list of situation frames a NI can apply to the sentence and 8.2b shows the likelihood the NI can apply to the situation frame label.

(CADET) workbench (Van Durme et al., 2017).[1] Figure 8.2 illustrates the user interface for NIs to provide annotations using CADET.

Initially developed at the Human Language Technology Center of Excellence's 2016 Summer Camp for Applied Language Exploration program (SCALE),[2] CADET is a workbench for rapid discovery, annotation, and extraction on text. CADET includes an active learning component where a model is used to determine which are the most informative data points to label (Settles, 1995). Based upon new annotations, the model can be updated in real-time to update and re-prioritize the list of unlabeled data points to be annotated. This notion is very similar to the idea of *Machine Teaching* (Simard et al., 2017) and can be deployed to help a domain expert work more efficiently (Sunkle, Kholkar, and Kulkarni, 2016; Gooding and Briscoe, 2019).

[1] `https://github.com/hltcoe/cadet`

[2] `https://hltcoe.jhu.edu/research/scale/scale-2016/`

During the LORELEI evaluations, between each NI session, we used this framework to update the list of examples for NI's to annotate.

8.2 Methods

Baseline

For each emergency need $n \in \mathcal{N}$, a pre-defined set of possible needs (Figure 8.1), we train a binary classifier to predict whether n is expressed in sentence s. Each binary classifier consists of a Bi-LSTM encoder $g(s)$ that maps each sentence s to a vector representation v_s, and a MLP $f_n(v_s)$ that predicts whether n is expressed in s. To deal with large class imbalances due to that fact that most texts do not express an emergency need, we weight our loss function, specifically cross-entropy, based on the class imbalance of the training set. Our loss function for each binary classifier is $\mathcal{L}_n = \mathcal{L}(f_n(v_s), y)$, where y is a boolean indicating whether emergency need n is expressed in s.

Applying Adversarial Learning

Since each emergency situation may have different distributions of emergency needs and the needs may be expressed differently in different situations, applying these binary classifiers across events may not work well. During adversarial training,

Event	Date
2016 OromoProtest	2013-10-12 T15:09:30Z
2011 NabroEruption	2011-06-17 T15:49:19Z
2013 Iran Earthquake	2013-04-16 T12:26:11Z
2013 India Cyclone	2013-10-12 T15:09:30Z
2015 Paris Attacks	2015-11-14 T12:43:44Z
2014 Turkey Flash Floods	2014-09-02 T08:51:10Z
2011 EastAfricaDroughts	2011-07-21 T10:18:31Z
2013 EgyptCoupD'état	2013-04-16 T12:26:11

Table 8.1: Each disaster included in our dataset.

we additionally feed v_s to a new MLP $f_{situation}$ that predicts which disastrous event e occurred. We modify the loss function of our network to become $\mathcal{L} = \mathcal{L}_n + \lambda \mathcal{L}_{\text{Adv}}$, where $\mathcal{L}_{\text{Adv}} = \mathcal{L}(f_{situation}(\lambda_{enc} GRL(g(s))), e)$. λ and λ_{enc} respectively control the weight of the adversarial loss function and the gradient reversal to $g(s)$. We do not perturb the training data like in AdvDat as our results for transferring across RTE datasets using AdvDat varied more widely than when using AdvCls.

8.3 Experiments

Data

We use tweets associated with 8 disaster situations in the past ten years, that were internally annotated with the 11 situation frame emergency needs (Figure 8.1). using the EASL framework (Sakaguchi and Van Durme, 2018). Table 8.1 provides details regarding the disaster situations included in our dataset and Table 8.2 include sampled tweets. To test our hypothesis, we use a leave-one-out setup where we train

Event	Tweet
2013 Iran Earthquake	@USER but the actually earthquake happened in iran i feel bad for them omg Iran-Pakistan Border... 8 on Richter scale.. Praying there is no loss of life! #EarthQuake
2015 Paris Attacks	Sitting here reading reports whilst feeding my 6wk old baby wondering what kind of world have I brought her into :(#ParisAttacks Cousin of French international footballer Lassana Diarra was killed in the Paris attacks
2011 EastAfricaDroughts	Tens of thousands feared dead in south #Somalia famine http://t.co/NXczq4n via @globeandmail Please help your fellow people in #Africa. #Donate to an organisation who is active in the drought areas. #Fight the #famine.

Table 8.2: Example tweets sampled from the crowdsourced dataset.

our baseline & adversarially-trained binary classifiers on all but one disaster event and test on the held-out event. We repeat this process for all 8 events collected.

Results

Table 8.3 reports the difference in F1, accuracy, precision, and recall between the best performing adversarial model and the baseline model for each emergency need and disaster event.[3] In 42 of the 88 settings, we see no difference (in F1 score) between the best performing adversarial model vs the baseline. In one case, the best performing adversarial model does slightly worse than the baseline, and in the remaining 43 examples, the best adversarial model outperforms the baseline in F1.

[3] Here, we determine the best performing adversarial model based on F1.

		violence	food	utils	infra	water	shelter	regimechange	evac	terrorism
F1	2016 OromoProtest	0.00	0.00	0.34	0.00	0.19	0.00	2.66	0.08	0.08
	2011 NabroEruption	0.12	0.00	0.28	0.00	0.13	0.00	0.00	0.14	0.44
	2013 Iran Earthquake	0.00	2.37	0.00	0.72	0.00	0.63	4.34	0.00	0.20
	2013 India Cyclone	1.50	0.00	0.10	0.00	0.00	0.60	0.00	0.42	0.00
	2015 Paris Attacks	0.01	0.34	0.63	1.00	0.00	0.00	10.02	0.58	0.24
	2014 Turkey Flash Floods	0.00	0.00	0.27	0.02	1.68	0.18	0.00	3.54	0.00
	2011 EastAfricaDroughts	1.14	0.01	0.94	0.03	0.06	0.03	0.00	0.00	0.00
	2013 EgyptCoupD'état	0.01	0.00	0.68	0.72	1.88	0.00	0.10	-0.06	0.00
accuracy	2016 OromoProtest	0.00	0.00	5.61	0.00	0.45	0.65	2.54	0.04	0.00
	2011 NabroEruption	1.06	0.00	3.87	0.00	0.97	0.00	0.00	0.39	1.79
	2013 Iran Earthquake	0.00	0.08	0.04	0.04	0.00	0.96	9.25	1.08	1.95
	2013 India Cyclone	3.16	0.00	0.77	0.00	0.00	3.41	0.00	3.20	0.00
	2015 Paris Attacks	0.04	0.61	0.08	1.63	0.00	0.00	6.33	0.08	8.64
	2014 Turkey Flash Floods	0.22	0.00	0.89	0.06	0.39	8.42	0.00	0.00	0.00
	2011 EastAfricaDroughts	2.53	0.09	1.29	0.04	0.04	1.11	0.00	0.04	0.00
	2013 EgyptCoupD'état	0.12	0.00	0.00	0.12	0.08	0.00	0.08	-0.04	0.49
precision	2016 OromoProtest	0.00	0.00	4.47	0.00	2.25	2.89	14.01	0.11	0.05
	2011 NabroEruption	0.10	0.00	4.77	0.00	0.11	0.00	0.00	0.43	0.35
	2013 Iran Earthquake	0.00	5.88	8.33	0.28	0.00	0.63	15.40	0.31	0.20
	2013 India Cyclone	5.22	0.00	0.11	0.00	0.00	0.62	0.00	0.60	0.00
	2015 Paris Attacks	0.01	4.18	0.85	0.96	0.00	0.00	19.31	0.93	1.00
	2014 Turkey Flash Floods	0.39	0.00	0.19	0.40	6.94	1.68	0.00	0.00	0.00
	2011 EastAfricaDroughts	1.78	1.39	1.08	0.11	0.15	0.88	0.00	0.00	0.00
	2013 EgyptCoupD'état	0.01	0.00	0.41	0.97	2.02	0.00	4.09	-0.20	2.75
recall	2016 OromoProtest	0.00	0.00	0.39	0.00	0.00	0.00	0.00	0.00	0.36
	2011 NabroEruption	0.00	0.00	0.00	0.00	4.84	0.00	0.00	0.00	0.37
	2013 Iran Earthquake	0.00	1.36	0.00	1.25	0.00	0.64	0.00	0.00	0.00
	2013 India Cyclone	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2015 Paris Attacks	13.73	0.35	0.50	1.55	0.00	0.00	0.00	0.43	-0.32
	2014 Turkey Flash Floods	0.00	0.00	0.85	0.00	0.90	0.00	0.00	2.20	0.00
	2011 EastAfricaDroughts	22.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2013 EgyptCoupD'état	0.00	0.00	0.84	0.57	1.04	0.00	0.00	0.00	0.00

Table 8.3: Each row indicates the held-out event and each column represents the emergency need predicted. Numbers represent the difference in accuracy between the best performing model for each setting and the corresponding baseline binary classifier. The first column indicates the difference in which metric is reported. Note that for a given disaster event and SF type, the numbers do not correspond to the same model.

Inspecting our results, the binary classifiers that predict whether a tweet mentions a *"search"* or *"med"* need achieve the same F1 score as the baseline in all but one case.[4] When we test on the 2013 Iranian Earthquake, the best performing model for *"med"* outperforms the baseline by 0.04 in accuracy and 3.91 in recall. While each emergency need is often expressed in less than 50% of the tweets, it should be

[4] Therefore we do not include these in Table 8.3.

Event	Tweet	NEED	Baseline	Adv
2015 Paris Attacks	#ParisAttacks a specific group or a person cannot represent Islam.We Muslims r peace loving. There is no room for terrorism in our religion	regime change	✓	✗
2014 Turkey Flash Floods	Where is the water going to go if they cover what is left of Somerset Levels in solar panels	water	✗	✓

Table 8.4: Examples where the adversarial model correctly predicted whether the SF type applied to the tweet. ✓ and ✗ respectively indicate whether the model classified the need as applying or not applying to the tweet. In these examples, the baseline's predictions were incorrect and the adversarial model's predictions were correct.

noted that these two needs each appear in less than 10% of the training examples. For the *"regimechange"* need during the 2015 Paris Attack, we notice a 10+ absolute F1 improvement. Upon inspection, the model correctly predicted that a significantly smaller number of tweets represented a *"regimechange"* need compared to the baseline model's predictions. Further manual expection of the results indicates that the number of true negative predictions often increase for the best performing models.

Table 8.4 includes qualitative examples where the adversarial model correctly predicts whether the need is expressed in the text while the baseline model's prediction is incorrect. The second tweet in the table demonstrates examples of noise in this dataset. While the event likely describes issues related to flooding and water, this tweet does not describe the flash floods in Turkey in 2014 since the Somerset Levels are in Somerset England.

