A Survey on Recognizing Textual Entailment as an NLP Evaluation

Adam Poliak

Barnard College, Data Science Institute, Columbia University
3009 Broadway, New York, NY 10027
`apoliak@barnard.ed`

Abstract

Recognizing Textual Entailment (RTE) was proposed as a unified evaluation framework to compare semantic understanding of different NLP systems. In this survey paper, we provide an overview of different approaches for evaluating and understanding the reasoning capabilities of NLP systems. We then focus our discussion on RTE by highlighting prominent RTE datasets as well as advances in RTE dataset that focus on specific linguistic phenomena that can be used to evaluate NLP systems on a fine-grained level. We conclude by arguing that when evaluating NLP systems, the community should utilize newly introduced RTE datasets that focus on specific linguistic phenomena.

1 Introduction

As NLP technologies are more widely adopted, how to evaluate NLP systems and how to determine whether one model understands language or generates text better than another is an increasingly important question. Recognizing Textual Entailment (RTE Cooper et al., 1996; Dagan et al., 2006), the task of determining whether the meaning of one sentence can likely be inferred from another was introduced to answer this question.

We begin this survey by discussing different approaches over the past thirty years for evaluating and comparing NLP systems. Next, we will discuss how 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 then highlight recent RTE datasets that focus on specific semantic phenomena and conclude by arguing that they should be utilized for evaluating the reasoning capabilities of downstream NLP systems.

Natural Language Inference or Recognizing Textual Entailment?

The terms Natural Language Inference (NLI) and RTE are often used interchangeably. Many papers begin by explicitly mentioning that these terms are synonymous (Liu et al., 2016; Gong et al., 2018; Camburu et al., 2018).[1] 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 used in RTE. Earlier works used the term *natural language inference* in this way (Schwarcz et al., 1970; Wilks, 1975; Punyakanok et al., 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 (Manning, 2006). Zaenen et al. (2005) prefer the term *textual inference* because examples in RTE datasets often require a system to not only identify entailments but also conventional implicatures, conversational implicatures, and world knowledge.

If starting over, we would advocate for the 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.

---

[1] In fact, variants of the phrase "natural language inference, also known as recognizing textual entailment" appear in many papers (Chen et al., 2017; Williams et al., 2017; Naik et al., 2018; Chen et al., 2018; Tay et al., 2018, *i.a.*).

2 Evaluating NLP Systems

The question of how best to evaluate NLP systems is an open problem intriguing the community for decades. A 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. Some workshops focused on: 1) evaluations in general (Pastra, 2003); 2) different NLP tasks, e.g. machine translation (ws-, 2001; Goldstein et al., 2005) and summarization (Conroy et al., 2012; Giannakopoulos et al., 2017); or 3) contemporary NLP approaches that rely on vector space representations (Levy et al., 2016; Bowman et al., 2017; Rogers et al., 2019).

In the quest to develop an ideal evaluation framework for NLP systems, researchers proposed multiple evaluation methods, e.g. EAGLES (King et al., 1995), TSNLP (Oepen and Netter, 1995; Lehmann et al., 1996), *FraCas* (Cooper et al., 1996), SENSEVAL (Kilgarriff, 1998), CLEF (Agosti et al., 2007), and others. These approaches are often divided along multiple dimensions. Here, we will survey approaches along two dimensions: 1) intrinsic vs. extrinsic evaluations; 2) general purpose vs task specific evaluations.[2]

2.1 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)

When reviewing Sparck Jones and Galliers (1996)'s textbook on NLP evaluations, Estival (1997) comments 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, for example, "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 versus application-driven evaluations (Kováź et al., 2016).[3]

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 et al., 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.[4] Developing intrinsic evaluations that correlate with extrinsic evaluations remains an open problem in NLP.

---

[2]Resnik and Lin (2010) summarize other evaluation approaches and Paroubek et al. (2007) present a history and evolution of NLP evaluation methods.

[3]As another example, 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. This is the basis of evaluation benchmarks like SimLex (Hill et al., 2015), Verb (Baker et al., 2014), RW (Luong et al., 2013), MEN (Bruni et al., 2012), WordSim-353 (Finkelstein et al., 2001), and others. Extrinsic evaluations for word embeddings 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., 2019b), or named entity recognition (Wu et al., 2015; Nayak et al., 2016).