8.4 Discussion

In this chapter we demonstrated how adversarial learning can be applied to sentence classification models for real-world, applied tasks. In the 2019 LORELEI evaluation, our ensemble approach combined predictions from the adversarial trained model with other models developed by teammates (Yuan et al., 2019; Zhang, Fujinuma, and Boyd-Graber, 2020). Using adversarial learning resulted in a winning submission during the 2019 LORELEI challenge.

Chapter 9

Conclusion

9.1 Contributions

This thesis has made contributions to Recognizing Textual Entailment (RTE), primarily as an evaluation framework for exploring how well NLP systems capture a wide range of semantic phenomena. Chapter 2 delved into different NLP evaluations paradigms and provides motivation for why now is an ideal time to revisit RTE as a method to evaluate NLP systems. In Chapter 3, we introduced the Diverse Natural Language Inference Collection (DNC), a large scale test suite of RTE datasets that cover a wide range of linguistic phenomena. Next, we presented a general framework for using the DNC and related datasets to explore the reasoning capabilities of NLP systems. Chapter 4 included a thorough study exploring how well sentence encoders trained as part of a neural machine translation system capture phenomena such as paraphrastic inference, anaphora resolution, and semantic proto-roles. We also used this general framework to explore models trained to perform other tasks, including connecting images with captions, syntactic parsing, and discourse marking.

In Chapter 5, we discovered hypothesis-only biases in a large number of RTE datasets. This work inspired similar discoveries in the field as researchers began to question long held assumptions about gold-standard dataset. This work also enabled us to explore the limits of RTE as an evaluation framework.

Since many tasks require systems to perform semantic inferences similar to RTE, researchers often pre-train models on RTE datasets. In Chapter 7 we argued that hypothesis-only biases from Chapter 5 might negatively impact this common approach. We demonstrated how adversarial learning can be incorporated when training RTE models to be robust to such biases. Finally, we applied the lessons learned to incorporate adversarial learning to train models to detect emergency needs in disaster events.

9.2 Future Work

As this thesis focused on revisiting RTE as an evaluation framework, we raise opportunities for future research directions. As pointed out in Staliūnaitė (2018)'s master' thesis, the recasting methods introduced in Chapter 3 can be further refined and improved. Additionally, as the DNC only begins to scratch the surface of semantic phenomena that are important for understanding human language, there are many opportunities to add more semantic phenomena to the DNC.

In Chapter 4, we used the DNC to evaluate whether the sentence-representations extracted from the final layer of encoders capture different semantic phenomena. Future research can investigate whether these phenomena are captured at different layers of an encoder or even in specific neurons of a neural network. Additionally, exploring how attention-based mechanism capture these phenomena is an important open research question as many NLP models rely on attention mechanism and there is ongoing debate in the community regarding the interpretability of attention-based mechanisms.

Our methodology in Chapter 7 for using adversarial learning to develop robust RTE models can be extended to handle biases in other tasks where one is concerned with finding relationships between two objects, such as visual question answering, story cloze completion, and reading comprehension. We hope to encourage such investigation in the broader community.

Finally, the relationship between how well a model captures different semantic phenomena and how well a model performs on a downstream task remains an open question. As Vázquez et al. (2020) note, "it is not fully clear how the properties of the fixed-sized vector influence the tradeoff between the performance of the model in MT and the information it encodes as a meaning representation vector." This idea is relevant not just to machine translation but to all NLP systems. Broadly, this question can be viewed as asking how well do test suite evaluations correlate with either intrinsic or extrinsic evaluations? This is an important question that remains open in the field and is vital for developing efficient and effective NLP evaluation frameworks.

Bibliography

Abend, Omri and Ari Rappoport (2013). “Universal conceptual cognitive annotation (ucca)”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 228–238.

Fine-grained analysis of sentence embeddings using auxiliary prediction tasks (2017).

Adler, Meni, Jonathan Berant, and Ido Dagan (2012). “Entailment-based Text Exploration with Application to the Health-care Domain”. In: *Association for Computational Linguistics*, p. 79.

Agirre, Eneko, Lluís Màrquez, and Richard Wicentowski, eds. (June 2007). *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/S07-1000.

Agosti, Maristella, Giorgio Maria Di Nunzio, Nicola Ferro, Donna Harman, and Carol Peters (2007). “The Future of Large-scale Evaluation Campaigns for Information Retrieval in Europe”. In: *International Conference on Theory and Practice of Digital Libraries*. Springer, pp. 509–512.

BIBLIOGRAPHY

Aharon, Roni Ben, Idan Szpektor, and Ido Dagan (2010). "Generating entailment rules from framenet". In: *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, pp. 241–246.

Allen, James (1995). *Natural language understanding*. Pearson.

Amoia, Marilisa (Nov. 2008). "Linguistic-Based Computational Treatment of Textual Entailment Recognition". Theses. Université Henri Poincaré - Nancy 1. URL: `https://hal.univ-lorraine.fr/tel-01748535`.

Andreas, Jacob and Dan Klein (2017). "Analogs of Linguistic Structure in Deep Representations". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2893–2897.

Androutsopoulos, Ion and Prodromos Malakasiotis (2010). "A survey of paraphrasing and textual entailment methods". In: *Journal of Artificial Intelligence Research* 38, pp. 135–187.

Avramidis, Eleftherios, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit (Aug. 2019). "Linguistic Evaluation of German-English Machine Translation Using a Test Suite". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pp. 445–454. URL: `https://www.aclweb.org/anthology/W19-5351`.

BIBLIOGRAPHY

Aziz, Wilker, Marc Dymetman, Shachar Mirkin, Lucia Specia, Nicola Cancedda, and Ido Dagan (2010). "Learning an Expert from Human Annotations in Statistical Machine Translation: the Case of Out-of-VocabularyWords". In: *14th Annual Meeting of the European Association for Machine Translation (EAMT)*. Saint-Rapha, France.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Baker, Collin F, Charles J Fillmore, and John B Lowe (1998). "The berkeley framenet project". In: *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 86–90.

Baker, Mona (2018). *In other words: A coursebook on translation*. Routledge.

Baker, Simon, Roi Reichart, and Anna Korhonen (Oct. 2014). "An Unsupervised Model for Instance Level Subcategorization Acquisition". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 278–289. URL: https://www.aclweb.org/anthology/D14-1034.

Balkan, Lorna, Siety Meijer, Doug Arnold, Eva Dauphin, Dominique Estival, Kirsten Falkedal, Sabine Lehmann, and Sylvie Regnier-Prost (1994). "Test Suite Design Guidelines and Methodology". In:

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider

(2013). "Abstract meaning representation for sembanking". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186.

Bangalore, Srinivas and Aravind K. Joshi (1999). "Supertagging: An Approach to Almost Parsing". In: *Computational Linguistics* 25.2, pp. 237–265. URL: `https://www.aclweb.org/anthology/J99-2004`.

Bar-Haim, Roy, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, and Bernardo Magnini (2006). "The Second PASCAL Recognising Textual Entailment Challenge". In:

Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow (2017). "Evaluating Discourse Phenomena in Neural Machine Translation". In: *arXiv preprint arXiv:1711.00513*.

Bekinschtein, Tristan A, Matthew H Davis, Jennifer M Rodd, and Adrian M Owen (2011). "Why clowns taste funny: the relationship between humor and semantic ambiguity". In: *Journal of Neuroscience* 31.26, pp. 9665–9671.

Belinkov, Yonatan (2018). "On internal language representations in deep learning: An analysis of machine translation and speech recognition". PhD thesis. Massachusetts Institute of Technology.

Belinkov, Yonatan and Yonatan Bisk (2018). "Synthetic and Natural Noise Both Break Neural Machine Translation". In: *International Conference on Learning Representations*. URL: `https://openreview.net/forum?id=BJ8vJebC-`.

BIBLIOGRAPHY

Belinkov, Yonatan and James Glass (2017). "Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 2441–2451. URL: `http://papers.nips.cc/paper/6838-analyzing-hidden-representations-in-end-to-end-automatic-speech-recognition-systems.pdf`.

Belinkov, Yonatan and James Glass (2019). "Analysis Methods in Neural Language Processing: A Survey". In: *Transactions of the Association for Computational Linguistics* 7, pp. 49–72. eprint: `https://doi.org/10.1162/tacl_a_00254`. URL: `https://doi.org/10.1162/tacl_a_00254`.

Belinkov, Yonatan, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass (2017a). "Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 1–10.

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass (2017b). "What do Neural Machine Translation Models Learn about Morphology?" In: *Proceedings of the 55th Annual Meeting of the Association for Computa-*

tional Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, pp. 861–872.

Belinkov, Yonatan, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush (July 2019a). “Don’t Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 877–891. URL: `https://www.aclweb.org/anthology/P19-1084`.

Belinkov, Yonatan, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush (June 2019b). “On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference”. In: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 256–262. URL: `https://www.aclweb.org/anthology/S19-1028`.

Belz, Anja and Albert Gatt (June 2008). “Intrinsic vs. Extrinsic Evaluation Measures for Referring Expression Generation”. In: *Proceedings of ACL-08: HLT, Short Papers*. Columbus, Ohio: Association for Computational Linguistics, pp. 197–200. URL: `https://www.aclweb.org/anthology/P08-2050`.

Ben-David, Shai, John Blitzer, Koby Crammer, and Fernando Pereira (2007). “Analysis of Representations for Domain Adaptation”. In: *Advances in Neural Information Processing Systems 19*. Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman.

MIT Press, pp. 137–144. URL: http://papers.nips.cc/paper/2983-analysis-of-representations-for-domain-adaptation.pdf.

Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan (2010). “A theory of learning from different domains”. In: *Machine learning* 79.1-2, pp. 151–175.

Bentivogli, Luisa, Ido Dagan, and Bernardo Magnini (2017). “The Recognizing Textual Entailment Challenges: Datasets and Methodologies”. In: *Handbook of Linguistic Annotation*. Ed. by Nancy Ide and James Pustejovsky. Dordrecht: Springer Netherlands, pp. 1119–1147. URL: https://doi.org/10.1007/978-94-024-0881-2_42.

Bentivogli, Luisa, Peter Clark, Ido Dagan, and Danilo Giampiccolo (2011). “The Seventh PASCAL Recognizing Textual Entailment Challenge.” In: *Textual Analysis Conference (TAC)*.

Berant, Jonathan (2012). “Global Learning of Textual Entailment Graphs”. PhD thesis. Bar Ilan University.

Beutel, Alex, Jilin Chen, Zhe Zhao, and Ed H Chi (2017). “Data decisions and theoretical implications when adversarially learning fair representations”. In: *arXiv preprint arXiv:1707.00075*.

Bhagat, Rahul and Eduard Hovy (2013). “What is a paraphrase?” In: *Computational Linguistics* 39.3, pp. 463–472.

Bhagavatula, Chandra, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi (2020). "Abductive Commonsense Reasoning." In: *ICLR*. URL: https://openreview.net/forum?id=Byg1v1HKDB.

Binsted, Kim (1996). "Machine humour: An implemented model of puns". PhD thesis. Edinburgh, Scotland: University of Edinburgh. URL: http://www2.hawaii.edu/~binsted/papers/Binstedthesis.pdf.

Birch, Alexandra, Omri Abend, Ondřej Bojar, and Barry Haddow (Nov. 2016). "HUME: Human UCCA-Based Evaluation of Machine Translation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1264–1274. URL: https://www.aclweb.org/anthology/D16-1134.

Bird, Steven and Edward Loper (2004). "NLTK: the natural language toolkit". In: *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, p. 31.

Bisk, Yonatan, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi (2020). "PIQA: Reasoning about Physical Commonsense in Natural Language". In: *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Blake, Catherine (2007). "The Role of Sentence Structure in Recognizing Textual Entailment". In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. RTE '07. Prague, Czech Republic: Association for Computational Linguistics, pp. 101–106. URL: http://dl.acm.org/citation.cfm?id=1654536.1654557.

Blasband, M, N Bevan, M King, B Maegaard, L des Tombe, S Krauwer, S Manzi, and N Underwood (1999). "Expert advisory group on language engineering standards/evaluation working group final report 2". In:

Bojar, Ondrej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna (2014). "Findings of the 2014 Workshop on Statistical Machine Translation". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 12–58.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Advances in Neural Information Processing Systems*, pp. 4349–4357.

Bos, Johan, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva (2017). "The groningen meaning bank". In: *Handbook of Linguistic Annotation*. Springer, pp. 463–496.

Bosma, Wauter and Chris Callison-Burch (2006). "Paraphrase substitution for recognizing textual entailment". In: *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, pp. 502–509.

Bowman, Samuel (2016). “MODELING NATURAL LANGUAGE SEMANTICS IN LEARNED REPRESENTATIONS”. PhD thesis. Stanford University.

Bowman, Samuel, Yoav Goldberg, Felix Hill, Angeliki Lazaridou, Omer Levy, Roi Reichart, and Anders Søgaard, eds. (Sept. 2017). *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Copenhagen, Denmark: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W17-5300.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning (2015). “A large annotated corpus for learning natural language inference”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Brad, Florin and Traian Rebedea (Sept. 2017). “Neural Paraphrase Generation using Transfer Learning”. In: *Proceedings of the 10th International Conference on Natural Language Generation*. Santiago de Compostela, Spain: Association for Computational Linguistics, pp. 257–261. URL: https://www.aclweb.org/anthology/W17-3542.

Bras, Ronan Le, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi (2020). *Adversarial Filters of Dataset Biases*. arXiv: 2002.04108 [cs.LG].

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran (2012). “Distributional semantics in technicolor”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 136–145.

Bugert, Michael, Yevgeniy Puzikov, Andreas Rücklé, Judith Eckle-Kohler, Teresa Martin, Eugenio Martínez-Cámara, Daniil Sorokin, Maxime Peyrard, and Iryna Gurevych (Apr. 2017). “LSDSem 2017: Exploring Data Generation Methods for the Story Cloze Test”. In: *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. Valencia, Spain: Association for Computational Linguistics, pp. 56–61. URL: `https://www.aclweb.org/anthology/W17-0908`.

Cai, Zheng, Lifu Tu, and Kevin Gimpel (2017). “Pay Attention to the Ending: Strong Neural Baselines for the ROC Story Cloze Task”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Callison-Burch, Chris (2007). “Paraphrasing and Translation”. PhD thesis. Edinburgh, Scotland: University of Edinburgh. URL: `http://cis.upenn.edu/~ccb/publications/callison-burch-thesis.pdf`.

Camburu, Oana-Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom (2018). “e-SNLI: Natural Language Inference with Natural Language Explanations”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., pp. 9539–9549. URL: `http://papers.nips.cc/paper/`

8163-e-snli-natural-language-inference-with-natural-language-explanations.pdf.

Caporaso, J Gregory, Nita Deshpande, J Lynn Fink, Philip E Bourne, K Bretonnel Cohen, and Lawrence Hunter (2008). “Intrinsic evaluation of text mining tools may not predict performance on realistic tasks”. In: *Biocomputing 2008*. World Scientific, pp. 640–651.

Carpuat, Marine, Yogarshi Vyas, and Xing Niu (2017). “Detecting Cross-Lingual Semantic Divergence for Neural Machine Translation”. In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, pp. 69–79. URL: http://aclweb.org/anthology/W17-3209.

Caruana, Rich (1993). “Multitask Learning: A Knowledge-Based Source of Inductive Bias”. In: *ICML*.

Caruana, Rich (1997). “Multitask learning”. In: *Machine learning* 28.1, pp. 41–75.

Castillo, Julio Javier and Laura Alonso Alemany (2008). “An approach using named entities for recognizing textual entailment”. In: *Notebook Papers of the Text Analysis Conference, TAC Workshop*.

Chan, Seng Yee, Tou Hwee Ng, and David Chiang (2007). “Word Sense Disambiguation Improves Statistical Machine Translation”. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 33–40.

Chang, Shiyu, Yang Zhang, Mo Yu, and Tommi S Jaakkola (2020). “Invariant rationalization”. In: *arXiv preprint arXiv:2003.09772*.

Chang, Tyler A. and Anna N. Rafferty (2020). “Encodings of Source Syntax: Similarities in NMT Representations Across Target Languages”. In:

Chatzikyriakidis, Stergios, Robin Cooper, Simon Dobnik, and Staffan Larsson (2017). “An overview of Natural Language Inference Data Collection: The way forward?” In: *Proceedings of the Computing Natural Language Inference Workshop*. URL: http://aclweb.org/anthology/W17-7203.

Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson (2013). “One billion word benchmark for measuring progress in statistical language modeling”. In: *arXiv preprint arXiv:1312.3005*.

Chen, Long, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang (2020a). “Counterfactual samples synthesizing for robust visual question answering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10800–10809.

Chen, Qian, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen (2017a). “Enhanced LSTM for Natural Language Inference”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1657–1668. URL: http://www.aclweb.org/anthology/P17-1152.

Chen, Qian, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen (Sept. 2017b). "Recurrent Neural Network-Based Sentence Encoder with Gated Attention for Natural Language Inference". In: *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 36–40. URL: https://www.aclweb.org/anthology/W17-5307.

Chen, Qian, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei (July 2018a). "Neural Natural Language Inference Models Enhanced with External Knowledge". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2406–2417. URL: https://www.aclweb.org/anthology/P18-1224.

Chen, Tongfei, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme (2020b). "Uncertain Natural Language Inference". In: *ACL*.

Chen, Xilun, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger (2018b). "Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification". In: *Transactions of the Association for Computational Linguistics* 6, pp. 557–570. URL: https://www.aclweb.org/anthology/Q18-1039.

Cheng, Jianpeng, Li Dong, and Mirella Lapata (2016). "Long Short-Term Memory-Networks for Machine Reading". In: *Proceedings of the 2016 Conference on Em-*

pirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, pp. 551–561.

Chinchor, Nancy (1991). “MUC-3 Linguistic Phenomena Test Experiment”. In: *Third Message Uunderstanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991*. URL: `https://www.aclweb.org/anthology/M91-1004`.

Chinchor, Nancy, Lynette Hirschman, and David D. Lewis (1993). “Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)”. In: *Computational Linguistics* 19.3, pp. 409–450. URL: `https://www.aclweb.org/anthology/J93-3001`.

Chiu, Billy, Anna Korhonen, and Sampo Pyysalo (Aug. 2016). “Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance”. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany: Association for Computational Linguistics, pp. 1–6. URL: `https://www.aclweb.org/anthology/W16-2501`.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio (2014). “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches”. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational Linguistics, pp. 103–111. URL: `http://www.aclweb.org/anthology/W14-4012`.

Choshen, Leshem and Omri Abend (Nov. 2019). “Automatically Extracting Challenge Sets for Non-Local Phenomena in Neural Machine Translation”. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 291–303. URL: `https://www.aclweb.org/anthology/K19-1028`.

Christianson, Caitlin, Jason Duncan, and Boyan Onyshkevych (June 2018). “Overview of the DARPA LORELEI Program”. In: *Machine Translation* 32.1-2, pp. 3–9. URL: `https://doi.org/10.1007/s10590-017-9212-4`.

Clark, Christopher, Mark Yatskar, and Luke Zettlemoyer (Nov. 2019). “Don’t Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 4069–4082. URL: `https://www.aclweb.org/anthology/D19-1418`.

Clark, Peter, Phil Harrison, and Niranjan Balasubramanian (2012). “Answering Biology Questions using Textual Reasoning”. In: *Pacific Northwest Regional NLP Workshop (NW-NLP 2012)*.

Clark, Peter, Phil Harrison, and Xuchen Yao (2012). “An Entailment-Based Approach to the QA4MRE Challenge”. In: *Proc. CLEF 2012 (Conference and Labs of the*

Evaluation Forum) - QA4MRE (Question Answering for Machine Reading Evaluation) Lab.

Clark, Peter, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D Turney, and Daniel Khashabi (2016). “Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions.” In: *AAAI*.

Clark, Peter, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, and Michael Schmitz (2019). “From ’F’ to ’A’ on the N.Y. Regents Science Exams: An Overview of the Aristo Project”. In: *ArXiv* abs/1909.01958.

Clark, Stephen (May 2002). “Supertagging for Combinatory Categorial Grammar”. In: *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+6)*. Universitá di Venezia: Association for Computational Linguistics, pp. 19–24. URL: https://www.aclweb.org/anthology/W02-2203.

Condoravdi, Cleo, Dick Crouch, Valeria De Paiva, Reinhard Stolle, and Daniel G Bobrow (2003). “Entailment, intensionality and text understanding”. In: *Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9*. Association for Computational Linguistics, pp. 38–45.

Conneau, Alexis and Douwe Kiela (May 2018). “SentEval: An Evaluation Toolkit for Universal Sentence Representations”. In: *Proceedings of the Eleventh Interna-*

tional Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://www.aclweb.org/anthology/L18-1269.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes (2017). "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 670–680. URL: https://www.aclweb.org/anthology/D17-1070.

Conneau, Alexis, GermÃ¡n Kruszewski, Guillaume Lample, LoÃc Barrault, and Marco Baroni (2018). "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2126–2136. URL: http://www.aclweb.org/anthology/P18-1198.

Conroy, John M., Hoa Trang Dang, Ani Nenkova, and Karolina Owczarzak, eds. (June 2012). *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*. Montréal, Canada: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W12-2600.

Cooper, Robin, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. (1996). “Using the framework”. In:

Dagan, Ido, Oren Glickman, and Bernardo Magnini (2006). “The PASCAL recognising textual entailment challenge”. In: *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment.* Springer, pp. 177–190.