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

2.2 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 et al., 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, or metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) in generation tasks. 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 distinction 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 et al., 2015), machine translation (King and Falkedal, 1990; Koh et al., 2001; Isabelle et al., 2017; Choshen and Abend, 2019; Popović and Castilho, 2019; Avramidis et al., 2019), summarization (Pitler et al., 2010), and others (Chinchor, 1991; Chinchor et al., 1993).

Test Suites vs. Test Corpora This distinction can also be described 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 [but] 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 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.

2.3 Probing Deep Learning NLP Models

In recent years, interpreting and analysing NLP models has become prominent in many research agendas. Contemporary and successful deep learning NLP methods are not as interpretable as previously popular NLP approaches relying on feature engineering. Approaches for interpreting and analysing how well NLP models capture linguistic phenomena often leverage auxiliary or diagnostic classifiers. Contemporary deep learning 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 sen-

tence length are captured in pre-trained sentence representations.

As Adi et al.'s general methodology "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 et al., 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., 2017a), and syntax (Belinkov et al., 2017b).

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) are used to evaluate how different sentence representations perform on paraphrase detection, semantic textual similarity, and a wide range of other binary and multiclass classification problems. We categorize these datasets as extrinsic evaluations since they often treat learned sentence-representations as features to train a classifier for an external task. However, most of these do not count as test suites, since the data is not tightly controlled to evaluate specific linguistic phenomena. Rather, resources like GLUE and SuperGLUE (Wang et al., 2019a) 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.

3   Recognizing Textual Entailment

> 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 et al., 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 a framework to evaluate NLP systems. 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. For example, annotation guidelines for an RTE dataset[5] 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 et al., 2006)

Starting with *FraCas*, we will discuss influential 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,[6] *FraCas* is a "semantic test suite" that covers a range of semantic phenomena categorized into 9 classes. These are generalized quantifiers, plurals, anaphora, ellipsis, adjectives, comparatives, temporal reference, verbs, and attitudes. Based on the descriptions in §2, we would classify *FraCas* as an intrinsic evaluation and a general purpose test suite.

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.[7] Table 4 (in the appendix) 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, the limited number of examples in *FraCas*

---

[5]These were the guidelines in RTE-1.

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

[7]https://nlp.stanford.edu/~wcmac/downloads/fracas.xml

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 1: 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 example's label.

prevents its use as a dataset to train data hungry deep learning models.

Pascal RTE Challenges With a similar broad goal as *FraCas*, the Pascal Recognizing Textual Entailment challenges began as a "generic evaluation framework" to compare the inference capabilities of models designed to perform different tasks, based on the intuition "that major inferences, as needed by multiple applications, can indeed be cast in terms of textual entailment" (Dagan et al., 2006). Unlike *FraCas*'s goal of determining whether a model performs distinct types of reasoning, the Pascal RTE Challenges 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 (including the quote above (Dagan et al., 2006)).[8] For example, the RTE1 data came from 7 tasks: comparable documents, reading comprehension, question answering, information extraction, machine translation, information retrieval, and paraphrase acquisition.[9] Starting with Dagan et al. (2006), there have been eight iterations of the RTE challenge, with the most recent being Dzikovska et al. (2013).

SNLI and MNLI The most popular recent RTE datasets, Stanford Natural Language Inference (SNLI; Bowman et al., 2015) and its successor Multi-NLI (Williams et al., 2017), each contain over half a million examples and enabled re-

searchers 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., 2014). 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 provides such examples for both datasets. Rudinger et al. (2017) illustrated how eliciting textual data in this fashion creates stereotypical biases in SNLI. Some of the biases are gender-, age-, and race-based. Poliak et al. (2018c) argue that this may cause additional biases enabling a hypothesis-only model to outperform the majority baseline on SNLI by 100 percent (Gururangan et al., 2018; Tsuchiya, 2018).

3.1 Entailment as a Downstream NLP Task

The datasets in the PASCAL RTE Challenges were primarily treated as test corpora. Teams participated in those challenges by developing models to achieve increasingly high scores on each challenges' datasets. Since RTE was motivated as a diagnostic, researchers analyzed the RTE challenge datasets. de Marneffe et al. (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[10] 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." Addi-

---

[8]See Appendix A for the annotation guidelines for RTE1, RTE2, and RTE3.