Dai, Andrew M and Quoc V Le (2015). “Semi-supervised sequence learning”. In: *Advances in neural information processing systems*, pp. 3079–3087.

Dalvi, Fahim, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel (2017). “Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder”. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 142–151.

Dasgupta, I., D. Guo, A. Stuhlmüller, S. J. Gershman, and N. D. Goodman (Feb. 2018). “Evaluating Compositionality in Sentence Embeddings”. In: *ArXiv e-prints.* arXiv: `1802.04302 [cs.CL]`.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo-*

gies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. URL: `https://www.aclweb.org/anthology/N19-1423`.

Dowty, David (1991). “Thematic proto-roles and argument selection”. In: *language*, pp. 547–619.

Duan, Chaoqun, Lei Cui, Xinchi Chen, Furu Wei, Conghui Zhu, and Tiejun Zhao (July 2018). “Attention-Fused Deep Matching Network for Natural Language Inference”. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, pp. 4033–4040. URL: `https://doi.org/10.24963/ijcai.2018/561`.

Dzikovska, Myroslava, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang (2013). “SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge”. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 263–274. URL: `http://www.aclweb.org/anthology/S13-2045`.

Eichler, Max, Gözde Gül Şahin, and Iryna Gurevych (Nov. 2019). “LINSPECTOR WEB: A Multilingual Probing Suite for Word Representations”. In: *Proceedings of* *the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, pp. 127–132. URL: https://www.aclweb.org/anthology/D19-3022.

Elazar, Yanai and Yoav Goldberg (2018). "Adversarial Removal of Demographic Attributes from Text Data". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 11–21. URL: http://aclweb.org/anthology/D18-1002.

Estival, Dominique (1997). "Karen Sparck Jones & Julia R. Galliers, Evaluating Natural Language Processing Systems: An Analysis and Review. Lecture Notes in Artificial Intelligence 1083". In: *Machine Translation* 12.4, pp. 375–379.

Ettinger, Allyson, Ahmed Elgohary, Colin Phillips, and Philip Resnik (Aug. 2018). "Assessing Composition in Sentence Vector Representations". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1790–1801. URL: https://www.aclweb.org/anthology/C18-1152.

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer (Aug. 2016). "Problems With Evaluation of Word Embeddings Using Word Similarity Tasks". In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations* *for NLP*. Berlin, Germany: Association for Computational Linguistics, pp. 30–35. URL: https://www.aclweb.org/anthology/W16-2506.

Farzindar, Atefeh and Guy Lapalme (2004). "Letsum, an automatic legal text summarizing system". In: *Legal knowledge and information systems, JURIX*.

Feng, Shi, Eric Wallace, and Jordan Boyd-Graber (July 2019). "Misleading Failures of Partial-input Baselines". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5533–5538. URL: https://www.aclweb.org/anthology/P19-1554.

Ferraro, Francis, Matt Post, and Benjamin Van Durme (June 2012). "Judging Grammaticality with Count-Induced Tree Substitution Grammars". In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Montréal, Canada: Association for Computational Linguistics, pp. 116–121. URL: https://www.aclweb.org/anthology/W12-2013.

Ferraro, Francis, Benjamin Van Durme, and Matt Post (2012). "Toward tree substitution grammars with latent annotations". In: *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pp. 23–30.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin (2001). "Placing search in context: The concept revisited". In: *Proceedings of the 10th international conference on World Wide Web*, pp. 406–414.

Flickinger, Daniel, John Nerbonne, Ivan Sag, and Tom Wasow (1987). “Toward evaluation of NLP systems”. In: *Hewlett Packard Laboratories, Palo Alto, CA*.

Fonseca, Erick R. and Sandra Maria Aluísio (Nov. 2015). “Semi-Automatic Construction of a Textual Entailment Dataset: Selecting Candidates with Vector Space Models”. In: *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*. Natal, Brazil: Sociedade Brasileira de Computação, pp. 201–210. URL: `https://www.aclweb.org/anthology/W15-5624`.

Gabrilovich, Evgeniy, Michael Ringgaard, and Amarnag Subramanya (2013). “FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0)”. In: *Note: http://lemurproject. org/clueweb09/FACC1/Cited by* 5.

Gaizauskas, Robert (1998). “Evaluation in language and speech technology”. In: *Computer Speech & Language* 12.4, pp. 249–262.

Ganin, Yaroslav and Victor Lempitsky (2015). “Unsupervised Domain Adaptation by Backpropagation”. In: *International Conference on Machine Learning*, pp. 1180–1189.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky (2016). “Domain-adversarial training of neural networks”. In: *The Journal of Machine Learning Research* 17.1, pp. 2096–2030.

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch (2013). “PPDB: The paraphrase database”. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 758–764.

Gao, Qin and Stephan Vogel (2011). “Utilizing Target-Side Semantic Role Labels to Assist Hierarchical Phrase-based Machine Translation”. In: *Proceedings of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 107–115.

Giampiccolo, Danilo, Bernardo Magnini, Ido Dagan, and Bill Dolan (June 2007). “The Third PASCAL Recognizing Textual Entailment Challenge”. In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Prague: Association for Computational Linguistics, pp. 1–9. URL: https://www.aclweb.org/anthology/W07-1401.

Giannakopoulos, George, Elena Lloret, John M. Conroy, Josef Steinberger, Marina Litvak, Peter Rankel, and Benoit Favre, eds. (Apr. 2017). *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*. Valencia, Spain: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W17-1000.

Gildea, Daniel and Martha Palmer (2002). “The necessity of parsing for predicate argument recognition”. In: *Proceedings of the 40th Annual Meeting on Associa-*

tion for Computational Linguistics. Association for Computational Linguistics, pp. 239–246.

Glickman, Oren (2006). “Applied textual entailment”. PhD thesis. Bar Ilan University.

Glockner, Max, Vered Shwartz, and Yoav Goldberg (2018). “Breaking NLI Systems with Sentences that Require Simple Lexical Inferences”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 650–655. URL: `http://www.aclweb.org/anthology/P18-2103`.

Goldberg, Yoav (2017). *Neural network methods in natural language processing*. Morgan & Claypool Publishers.

Goldstein, Jade, Alon Lavie, Chin-Yew Lin, and Clare Voss, eds. (June 2005). *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics. URL: `https://www.aclweb.org/anthology/W05-0900`.

Gong, Yichen, Heng Luo, and Jian Zhang (2018). “Natural Language Inference over Interaction Space”. In: *International Conference on Learning Representations*.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2015). “Explaining and Harnessing Adversarial Examples”. In: *International Conference on Learning Representations (ICLR)*.

Gooding, Sian and Ted Briscoe (Sept. 2019). “Active Learning for Financial Investment Reports”. In: *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019).* Turku, Finland: Linköping University Electronic Press, pp. 25–32. URL: `https://www.aclweb.org/anthology/W19-6404`.

Grand, Gabriel and Yonatan Belinkov (2019). “Adversarial Regularization for Visual Question Answering: Strengths, Shortcomings, and Side Effects”. In: *Proceedings of the 2nd Workshop on Shortcomings in Vision and Language (SiVL) at NAACL-HLT.*

Gu, Jiatao, Zhengdong Lu, Hang Li, and Victor O.K. Li (2016). “Incorporating Copying Mechanism in Sequence-to-Sequence Learning”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Berlin, Germany: Association for Computational Linguistics, pp. 1631–1640.

Guo, Han, Ramakanth Pasunuru, and Mohit Bansal (2018a). “Dynamic Multi-Level Multi-Task Learning for Sentence Simplification”. In: *Proceedings of the 27th International Conference on Computational Linguistics.* Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 462–476. URL: `http://www.aclweb.org/anthology/C18-1039`.

Guo, Han, Ramakanth Pasunuru, and Mohit Bansal (2018b). “Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Lin-*

guistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, pp. 687–697. URL: http://www.aclweb.org/anthology/P18-1064.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith (2018). "Annotation Artifacts in Natural Language Inference Data". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 107–112. URL: http://www.aclweb.org/anthology/N18-2017.

Harabagiu, Sanda and Andrew Hickl (July 2006). "Methods for Using Textual Entailment in Open-Domain Question Answering". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, pp. 905–912. URL: https://www.aclweb.org/anthology/P06-1114.

Harabagiu, Sanda, Andrew Hickl, and Finley Lacatusu (2007). "Satisfying information needs with multi-document summaries". In: *Information Processing & Management* 43.6, pp. 1619–1642.

Hartshorne, Joshua K, Claire Bonial, and Martha Palmer (2013). "The VerbCorner project: Toward an empirically-based semantic decomposition of verbs". In: *Pro-*

ceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1438–1442.

He, He, Sheng Zha, and Haohan Wang (Nov. 2019). “Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual”. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, pp. 132–142. URL: `https://www.aclweb.org/anthology/D19-6115`.

Hill, Felix, Roi Reichart, and Anna Korhonen (Dec. 2015). “SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation”. In: *Computational Linguistics* 41.4, pp. 665–695. URL: `https://www.aclweb.org/anthology/J15-4004`.

Hockenmaier, Julia and Mark Steedman (2007). “CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank”. In: *Computational Linguistics* 33.3, pp. 355–396. URL: `https://www.aclweb.org/anthology/J07-3004`.

Hu, J. Edward, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme (June 2019). “Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Minneapolis, Minnesota: Association for Computational Linguistics, pp. 839–850. URL: https://www.aclweb.org/anthology/N19-1090.

Hupkes, Dieuwke, Sara Veldhoen, and Willem Zuidema (2018). "Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure". In: *Journal of Artificial Intelligence Research* 61, pp. 907–926.

Isabelle, Pierre, Colin Cherry, and George Foster (Sept. 2017). "A Challenge Set Approach to Evaluating Machine Translation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2486–2496. URL: https://www.aclweb.org/anthology/D17-1263.

Iyyer, Mohit, John Wieting, Kevin Gimpel, and Luke Zettlemoyer (June 2018). "Adversarial Example Generation with Syntactically Controlled Paraphrase Networks". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1875–1885. URL: https://www.aclweb.org/anthology/N18-1170.

Janssen, Theo M. V. (2011). "Montague Semantics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2011. Metaphysics Research Lab, Stanford University.

Javadpour, Seyedeh Leili (2013). "Resolving pronominal anaphora using commonsense knowledge". PhD thesis. Louisiana State University.

Jeretic, Paloma, Alex Warstadt, Suvrat Bhooshan, and Adina Williams (July 2020a). "Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8690–8705. URL: `https://www.aclweb.org/anthology/2020.acl-main.768`.

Jeretic, Paloma, Alex Warstadt, Suvrat Bhooshan, and Adina Williams (July 2020b). "Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8690–8705. URL: `https://www.aclweb.org/anthology/2020.acl-main.768`.