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

[10]In Chapter 2.2 of his thesis

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: 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.

tionally, Vanderwende and Dolan (2006) and Blake (2007) demonstrate how sentence structure alone can provide a high signal for some RTE datasets.[11] Despite these analyses, researchers primarily built models to perform the task on the PASCAL RTE datasets rather than leveraging these datasets to evaluate models built for other tasks.

Coinciding with the recent "deep learning wave" that has taken over NLP and Machine Learning (Manning, 2015), the introduction of large scale RTE datasets, specifically SNLI and MNLI, 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 et al., 2018) or medical information (Romanov and Shivade, 2018), emerged as well. However, 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 for new RTE datasets.

4 Revisiting RTE as an NLP Evaluation

> There has been little evidence to suggest [that RTE models] capture the type of compositional or world knowledge tested by datasets like the FraCas test suite.
>
> (Pavlick, 2017)

As large scale RTE datasets, like SNLI and MNLI, 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. As noted by White et al. (2017), "researchers compete on which system achieves the highest score on

a test set, but this itself does not lead to an understanding of which linguistic properties are better captured by a quantitatively superior system." In other words, 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 regarding the earlier PASCAL RTE datasets. In her thesis that presented "a test suite for adjectival inference developed as a resource for the evaluation of computational systems handling natural language inference." Amoia (2008) blamed "the difficulty of defining the linguistic phenomena which are responsible for inference" as the reason why previous RTE resources "concentrated on the creation of applications coping with textual entailment" rather than "resources for the evaluation of such applications."

As current studies began exploring what linguistic phenomena are captured by neural NLP models and auxiliary diagnostic classifiers became a common tool to evaluate sentence representations in NLP systems, (§2.3), the community saw an interest in developing RTE datasets that can provide insight into what type of linguistic phenomena are captured by neural, deep learning models. In turn, the community is answering Chatzikyriakidis et al. (2017) plea to the community to test "more kinds of inference" than in previous RTE challenge sets. Here, we will highlight recent efforts in creating datasets that demonstrate how the community has started answering Chatzikyriakidis et al.'s call. We group these different datasets based on how they were created. Table 3 includes additional RTE datasets focused on specific linguistic phenomena.

---

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

4.1 Automatically Created

White et al. (2017) advocate for using RTE as a single framework to evaluate different linguistic phenomena. They argue for creating RTE datasets focused on specific phenomena by *recasting* existing annotations for different semantic phenomena into RTE. Poliak et al. (2018b) introduce the Diverse Natural Language Inference Collection (DNC) of over half a million RTE examples. They create the DNC by converting 7 semantic phenomena from 13 existing datasets into RTE. These phenomena include event factuality, named entity recognition, gendered anaphora resolution, sentiment analysis, relationship extraction, pun detection, and lexicosyntactic inference. Staliūnaitė (2018)'s master's thesis improved Poliak et al. (2018b)'s method used to recast annotations for factuality into RTE. Other efforts have created recast datasets in Hindi that focus on sentiment and emotion detection.[12]

Concurrent to the DNC, Naik et al. (2018) released the "NLI Stress Tests" that included RTE datasets focused on negation, word overlap between premises and hypotheses, numerical reasoning, amongst other phenomena. Naik et al. (2018) similarly create their stress tests automatically using different methods for each phenomena. They then used these datasets to evaluate how well a wide class of RTE models capture these phenomena. Other RTE datasets that target more specific phenomena were created using automatic methods, including Jeretic et al. (2020)'s "IMPRES" diagnostic RTE dataset that tests for IMPlicatures and PRESuppositions.

If not done with thorough testing and care, recasting or other automatic methods for creating these RTE datasets can lead to annotation artifacts unrelated to RTE that limit how well a dataset tests for a specific semantic phenomena. For example, 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.[13] This limits the utility of the recast version of FN+ to be used when evaluating how well models capture paraphrastic inference.

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; Gardner et al., 2019). Recasting problems into RTE, as opposed to question-answering, has deeper roots in linguistic theory (Seuren, 1998; Chierchia and McConnell-Ginet, 2000; Brinton, 2000), and continues a rich history within the NLP community.