Jia, Robin and Percy Liang (Sept. 2017). "Adversarial Examples for Evaluating Reading Comprehension Systems". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2021–2031. URL: `https://www.aclweb.org/anthology/D17-1215`.

Joachims, Thorsten (1998). "Text categorization with support vector machines: Learning with many relevant features". In: *European conference on machine learning*. Springer, pp. 137–142.

Johnson, Melvin, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a ViÃ©gas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2017). “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation”. In: *Transactions of the Association for Computational Linguistics* 5, pp. 339–351.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov (Apr. 2017). “Bag of Tricks for Efficient Text Classification”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 427–431. URL: `https://www.aclweb.org/anthology/E17-2068`.

Kang, Dongyeop, Tushar Khot, Ashish Sabharwal, and Eduard Hovy (2018). “AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2418–2428. URL: `http://www.aclweb.org/anthology/P18-1225`.

Karimi Mahabadi, Rabeeh, Yonatan Belinkov, and James Henderson (July 2020). “End-to-End Bias Mitigation by Modelling Biases in Corpora”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8706–8716. URL: `https://www.aclweb.org/anthology/2020.acl-main.769`.

Khot, Tushar, Ashish Sabharwal, and Peter Clark (2018). “SciTail: A Textual Entailment Dataset from Science Question Answering”. In: *AAAI*.

Kiela, Douwe, Alexis Conneau, Allan Jabri, and Maximilian Nickel (2018). “Learning Visually Grounded Sentence Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 408–418.

Kifer, Daniel, Shai Ben-David, and Johannes Gehrke (2004). “Detecting change in data streams”. In: *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp. 180–191.

Kilgarriff, Adam (1998). “SENSEVAL: an exercise in evaluating world sense disambiguation programs”. In: *First International Conference on language resources & evaluation: Granada, Spain, 28-30 May 1998*. European Language Resources Association, pp. 581–588.

Kilgarriff, Adam and Martha Palmer (2000). “Introduction to the special issue on SENSEVAL”. In: *Computers and the Humanities* 34.1-2, pp. 1–13.

Kim, Najoung, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick (June 2019). “Probing What Different NLP Tasks Teach Machines about Function Word Comprehension”. In: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 235–249. URL: https://www.aclweb.org/anthology/S19-1026.

Kim, Najoung, Song Feng, Chulaka Gunasekara, and Luis Lastras (July 2020). "Implicit Discourse Relation Classification: We Need to Talk about Evaluation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5404–5414. URL: https://www.aclweb.org/anthology/2020.acl-main.480.

King, Maghi, Bente MAEGAARD, Jorg SCHÜTZ, Louis des TOMBE, Annelise BECH, Ann NEVILLE, Antti ARPPE, Lorna BALKAN, Colin BRACE, Harry BUNT, Lauri CARLSON, Shona DOUGLAS, Monika HÖGE, Steven KRAUWER, Sandra MANZI, Cristina MAZZI, Ann June SIELEMANN, and Ragna STEENBAKKERS (1995). *EAGLES: Evaluation of Natural Language Processing Systems. Final Report.* Tech. rep. URL: https://www.issco.unige.ch/en/research/projects/ewg95.

King, Margaret (1996). "Evaluating natural language processing systems". In: *Communications of the ACM* 39.1, pp. 73–79.

King, Margaret and Kirsten Falkedal (1990). "Using Test Suites in Evaluation of Machine Translation Systems". In: *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*. URL: https://www.aclweb.org/anthology/C90-2037.

Knowles, Rebecca (2019). “Interactive and Adaptive Neural Machine Translation”. PhD thesis. Johns Hopkins University.

Koehn, Philipp (2017). “Neural Machine Translation”. In: *arXiv preprint arXiv:1709.07809*.

Koh, Sungryong, Jinee Maeng, Ji-Young Lee, Young-Sook Chae, and Key-Sun Choi (2001). “A test suite for evaluation of English-to-Korean machine translation systems”. In: *MT Summit’conference, Santiago de Compostela*.

Kotzias, Dimitrios, Misha Denil, Nando De Freitas, and Padhraic Smyth (2015). “From group to individual labels using deep features”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 597–606.

Kováź, Vojtźch, Miloź Jakubíźek, and Aleź Horák (2016). “On Evaluation of Natural Language Processing Tasks”. In: *Proceedings of the 8th International Conference on Agents and Artificial Intelligence*. SCITEPRESS-Science and Technology Publications, Lda, pp. 540–545.

Lai, Alice, Yonatan Bisk, and Julia Hockenmaier (2017). “Natural Language Inference from Multiple Premises”. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 100–109. URL: http://www.aclweb.org/anthology/I17-1011.

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2018). "Word translation without parallel data". In: *International Conference on Learning Representations.*

Lee, Kenton, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer (2015). "Event detection and factuality assessment with non-expert supervision". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1643–1648.

Lehmann, Sabine, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Herve Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold (1996). "TSNLP - Test Suites for Natural Language Processing". In: *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics.* URL: https://www.aclweb.org/anthology/C96-2120.

Levy, Omer, Felix Hill, Anna Korhonen, Kyunghyun Cho, Roi Reichart, Yoav Goldberg, and Antione Bordes, eds. (Aug. 2016). *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W16-2500.

Li, Bowen, Lili Mou, and Frank Keller (July 2019). "An Imitation Learning Approach to Unsupervised Parsing". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Florence, Italy: Association for Computational Linguistics, pp. 3485–3492. URL: `https://www.aclweb.org/anthology/P19-1338`.

Li, Yitong, Timothy Baldwin, and Trevor Cohn (2018). “Towards Robust and Privacy-preserving Text Representations”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 25–30. URL: `http://aclweb.org/anthology/P18-2005`.

Liang, Paul Pu, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency (July 2020). “Towards Debiasing Sentence Representations”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5502–5515. URL: `https://www.aclweb.org/anthology/2020.acl-main.488`.

Lin, Chin-Yew (July 2004). “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: `https://www.aclweb.org/anthology/W04-1013`.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick (2014). “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer, pp. 740–755.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg (2016). “Assessing the ability of LSTMs to learn syntax-sensitive dependencies”. In: *Transactions of the Association for Computational Linguistics* 4, pp. 521–535.

Liu, Frederick, Han Lu, and Graham Neubig (2018). “Handling Homographs in Neural Machine Translation”. In: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana.

Liu, Yang, Chengjie Sun, Lei Lin, and Xiaolong Wang (2016). “Learning natural language inference using bidirectional LSTM model and inner-attention”. In: *arXiv preprint arXiv:1605.09090*.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2020). *Ro{BERT}a: A Robustly Optimized {BERT} Pretraining Approach*. URL: `https://openreview.net/forum?id=SyxS0T4tvS`.

Lloberes, Marina, Irene Castellón, and Lluís Padró (July 2015). “Suitability of ParTes Test Suite for Parsing Evaluation”. In: *Proceedings of the 14th International Conference on Parsing Technologies*. Bilbao, Spain: Association for Computational Linguistics, pp. 61–65. URL: `https://www.aclweb.org/anthology/W15-2207`.

Lo, Chi-kiu (Sept. 2017). “MEANT 2.0: Accurate semantic MT evaluation for any output language”. In: *Proceedings of the Second Conference on Machine Transla-*

tion. Copenhagen, Denmark: Association for Computational Linguistics, pp. 589–597. URL: https://www.aclweb.org/anthology/W17-4767.

Lo, Chi-kiu and Dekai Wu (2011a). "MEANT: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 220–229.

Lo, Chi-kiu and Dekai Wu (June 2011b). "Structured vs. Flat Semantic Role Representations for Machine Translation Evaluation". In: *Proceedings of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 10–20. URL: https://www.aclweb.org/anthology/W11-1002.

Lo, Chi-kiu, Meriem Beloucif, Markus Saers, and Dekai Wu (2014). "XMEANT: Better semantic MT evaluation without reference translations". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2, pp. 765–771.

Lopez, Adam, Matt Post, Chris Callison-Burch, Jonathan Weese, Juri Ganitkevitch, Narges Ahmidi, Olivia Buzek, Leah Hanson, Beenish Jamil, Matthias Lee, et al. (2013). "Learning to translate with products of novices: a suite of open-ended challenge problems for teaching MT". In: *Transactions of the Association for Computational Linguistics* 1, pp. 165–178.

Lowd, Daniel and Christopher Meek (2005). “Adversarial learning”. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 641–647.

Luong, Thang, Richard Socher, and Christopher Manning (Aug. 2013). “Better Word Representations with Recursive Neural Networks for Morphology”. In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 104–113. URL: `https://www.aclweb.org/anthology/W13-3512`.

Maas, Andrew L, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts (2011). “Learning word vectors for sentiment analysis”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics, pp. 142–150.

Mac Kim, Sunghwan and Steve Cassidy (2015). “Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers”. In: *Proceedings of the Australasian Language Technology Association Workshop 2015*, pp. 57–65.

MacCartney, Bill (2009). “Natural language inference”. PhD thesis. Stanford University.

MacCartney, Bill, Michel Galley, and Christopher D Manning (2008). “A phrase-based alignment model for natural language inference”. In: *Proceedings of the conference* *on empirical methods in natural language processing.* Association for Computational Linguistics, pp. 802–811.

Manning, Christopher D (2006). “Local textual inference: it’s hard to circumscribe, but you know it when you see it–and NLP needs it”. In:

Manning, Christopher D (2015). “Computational linguistics and deep learning”. In: *Computational Linguistics* 41.4, pp. 701–707.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky (2014). “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60. URL: `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella bernardi, and Roberto Zamparelli (2014). “A SICK cure for the evaluation of compositional distributional semantic models”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA). URL: `http://www.aclweb.org/anthology/L14-1314`.

Marneffe, Marie-Catherine de, Anna N. Rafferty, and Christopher D. Manning (June 2008). “Finding Contradictions in Text”. In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, pp. 1039–1047. URL: `https://www.aclweb.org/anthology/P08-1118`.

Marvin, Rebecca and Philipp Koehn (2018). “Exploring Word Sense Disambiguation Abilities of Neural Machine Translation Systems”. In: *Proceedings of the 13th Conference of The Association for Machine Translation in the Americas (Volume 1: Research Track*. Boston, pp. 125–131.

McAuley, Julian and Jure Leskovec (2013). “Hidden factors and hidden topics: understanding rating dimensions with review text”. In: *Proceedings of the 7th ACM conference on Recommender systems*. ACM, pp. 165–172.

McCann, Bryan, James Bradbury, Caiming Xiong, and Richard Socher (2017). “Learned in Translation: Contextualized Word Vectors”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 6297–6308.

McCann, Bryan, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher (2018). “The natural language decathlon: Multitask learning as question answering”. In: *arXiv preprint arXiv:1806.08730*.