4.2 Semi-Automatically Created

Other RTE datasets focused on specific phenomena rely on semi-automatic methods. RTE pairs are often generated automatically using well developed heuristics. Instead of automatically labeling the RTE example pairs (like in the approaches previously discussed), the automatically created examples are often labeled by crowdsource workers. For example, Kim et al. (2019) use hueristics to create RTE pairs that test for prepositions, comparatives, quantification, spacial reasoning, and negation and then present these examples to crowdsource workers on Amazon Mechanical Turk. Similarly, Ross and Pavlick (2019) generate two premise-hypothesis pairs for each RTE example in MNLI that satisfy their set of constraints. Next, they rely on crowdsource workers to annotated whether the premise likely entails the hypothesis on a 5-point Likert scale.

Some methods instead first manually annotate their data and then rely on automatic methods to construct hypotheses and label RTE pairs. When generating RTE examples testing for monotonicity, Richardson et al. (2020) first manually encode the "monotonicity information of each token in the lexicon and built sentences via a controlled set of grammar rules." They then "substitute upward entailing tokens or constituents with something 'greater than or equal to' them, or downward entailing ones with something 'less than or equal to' them."

4.3 Manually Created

While most of these datasets rely on varying degrees of automation, some RTE datasets focused on evaluating how well models capture specific phenomena rely on manual annotations. The GLUE and SuperGlue datasets include diagnostic sets where annotators manually labeled samples of examples as requiring a broad range of linguistic phenomena. The types of phenomena manu-

---

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

[13] Table 5 (in the appendix) 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.

Proto-Roles (White et al., 2017), Paraphrastic Inference (White et al., 2017), Event Factuality (Poliak et al., 2018b; Staliūnaitė, 2018), Anaphora Resolution (White et al., 2017; Poliak et al., 2018b), Lexicosyntactic Inference (Pavlick and Callison-Burch, 2016; Poliak et al., 2018b; Glockner et al., 2018), Compositionality (Dasgupta et al., 2018), Prepositions (Kim et al., 2019), Comparatives (Kim et al., 2019; Richardson et al., 2020), Quantification/Numerical Reasoning (Naik et al., 2018; Kim et al., 2019; Richardson et al., 2020), Spatial Expressions (Kim et al., 2019), Negation (Naik et al., 2018; Kim et al., 2019; Richardson et al., 2020), Tense & Aspect (Kober et al., 2019), Veridicality (Poliak et al., 2018b; Ross and Pavlick, 2019), Monotonicity (Yanaka et al., 2019, 2020; Richardson et al., 2020), Presupposition (Jeretic et al., 2020), Implicatures (Jeretic et al., 2020), Temporal Reasoning (Vashishtha et al., 2020)

Table 3: List of different semantic phenomena tested for in recent RTE datasets.

ally labeled include lexical semantics, predicate-argument structure, logic, and common sense or world knowledge.[14]

5 Recommendations

These efforts resulted in a consistent format and framework for testing how well contemporary, deep learning NLP systems capture a wide-range of linguistic phenomena. However, so far, most of these datasets that target specific linguistic phenomena have been used to solely evaluate how well RTE models capture a wide range of phenomena, as opposed to evaluating how well systems trained for more applied NLP tasks capture these phenomena. Since RTE was introduced as a framework to evaluate how well NLP models cope with inferences, these newly created datasets have not been used to their full potential.

A limited number of studies used some of these datasets to evaluate how well models trained for other tasks capture these phenomena. Poliak et al. (2018a) evaluated how well a BiLSTM encoder trained as part of a neural machine translation system capture phenomena like semantic proto-roles, paraphrastic inference, and anaphora resolution. Kim et al. (2019) used their RTE datasets focused on function words to evaluate different encoders trained for tasks like CCG parsing, image-caption matching, predicting discourse markers, and others. Those studies relied on the use of auxiliary classifiers as a common probing technique to evaluate sentence representations. As the community's interest in analyzing deep learning systems increases, demonstrated by the recent work relying on (Linzen et al., 2018, 2019) and improving upon (Hewitt and Liang, 2019; Voita and Titov, 2020; Pimentel et al., 2020; Mu and Andreas, 2020) the popular auxiliary

classifier-based diagnostic technique, we call on the community to leverage the increasing number of RTE datasets focused on different semantic phenomena (Table 3) to thoroughly study the representations learned by downstream, applied NLP systems. The increasing number of RTE datasets focused on different phenomena can help researchers use one standard format to analyze how well models capture different phenomena, and in turn answer Sammons et al. (2010)'s challenge to make RTE "a central component of evaluation for relevant NLP tasks."

Another recent line of work uses RTE to evaluate the output of text generation systems. For example, Falke et al. (2019) explore "whether textual entailment predictions can be used to detect errors" in abstractive summarization systems and if errors "can be reduced by reranking alternative predicted summaries" with a textual entailment system trained on SNLI. While Falke et al. (2019) results demonstrated that current models might not be accurate enough to rank generated summaries, Barrantes et al. (2020) demonstrate that contemporary transformer models trained on the Adversarial NLI dataset (Nie et al., 2020) "achieve significantly higher accuracy and have the potential of selecting a coherent summary." Therefore, we are encouraged that researchers might be able to use many of these new RTE datasets focused on specific phenomena to evaluate the coherency of machine generated text based on multiple linguistic phenomena that are integral to entailment and NLU. This approach can help researchers use the RTE datasets to evaluate a wider class of models, specifically non-neural models, unlike the auxiliary classifier or probing methods previously discussed.

The overwhelming majority, if not all, of these RTE datasets targeting specific phenomena rely on categorical RTE labels, following the common for-

---

[14] https://gluebenchmark.com/diagnostics

mat of the task. However, as Chen et al. (2020) recently illustrated, categorical RTE labels do not capture the subjective nature of the task. Instead, they argue for scalar RTE labels that indicate how likely a hypothesis could be inferred by a premise. Pavlick and Kwiatkowski (2019) similarly lament how labels are currently used in RTE datasets. Pavlick and Kwiatkowski demonstrate that a single label aggregated from multiple annotations for one RTE example minimizes the "type of uncertainty present in [valid] human disagreements." Instead, they argue that a "representation should be evaluated in terms of its ability to predict the full distribution of human inferences (e.g., by reporting crossentropy against a distribution of human ratings), rather than to predict a single aggregate score (e.g., by reporting accuracy against a discrete majority label or correlation with a mean score)." Future RTE datasets targeting specific phenomena that contain scalar RTE labels from multiple annotators (following Chen et al.'s and Pavlick and Kwiatkowski's recommendations) can provide more insight into contemporary NLP models.

6 Conclusion

With the current zeitgeist of NLP research where researchers are interested in analyzing state-of-the-art deep learning models, now is a prime time to revisit RTE as a method to evaluate the inference capabilities of NLP models. In this survey, we discussed recent advances in RTE datasets that focus on specific linguistic phenomena that are integral for determining whether one sentence is likely inferred by another. Since RTE was primarily motivated as an evaluation framework, we began this survey with a broad overview of prior approaches for evaluating NLP systems. This included the distinctions between instrinsic vs extrinsic evaluations and general purpose vs task specific evaluations.

We discussed foundational RTE datasets that greatly impacted the NLP community and included critiques of why they do not fulfill the promise of RTE as an evaluation framework. We highlighted recent efforts to create RTE datasets that focus on specific linguistic phenomena. By using these datasets to evaluate sentence representations from neural models or rank generated text from NLP systems, researchers can help fulfil the promise of RTE as unified evaluation framework. Ultimately, this will help us determine how well models understand language on a fine-grained level.

References

2001. *Workshop on MT Evaluation: Hands-On Evaluation*.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *ICLR*.

Maristella Agosti, 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*, pages 509–512. Springer.

Marilisa Amoia. 2008. *Linguistic-Based Computational Treatment of Textual Entailment Recognition*. Theses, Université Henri Poincaré - Nancy 1.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 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)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 278–289, Doha, Qatar. Association for Computational Linguistics.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, and Bernardo Magnini. 2006. The second pascal recognising textual entailment challenge.

Mario Barrantes, Benedikt Herudek, and Richard Wang. 2020. Adversarial nli for factual correctness in text summarisation models.

Yonatan Belinkov. 2018. *On internal language representations in deep learning: An analysis of machine translation and speech recognition*. Ph.D. thesis, Massachusetts Institute of Technology.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 861–872. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 2441–2451. Curran Associates, Inc.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7:49–72.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. 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), pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In Proceedings of ACL-08: HLT, Short Papers, pages 197–200, Columbus, Ohio. Association for Computational Linguistics.

Catherine Blake. 2007. The role of sentence structure in recognizing textual entailment. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, RTE '07, pages 101–106, Stroudsburg, PA, USA. Association for Computational Linguistics.