McRae, Ken, George S Cree, Mark S Seidenberg, and Chris McNorgan (2005). “Semantic feature production norms for a large set of living and nonliving things”. In: *Behavior research methods* 37.4, pp. 547–559.

Miller, Tristan and Iryna Gurevych (2015). “Automatic disambiguation of English puns”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1, pp. 719–729.

Miller, Tristan, Christian Hempelmann, and Iryna Gurevych (2017). “SemEval-2017 Task 7: Detection and Interpretation of English Puns”. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).* Vancouver, Canada: Association for Computational Linguistics, pp. 58–68. URL: `http://www.aclweb.org/anthology/S17-2005`.

Miller, Tristan and Mladen Turković (2016). “Towards the automatic detection and identification of English puns”. In: *The European Journal of Humour Research* 4.1, pp. 59–75.

Minard, Anne-Lyse, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son (2016). “MEANTIME, the NewsReader Multilingual Event and Time Corpus”. In: *Language Resources and Evaluation Conference (LREC).*

Minervini, Pasquale and Sebastian Riedel (2018). “Adversarially Regularising Neural NLI Models to Integrate Logical Background Knowledge”. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018).* Association for Computational Linguistics.

Mirkin, Shachar (2011). “Context and Discource Textual Entailment Inference”. PhD thesis. Bar Ilan University.

Mirkin, Shachar, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor (Aug. 2009). “Source-Language Entailment Modeling for Translating Unknown Terms”. In: *Proceedings of the Joint Conference of the 47th An-*

nual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, pp. 791–799. URL: https://www.aclweb.org/anthology/P09-1089.

Mishev, Kostadin, Ana Gjorgjevikj, Riste Stojanov, Igor Mishkovski, Irena Vodenska, Ljubomir Chitkushev, and Dimitar Trajanov (2019). “Performance Evaluation of Word and Sentence Embeddings for Finance Headlines Sentiment Analysis”. In: *ICT Innovations 2019. Big Data Processing and Mining*. Ed. by Sonja Gievska and Gjorgji Madjarov. Cham: Springer International Publishing, pp. 161–172.

Miyao, Yusuke, Hideki Shima, Hiroshi Kanayama, and Teruko Mitamura (2012). “Evaluating textual entailment recognition for university entrance examinations”. In: *ACM Transactions on Asian Language Information Processing (TALIP)* 11.4, p. 13.

Mollá, Diego and Ben Hutchinson (Apr. 2003). “Intrinsic versus Extrinsic Evaluations of Parsing Systems”. In: *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?* Columbus, Ohio: Association for Computational Linguistics, pp. 43–50. URL: https://www.aclweb.org/anthology/W03-2806.

Mostafazadeh, Nasrin, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen (2016). “A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories”. In:

Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, pp. 839–849. URL: http://www.aclweb.org/anthology/N16-1098.

Mostafazadeh, Nasrin, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen (Apr. 2017). “LSDSem 2017 Shared Task: The Story Cloze Test”. In: *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. Valencia, Spain: Association for Computational Linguistics, pp. 46–51. URL: https://www.aclweb.org/anthology/W17-0906.

Mou, Lili, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin (2016). “Natural Language Inference by Tree-Based Convolution and Heuristic Matching”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 130–136. URL: http://anthology.aclweb.org/P16-2022.

Munkhdalai, Tsendsuren and Hong Yu (2017). “Neural Tree Indexers for Text Understanding”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 11–21.

Naik, Aakanksha, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig (2018a). “Stress Test Evaluation for Natural Language Inference”. In: *Proceedings of the 27th International Conference on Computational Linguistics*.

Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2340–2353. URL: `http://www.aclweb.org/anthology/C18-1198`.

Naik, Aakanksha, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig (2018b). “Stress Test Evaluation for Natural Language Inference”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2340–2353. URL: `http://www.aclweb.org/anthology/C18-1198`.

Nangia, Nikita, Adina Williams, Angeliki Lazaridou, and Samuel Bowman (2017). “The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations”. In: *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pp. 1–10.

Napoles, Courtney (2018). “Monolingual Sentence Rewriting as Machine Translation: Generation and Evaluation”. PhD thesis. Johns Hopkins University.

Napoles, Courtney, Matthew Gormley, and Benjamin Van Durme (2012). “Annotated Gigaword”. In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*. Montréal, Canada: Association for Computational Linguistics, pp. 95–100. URL: `http://www.aclweb.org/anthology/W12-3018`.

Nayak, Neha, Gabor Angeli, and Christopher D Manning (2016). “Evaluating word embeddings using a representative suite of practical tasks”. In: *Proceedings of the 1st workshop on evaluating vector-space representations for nlp*, pp. 19–23.

Neubig, Graham (2017). “Neural Machine Translation and Sequence-to-sequence Models: A Tutorial”. In: *arXiv preprint arXiv:1703.01619*.

Nevěřilová, Zuzana (2014). “Paraphrase and textual entailment generation”. In: *International Conference on Text, Speech, and Dialogue*. Springer, pp. 293–300.

Nie, Allen, Erin Bennett, and Noah Goodman (July 2019). “DisSent: Learning Sentence Representations from Explicit Discourse Relations”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4497–4510. URL: https://www.aclweb.org/anthology/P19-1442.

Nie, Yixin and Mohit Bansal (2017). “Shortcut-Stacked Sentence Encoders for Multi-Domain Inference”. In: pp. 41–45. URL: http://www.aclweb.org/anthology/W17-5308.

Oepen, Stephan and Klaus Netter (1995). “TSNLP - Test Suites for Natural Language Processing”. In: *In J. Nerbonne (Ed.), Linguistic Databases (pp. 13 – 36*. CSLI Publications, pp. 711–716.

Pado, Sebastian, Michel Galley, Dan Jurafsky, and Christopher D. Manning (Aug. 2009). “Robust Machine Translation Evaluation with Entailment Features”. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, pp. 297–305. URL: https://www.aclweb.org/anthology/P09-1034.

Pakray, Partha, Santanu Pal, Soujanya Poria, Sivaji Bandyopadhyay, and Alexander F Gelbukh (2010). “JU_CSE_TAC: Textual Entailment Recognition System at TAC RTE-6.” In: *TAC Workshop*.

Palmer, Martha and Tim Finin (1990). “Workshop on the Evaluation of Natural Language Processing Systems”. In: *Computational Linguistics* 16.3, pp. 175–181. URL: https://www.aclweb.org/anthology/J90-3005.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury (2005). “The Proposition Bank: An Annotated Corpus of Semantic Roles”. In: *Computational Linguistics* 31.1, pp. 71–106. URL: https://www.aclweb.org/anthology/J05-1004.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (July 2002). “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. URL: https://www.aclweb.org/anthology/P02-1040.

Parikh, Ankur, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit (2016). “A Decomposable Attention Model for Natural Language Inference”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2249–2255. URL: http://www.aclweb.org/anthology/D16-1244.

Paroubek, Patrick, Stéphane Chaudiron, and Lynette Hirschman (2007). “Principles of Evaluation in Natural Language Processing”. In: *Traitement Automatique des Langues* 48.1, pp. 7–31.

Pastra, Katerina, ed. (Apr. 2003). *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?* Columbus, Ohio: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W03-2800.

Pasunuru, Ramakanth and Mohit Bansal (2017). “Multi-Task Video Captioning with Video and Entailment Generation”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 1273–1283.

Pasunuru, Ramakanth and Mohit Bansal (2018). “Multi-Reward Reinforced Summarization with Saliency and Entailment”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 646–653. URL: http://www.aclweb.org/anthology/N18-2102.

Pavlick, Ellie (2017). “Compositional Lexical Entailment for Natural Language Inference”. PhD thesis. University of Pennsylvania.

Pavlick, Ellie and Chris Callison-Burch (2016). “Most “babies” are “little” and most “problems” are “huge”: Compositional Entailment in Adjective-Nouns”. In: *Proceed-*

ings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, pp. 2164–2173. URL: http://www.aclweb.org/anthology/P16-1204.

Pavlick, Ellie, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme (2015). "FrameNet+: Fast Paraphrastic Tripling of FrameNet". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, pp. 408–413.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

Pepicello, William J and Thomas A Green (1984). *Language of riddles: new perspectives.* The Ohio State University Press.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237.

Petrolito, Ruggero (2018). “Word Embeddings in Sentiment Analysis.” In: *Italian Conference on Computational Linguistics.*

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein (2006). “Learning accurate, compact, and interpretable tree annotation”. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, pp. 433–440.

Phang, Jason, Thibault Févry, and Samuel R Bowman (2018). “Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks”. In: *arXiv preprint arXiv:1811.01088.*

Pitler, Emily, Annie Louis, and Ani Nenkova (July 2010). “Automatic Evaluation of Linguistic Quality in Multi-Document Summarization”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* Uppsala, Sweden: Association for Computational Linguistics, pp. 544–554. URL: `https://www.aclweb.org/anthology/P10-1056`.

Plummer, Bryan A, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik (2015). “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models”. In: *Computer Vision (ICCV), 2015 IEEE International Conference on.* IEEE, pp. 2641–2649.

Poliak, Adam and Benjamin Van Durme (2019). “Adversarial Learning for Robust Emergency Need Discovery in Low Resource Settings”. In: *Second Annual West Coast NLP (WeCNLP) Summit*.

Poliak, Adam, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme (2018a). “Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 67–81. URL: http://aclweb.org/anthology/D18-1007.

Poliak, Adam, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme (2018b). “Hypothesis Only Baselines in Natural Language Inference”. In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 180–191. URL: http://aclweb.org/anthology/S18-2023.

Poliak, Adam, Yonatan Belinkov, James Glass, and Benjamin Van Durme (2018c). “On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 513–523. URL: http://www.aclweb.org/anthology/N18-2082.

Popović, Maja and Sheila Castilho (Aug. 2019). “Challenge Test Sets for MT Evaluation”. In: *Proceedings of Machine Translation Summit XVII Volume 3: Tutorial Abstracts*. Dublin, Ireland: European Association for Machine Translation.

Prince, Ellen F (1978). “On the function of existential presupposition in discourse”. In: *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill.* Vol. 14, pp. 362–376.

Punyakanok, Vasin, Dan Roth, and Wen-tau Yih (2004). *Natural language inference via dependency tree mapping: An application to question answering*. Tech. rep.

Punyakanok, Vasin, Dan Roth, and Wen-tau Yih (2008). “The importance of syntactic parsing and inference in semantic role labeling”. In: *Computational Linguistics* 34.2, pp. 257–287.

Qiu, Yuanyuan, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang (2018). “Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings”. In: *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, pp. 209–221.

Rahman, Altaf and Vincent Ng (2012). “Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge”. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 777–789. URL: http://www.aclweb.org/anthology/D12-1071.