Samuel Bowman, Yoav Goldberg, Felix Hill, Angeliki Lazaridou, Omer Levy, Roi Reichart, and Anders Søgaard, editors. 2017. Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP. Association for Computational Linguistics, Copenhagen, Denmark.

Samuel R. Bowman, 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.

L. Brinton. 2000. The Structure of Modern English: A linguistic introduction.

Elia Bruni, 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), pages 136–145.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle,

K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 9539–9549. Curran Associates, Inc.

J Gregory Caporaso, 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, pages 640–651. World Scientific.

Stergios Chatzikyriakidis, 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.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. 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), pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. 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, pages 36–40, Copenhagen, Denmark. Association for Computational Linguistics.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In ACL.

Gennaro Chierchia and Sally McConnell-Ginet. 2000. Meaning and grammar: An introduction to semantics.

Nancy Chinchor. 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.

Nancy Chinchor, Lynette Hirschman, and David D. Lewis. 1993. Evaluating message understanding systems: An analysis of the third message understanding conference (MUC-3). Computational Linguistics, 19(3):409–450.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, pages 1–6, Berlin, Germany. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 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), pages 291–303,

Hong Kong, China. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc 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)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

John M. Conroy, Hoa Trang Dang, Ani Nenkova, and Karolina Owczarzak, editors. 2012. *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*. Association for Computational Linguistics, Montréal, Canada.

Robin Cooper, 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.

Ido Dagan, 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*, pages 177–190. Springer.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Myroslava Dzikovska, 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)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Dominique Estival. 1997. Karen sparck jones & julia r. galliers, evaluating natural language processing systems: An analysis and review. lecture notes in artificial intelligence 1083. *Machine Translation*, 12(4):375–379.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

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

Lev Finkelstein, 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*, pages 406–414.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question answering is a format; when is it useful? *arXiv preprint arXiv:1909.11291*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

George Giannakopoulos, Elena Lloret, John M. Conroy, Josef Steinberger, Marina Litvak, Peter Rankel, and Benoit Favre, editors. 2017. *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*. Association for Computational Linguistics, Valencia, Spain.

Oren Glickman. 2006. *Applied textual entailment*. Ph.D. thesis, Bar Ilan University.

Max Glockner, 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)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors. 2005. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *International Conference on Learning Representations*.

Suchin Gururangan, 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)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

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

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Theo M. V. Janssen. 2011. Montague semantics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, winter 2011 edition. Metaphysics Research Lab, Stanford University.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Adam Kilgarriff. 1998. Senseval: an exercise in evaluating world sense disambiguation programs. In *First International Conference on language resources & evaluation: Granada, Spain, 28-30 May 1998*, pages 581–588. European Language Resources Association.

Najoung Kim, 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. 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)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Maghi King, 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. Technical report.

Margaret King 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*.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

Sungryong Koh, 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*.

Vojtźch Kováź, 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*, pages 540–545. SCITEPRESS-Science and Technology Publications, Lda.

Sabine Lehmann, 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*.

Omer Levy, Felix Hill, Anna Korhonen, Kyunghyun Cho, Roi Reichart, Yoav Goldberg, and Antione Bor-

des, editors. 2016. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Association for Computational Linguistics, Berlin, Germany.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors. 2018. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium.

Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors. 2019. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv preprint arXiv:1605.09090*.

Marina Lloberes, Irene Castellón, and Lluís Padró. 2015. Suitability of ParTes test suite for parsing evaluation. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 61–65, Bilbao, Spain. Association for Computational Linguistics.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Stanford University.

Christopher D Manning. 2006. Local textual inference: it's hard to circumscribe, but you know it when you see it–and nlp needs it.

Christopher D Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Kostadin Mishev, 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*, pages 161–172, Cham. Springer International Publishing.

Diego Mollá and Ben Hutchinson. 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?*, pages 43–50, Columbus, Ohio. Association for Computational Linguistics.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Neha Nayak, 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*, pages 19–23.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Stephan Oepen and Klaus Netter. 1995. Tsnlp - test suites for natural language processing. In *In J. Nerbonne (Ed.), Linguistic Databases (pp. 13 – 36*, pages 711–716. CSLI Publications.

Martha Palmer and Tim Finin. 1990. Workshop on the evaluation of natural language processing systems. *Computational Linguistics*, 16(3):175–181.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Patrick Paroubek, Stéphane Chaudiron, and Lynette Hirschman. 2007. Principles of evaluation in natural language processing. *Traitement Automatique des Langues*, 48(1):7–31.