Ramakrishnan, Sainandan, Aishwarya Agrawal, and Stefan Lee (2018). "Overcoming language priors in visual question answering with adversarial regularization". In: *Advances in Neural Information Processing Systems*, pp. 1548–1558.

Rastogi, Pushpendre and Benjamin Van Durme (2014). "Augmenting framenet via PPDB". In: *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pp. 1–5.

Ravfogel, Shauli, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg (July 2020). "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7237–7256. URL: `https://www.aclweb.org/anthology/2020.acl-main.647`.

Read, Walter, Alex Quilici, John Reeves, and Michael Dyer (1988). "Evaluating Natural Language Systems: A Sourcebook Approach". In: *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*. URL: `https://www.aclweb.org/anthology/C88-2112`.

Reisinger, Drew, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme (2015). "Semantic Proto-Roles". In: *Transactions of the Association for Computational Linguistics* 3, pp. 475–488.

Reiter, Ehud (2018). "A Structured Review of the Validity of BLEU". In: *Computational Linguistics* 44.3, pp. 393–401.

Resnik, Philip and Jimmy Lin (2010). “11 Evaluation of NLP Systems”. In: *The handbook of computational linguistics and natural language processing* 57.

Resnik, Philip, Michael Niv, Michael Nossal, and Gregory Schnitzer (2006). “Using intrinsic and extrinsic metrics to evaluate accuracy and facilitation in computer-assisted coding”. In: *Perspectives in Health Information Management Computer Assisted Coding Conference Proceedings*.

Reyes, Antonio, Paolo Rosso, and Davide Buscaldi (2012). “From humor recognition to irony detection: The figurative language of social media”. In: *Data & Knowledge Engineering* 74, pp. 1–12.

Richardson, Kyle, Hai Na Hu, Lawrence S. Moss, and Ashish Sabharwal (2020). “Probing Natural Language Inference Models through Semantic Fragments”. In: *AAAI*. Vol. abs/1909.07521.

Riloff, Ellen, Janyce Wiebe, and Theresa Wilson (2003). “Learning subjective nouns using extraction pattern bootstrapping”. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, pp. 25–32.

Rocktäschel, Tim, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom (2015). “Reasoning about entailment with neural attention”. In: *arXiv preprint arXiv:1509.06664*.

Rodrigo, Álvaro, Anselmo Peñas, and Felisa Verdejo (2009). “Overview of the Answer Validation Exercise 2008”. In: *Evaluating Systems for Multilingual and Multimodal*

Information Access. Ed. by Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 296–313.

Roemmele, Melissa, Cosmin Adrian Bejan, and Andrew S Gordon (2011). "Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning." In: *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Rogers, Anna, Aleksandr Drozd, Anna Rumshisky, and Yoav Goldberg, eds. (June 2019). *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Minneapolis, USA: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W19-2000.

Romano, Lorenza, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli (2006). "Investigating a Generic Paraphrase-Based Approach for Relation Extraction". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. URL: https://www.aclweb.org/anthology/E06-1052.

Romanov, Alexey and Chaitanya Shivade (2018). "Lessons from Natural Language Inference in the Clinical Domain". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1586–1596. URL: https://www.aclweb.org/anthology/D18-1187.

Ross, Alexis and Ellie Pavlick (Nov. 2019). "How well do NLI models capture verb veridicality?" In: *Proceedings of the 2019 Conference on Empirical Methods in*

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, pp. 2230–2240. URL: `https://www.aclweb.org/anthology/D19-1228`.

Roth, Dan, Mark Sammons, and V.G.Vinod Vydiswaran (Aug. 2009). "A Framework for Entailed Relation Recognition". In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore: Association for Computational Linguistics, pp. 57–60. URL: `https://www.aclweb.org/anthology/P09-2015`.

Rudinger, Rachel (2019). "Decompositional Semantics for Events, Participants, and Scripts in Text". PhD thesis. Johns Hopkins University.

Rudinger, Rachel, Chandler May, and Benjamin Van Durme (2017). "Social Bias in Elicited Natural Language Inferences". In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, pp. 74–79.

Rudinger, Rachel, Aaron Steven White, and Benjamin Van Durme (2018). "Neural Models of Factuality". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 731–744. URL: `http://www.aclweb.org/anthology/N18-1067`.

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme (2018). “Gender Bias in Coreference Resolution”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 8–14. URL: http://www.aclweb.org/anthology/N18-2002.

Sakaguchi, Keisuke and Benjamin Van Durme (July 2018). “Efficient Online Scalar Annotation with Bounded Support”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 208–218. URL: https://www.aclweb.org/anthology/P18-1020.

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi (2020). “WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale”. In: *AAAI*.

Sammons, Mark, VG Vinod Vydiswaran, Tim Vieira, Nikhil Johri, Ming-Wei Chang, Dan Goldwasser, Vivek Srikumar, Gourab Kundu, Yuancheng Tu, Kevin Small, et al. (2009). “Relation Alignment for Textual Entailment Recognition.” In: *TAC Workshop*.

Sauri, Roser and James Pustejovsky (2007). “Determining modality and factuality for text entailment”. In: *Semantic Computing, 2007. ICSC 2007. International Conference on*. IEEE, pp. 509–516.

Schluter, Natalie and Daniel Varab (2018). “When data permutations are pathological: the case of neural natural language inference”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4935–4939. URL: `https://www.aclweb.org/anthology/D18-1534`.

Schubert, Lenhart K. and Chung Hee Hwang (2000). “Episodic Logic Meets Little Red Riding Hood: A Comprehensive Natural Representation for Language Understanding”. In: *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. Cambridge, MA, USA: MIT Press, 111–174.

Schuler, Karin Kipper (2005). “VerbNet: A broad-coverage, comprehensive verb lexicon”. In:

Schuster, Tal, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay (2019). “Towards Debiasing Fact Verification Models”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3410–3416.

Schwarcz, Robert M, John F Burger, and Robert F Simmons (1970). “A deductive question-answerer for natural language inference”. In: *Communications of the ACM* 13.3, pp. 167–183.

Schwartz, Roy, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith (2017a). “Story Cloze Task: UW NLP System”. In: *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. Valencia, Spain: Association for Computational Linguistics, pp. 52–55. URL: `http://aclweb.org/anthology/W17-0907`.

Schwartz, Roy, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith (2017b). “The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task”. In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 15–25. URL: `http://aclweb.org/anthology/K17-1004`.

Seethamol, S. and K. Manju (2017). “Paraphrase identification using textual entailment recognition”. In: *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, pp. 1071–1074.

Sekizawa, Yuuki, Tomoyuki Kajiwara, and Mamoru Komachi (2017). “Improving Japanese-to-English Neural Machine Translation by Paraphrasing the Target Language”. In: *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 64–69.

Settles, Burr (1995). “Active Learning Literature Survey”. In: *Science* 10.3, pp. 237–304.

Sharma, Rishi, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh (2018). "Tackling the Story Ending Biases in The Story Cloze Test". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 752–757. URL: `http://www.aclweb.org/anthology/P18-2119`.

Shi, Xing, Inkit Padhi, and Kevin Knight (2016). "Does String-Based Neural MT Learn Source Syntax?" In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1526–1534. URL: `https://aclweb.org/anthology/D16-1159`.

Sileo, Damien, Tim Van De Cruys, Camille Pradel, and Philippe Muller (June 2019). "Mining Discourse Markers for Unsupervised Sentence Representation Learning". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3477–3486. URL: `https://www.aclweb.org/anthology/N19-1351`.

Simard, Patrice Y, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. (2017). "Machine teaching: A new paradigm for building machine learning systems". In: *arXiv preprint arXiv:1707.06742*.

Sparck Jones, Karen (1994). "Natural language processing: a historical review". In: *Current issues in computational linguistics: in honour of Don Walker*. Springer, pp. 3–16.

Sparck Jones, Karen (1994). "Towards Better NLP System Evaluation". In: *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. URL: `https://www.aclweb.org/anthology/H94-1018`.

Spärck Jones, Karen (2005). "ACL Lifetime Achievement Award: Some Points in a Time". In: *Computational Linguistics* 31.1, pp. 1–14. URL: `https://www.aclweb.org/anthology/J05-1001`.

Sparck Jones, Karen and Julia R. Galliers (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Berlin, Heidelberg: Springer-Verlag.

Stacey, Joe, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel (2020). *There is Strength in Numbers: Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training*. arXiv: `2004.07790 [cs.LG]`.

Staliūnaitė, Ieva (2018). "Learning about Non-Veridicality in Textual Entailment". MA thesis. Utrecht University.

Sun, Shuo, Francisco Guzmán, and Lucia Specia (July 2020). "Are we Estimating or Guesstimating Translation Quality?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6262–6267. URL: https://www.aclweb.org/anthology/2020.acl-main.558.

Sunkle, Sagar, Deepali Kholkar, and Vinay Kulkarni (2016). “Informed active learning to aid domain experts in modeling compliance”. In: *2016 IEEE 20th International Enterprise Distributed Object Computing Conference (EDOC)*. IEEE, pp. 1–10.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). “Sequence to sequence learning with neural networks”. In: *Advances in neural information processing systems*, pp. 3104–3112.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus (2014). “Intriguing properties of neural networks”. In: *International Conference on Learning Representations (ICLR)*.

Taboada, Maite (2016). “Sentiment analysis: an overview from linguistics”. In: *Annual Review of Linguistics* 2, pp. 325–347.

Tay, Yi, Anh Tuan Luu, and Siu Cheung Hui (2018). “Compare, Compress and Propagate: Enhancing Neural Architectures with Alignment Factorization for Natural Language Inference”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1565–1575. URL: https://www.aclweb.org/anthology/D18-1185.

Teichert, Adam, Adam Poliak, Benjamin Van Durme, and Matthew R Gormley (2017). “Semantic Proto-Role Labeling”. In: *Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*.

Thawani, Avijit, Biplav Srivastava, and Anil Singh (June 2019). “SWOW-8500: Word Association task for Intrinsic Evaluation of Word Embeddings”. In: *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Minneapolis, USA: Association for Computational Linguistics, pp. 43–51. URL: `https://www.aclweb.org/anthology/W19-2006`.

Thorne, James and Andreas Vlachos (2020). “Avoiding catastrophic forgetting in mitigating model biases in sentence-pair classification with elastic weight consolidation”. In: *arXiv preprint arXiv:2004.14366*.

Tjong Kim Sang, Erik F. and Fien De Meulder (2003). “Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition”. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*. CONLL ’03. Edmonton, Canada: Association for Computational Linguistics, pp. 142–147. URL: `https://doi.org/10.3115/1119176.1119195`.

Tsuchiya, Masatoshi (2018). “Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment”. In: *11th International Conference on Language Resources and Evaluation (LREC2018)*.

Tsvetkov, Yulia, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer (Sept. 2015). “Evaluation of Word Vector Representations by Subspace Alignment". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2049–2054. URL: https://www.aclweb.org/anthology/D15-1243.

Van Durme, Benjamin (2010). "Extracting Implicit Knowledge from Text". PhD thesis. Rochester, NY 14627: University of Rochester. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.3964&rep=rep1&type=pdf.

Van Durme, Benjamin, Tom Lippincott, Kevin Duh, Deana Burchfield, Adam Poliak, Cash Costello, Tim Finin, Scott Miller, James Mayfield, Philipp Koehn, Craig Harman, Dawn Lawrie, Chandler May, Max Thomas, Annabelle Carrell, Julianne Chaloux, Tongfei Chen, Alex Comerford, Mark Dredze, Benjamin Glass, Shudong Hao, Patrick Martin, Pushpendre Rastogi, Rashmi Sankepally, Travis Wolfe, Ying-Ying Tran, and Ted Zhang (Nov. 2017). "CADET: Computer Assisted Discovery Extraction and Translation". In: *Proceedings of the IJCNLP 2017, System Demonstrations*. Tapei, Taiwan: Association for Computational Linguistics, pp. 5–8. URL: https://www.aclweb.org/anthology/I17-3002.

Vanderwende, Lucy and William B Dolan (2006). "What syntax can contribute in the entailment task". In: *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Springer, pp. 205–216.

Vashishtha, Siddharth, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White (2020). *Temporal Reasoning in Natural Language Inference.*

Víta, Martin (2015). “Computing Semantic Textual Similarity based on Partial Textual Entailment”. In: *Doctoral Consortium on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Vol. 2. SCITEPRESS, pp. 3–12.

Víta, Martin and Jakub Klímek (Sept. 2019). “Exploiting Open IE for Deriving Multiple Premises Entailment Corpus”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd., pp. 1257–1264. URL: https://www.aclweb.org/anthology/R19-1144.

Vázquez, Raúl, Alessandro Raganato, Mathias Creutz, and Jörg Tiedemann (2020). “A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation”. In: *Computational Linguistics* 0.0, pp. 1–38. eprint: https://doi.org/10.1162/coli_a_00377. URL: https://doi.org/10.1162/coli_a_00377.

Walker, Marilyn A, Pranav Anand, Robert Abbott, and Ricky Grant (2012). “Stance classification using dialogic properties of persuasion”. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 592–596.

Walter, Sharon M. (1992). “Neal-Montgomery NLP System Evaluation Methodology”. In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman,* *New York, February 23-26, 1992.* URL: https://www.aclweb.org/anthology/H92-1064.

Walter, Sharon M. (1998). “Book Reviews: Evaluating Natural Language Processing Systems: An Analysis and Review”. In: *Computational Linguistics* 24.2. URL: https://www.aclweb.org/anthology/J98-2013.

Wang, Alex, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman (2018). “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *arXiv preprint arXiv:1804.07461.*

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2019a). “Superglue: A stickier benchmark for general-purpose language understanding systems”. In: *Advances in Neural Information Processing Systems*, pp. 3261–3275.

Wang, Alex, Ian F. Tenney, Yada Pruksachatkun, Phil Yeres, Jason Phang, Haokun Liu, Phu Mon Htut, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Edouard Grave, Najoung Kim, Thibault Févry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman (2019b). *jiant 1.3: A software toolkit for research on general-purpose text understanding models.* http://jiant.info/.

Wang, Bin, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo (2019c). “Evaluating word embedding models: Methods and experimental results”. In: *APSIPA transactions on signal and information processing* 8.

Welbl, Johannes, Nelson F Liu, and Matt Gardner (2017). “Crowdsourcing Multiple Choice Science Questions”. In: *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106.

Welleck, Sean, Jason Weston, Arthur Szlam, and Kyunghyun Cho (2019). “Dialogue Natural Language Inference”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3731–3741.

White, Aaron Steven and Kyle Rawlins (2016). “A computational model of S-selection”. In: *Semantics and linguistic theory*. Vol. 26, pp. 641–663.

White, Aaron Steven and Kyle Rawlins (2018). “The role of veridicality and factivity in clause selection”. In: *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*. Amherst, MA: GLSA Publications, to appear.

White, Aaron Steven, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme (2016). “Universal Decompositional Semantics on Universal Dependencies”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1713–1723. URL: `https://aclweb.org/anthology/D16-1177`.

White, Aaron Steven, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme (2017). "Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 996–1005.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie (2005). "Annotating expressions of opinions and emotions in language". In: *Language resources and evaluation* 39.2-3, pp. 165–210.

Wieting, John and Kevin Gimpel (2017). "Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations". In: *arXiv preprint arXiv:1711.05732*.

Wieting, John and Kevin Gimpel (July 2018). "ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 451–462. URL: `https://www.aclweb.org/anthology/P18-1042`.

Wieting, John, Jonathan Mallinson, and Kevin Gimpel (2017). "Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 274–285. URL: `https://www.aclweb.org/anthology/D17-1026`.

Wilks, Yorick (1975). “A preferential, pattern-seeking, semantics for natural language inference”. In: *Artificial intelligence* 6.1, pp. 53–74.

Williams, Adina, Nikita Nangia, and Samuel R Bowman (2017). “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”. In: *arXiv preprint arXiv:1704.05426*.

Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa (2006). “Recognizing strong and weak opinion clauses”. In: *Computational intelligence* 22.2, pp. 73–99.

Wooden, John and Steve Jamison (1997). *Wooden: a lifetime of observations and reflections on and off the court.*

Workshop on MT Evaluation: Hands-On Evaluation (2001).

Wu, Yonghui, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu (2015). “A study of neural word embeddings for named entity recognition in clinical text”. In: *AMIA Annual Symposium Proceedings.* Vol. 2015. American Medical Informatics Association, p. 1326.

Xie, Qizhe, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig (2017). “Controllable invariance through adversarial feature learning”. In: *Advances in Neural Information Processing Systems*, pp. 585–596.

Yanaka, Hitomi, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos (Aug. 2019). “Can Neural Networks Understand Monotonicity Reasoning?" In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 31–40. URL: `https://www.aclweb.org/anthology/W19-4804`.

Yanaka, Hitomi, Koji Mineshima, Daisuke Bekki, and Kentaro Inui (2020). "Do Neural Models Learn Systematicity of Monotonicity Inference in Natural Language?" In: *ACL*.

Yang, Diyi, Alon Lavie, Chris Dyer, and Eduard Hovy (2015). "Humor Recognition and Humor Anchor Extraction". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2367–2376. URL: `http://www.aclweb.org/anthology/D15-1284`.

Yao, Xuchen (2014). "Feature-driven question answering with natural language alignment". PhD thesis. Johns Hopkins University.

Yin, Wenpeng and Hinrich Schütze (2018). "Attentive Convolution: Equipping CNNs with RNN-style Attention Mechanisms". In: *Transactions of the Association for Computational Linguistics* 6, pp. 687–702.

Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier (2014a). "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions". In: *Transactions of the Association for Computational Linguistics* 2, pp. 67–78.

Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier (2014b). "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions". In: *Transactions of the Association for Computational Linguistics* 2, pp. 67–78.

Yu, Hong and Vasileios Hatzivassiloglou (2003). "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences". In: *Proceedings of the 2003 conference on Empirical methods in natural language processing.* Association for Computational Linguistics, pp. 129–136.

Yu, Weihao, Zihang Jiang, Yanfei Dong, and Jiashi Feng (2020). "ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning". In: *International Conference on Learning Representations (ICLR).*

Yuan, Michelle, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber (2019). "Interactive Refinement of Cross-Lingual Word Embeddings". In: *arXiv*, arXiv–1911.

Zaenen, Annie, Lauri Karttunen, and Richard Crouch (June 2005). "Local Textual Inference: Can it be Defined or Circumscribed?" In: *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment.* Ann Arbor, Michigan: Association for Computational Linguistics, pp. 31–36. URL: `https://www.aclweb.org/anthology/W05-1206`.

Zanzotto, Fabio Massimo, Marco Pennaccchiotti, and Kostas Tsioutsiouliklis (July 2011). "Linguistic Redundancy in Twitter". In: *Proceedings of the 2011 Conference* *on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, pp. 659–669. URL: `https://www.aclweb.org/anthology/D11-1061`.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell (2018). "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, pp. 335–340.

Zhang, Chen (2010). "Natural Language Interference from Textual Entailment to Conversation Entailment". AAI3435149. PhD thesis. East Lansing, MI, USA: Michigan State University.

Zhang, Chen and Joyce Chai (Oct. 2010). "Towards Conversation Entailment: An Empirical Investigation". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, pp. 756–766. URL: `https://www.aclweb.org/anthology/D10-1074`.

Zhang, Chen and Joyce Y. Chai (2009). "What Do We Know About Conversation Participants: Experiments on Conversation Entailment". In: *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. SIGDIAL '09. London, United Kingdom: Association for Computational Linguistics, pp. 206–215. URL: `http://dl.acm.org/citation.cfm?id=1708376.1708406`.

Zhang, Mozhi, Yoshinari Fujinuma, and Jordan Boyd-Graber (2020). "Exploiting Cross-Lingual Subword Similarities in Low-Resource Document Classification". In: *Proceedings of the AAAI Conference on Artificial Intelligence.*

Zhang, Sheng, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme (2017). "Ordinal Common-sense Inference". In: *Transactions of the Association of Computational Linguistics* 5.1, pp. 379–395.

Zhang, Yuan, Regina Barzilay, and Tommi Jaakkola (2017). "Aspect-augmented Adversarial Networks for Domain Adaptation". In: *Transactions of the Association for Computational Linguistics* 5, pp. 515–528. URL: https://www.aclweb.org/anthology/Q17-1036.

Zhou, Hao, Zhaopeng Tu, Shujian Huang, Xiaohua Liu, Hang Li, and Jiajun Chen (2017). "Chunk-Based Bi-Scale Decoder for Neural Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Vancouver, Canada: Association for Computational Linguistics, pp. 580–586.

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen (2016). "The United Nations Parallel Corpus v1.0". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan

Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA).

Vita

Adam Poliak was born and grew up in South Florida. He graduated with a B.A. and M.S.E in Computer Science from Johns Hopkins University in 2016 and 2019 respectively. Adam is the recipeint of a 2017 National GEM Consortium Fellowship. During his Ph.D., Adam was a teaching assistant in Fall 2019 and sole course instructor in Spring 2020 for Artificial Intelligence (EN.601.464). Adam received Best Paper Awards in 2018 and 2019 at The Joint Conference on Lexical and Computational Semantics. Adam's research focuses on natural language understanding, exploring the limits and reasoning capabilities of NLP systems, and applying NLP to domains like public health and social sciences.