Katerina Pastra, editor. 2003. *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?* Association for Computational Linguistics, Columbus, Ohio.

Ellie Pavlick. 2017. *Compositional Lexical Entailment for Natural Language Inference*. Ph.D. thesis, University of Pennsylvania.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Ellie Pavlick, 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)*, pages 408–413, Beijing, China. Association for Computational Linguistics.

Ruggero Petrolito. 2018. Word embeddings in sentiment analysis. In *Italian Conference on Computational Linguistics*.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden. Association for Computational Linguistics.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018a. 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)*, pages 513–523, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018b. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018c. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191. Association for Computational Linguistics.

Maja Popović and Sheila Castilho. 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.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2004. Natural language inference via dependency tree mapping: An application to question answering. Technical report.

Yuanyuan Qiu, 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*, pages 209–221. Springer.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Philip Resnik and Jimmy Lin. 2010. 11 evaluation of nlp systems. *The handbook of computational linguistics and natural language processing*, 57.

Philip Resnik, 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*.

Kyle Richardson, Hai Na Hu, Lawrence S. Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *AAAI*, volume abs/1909.07521.

Anna Rogers, Aleksandr Drozd, Anna Rumshisky, and Yoav Goldberg, editors. 2019. *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics, Minneapolis, USA.

Alexey Romanov 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*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Alexis Ross and Ellie Pavlick. 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)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Rachel Rudinger, 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*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. 2010. "ask not what textual entailment can do for you...". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1199–1208, Uppsala, Sweden. Association for Computational Linguistics.

Robert M Schwarcz, John F Burger, and Robert F Simmons. 1970. A deductive question-answerer for natural language inference. *Communications of the ACM*, 13(3):167–183.

P.A.M. Seuren. 1998. *Western Linguistics: An Historical Introduction*.

Karen Sparck Jones. 1994. Towards better NLP system evaluation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Karen Sparck Jones and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer-Verlag, Berlin, Heidelberg.

Ieva Staliūnaitė. 2018. Learning about non-veridicality in textual entailment. Master's thesis, Utrecht University.

Yi Tay, 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*, pages 1565–1575, Brussels, Belgium. Association for Computational Linguistics.

Avijit Thawani, Biplav Srivastava, and Anil Singh. 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*, pages 43–51, Minneapolis, USA. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *11th International Conference on Language Resources and Evaluation (LREC2018)*.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054, Lisbon, Portugal. Association for Computational Linguistics.

Lucy Vanderwende 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*, pages 205–216. Springer.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Proceedings of the Findings of EMNLP*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length.

Alex Wang, 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*, pages 3261–3275.

Alex Wang, 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. *arXiv preprint arXiv:1804.07461*.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. 2019b. Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8.

Aaron Steven White, 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)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Yonghui Wu, 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*, volume 2015, page 1326. American Medical Informatics Association.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *ACL*.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, Michigan. Association for Computational Linguistics.

A Pascal RTE Annotation Guidelines

In the first iteration of the PASCAL RTE challeges, the task organizers were frank in their view that they expected the task definition to change over time. They wrote that "finally, the task definition and evaluation methodologies are clearly not mature yet. We expect them to change over time and hope that participants' contributions, observations and comments will help shaping this evolving research direction." Here, we include snippets from the annotation guidelines for the first three PASCAL RTE challenges:

A.1 RTE1 Guidelines

Given that the text and hypothesis might originate from documents at different points in time, tense aspects are ignored. 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 at True. . . . To reduce the risk of unclear cases, annotators were guided to avoid vague examples for which inference has some positive probability that is not clearly very high. To keep the contexts in T and H self contained annotators replaced anaphors with the appropriate reference from preceding sentences where applicable. They also often shortened the hypotheses, and sometimes the texts, to reduce complexity.

(Dagan et al., 2006)

A.2 RTE2 Guidelines

The data collection and annotation guidelines were revised and expanded . . . We say that t entails h if, typically, a human reading t would infer that h is most likely true. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge. Textual entailment recognition is the task of deciding, given t and h, whether t entails h. Some additional judgment criteria and guidelines are listed below:

- *Entailment is a directional relation. The hypothesis must be entailed from the given text, but the text need not be entailed from the hypothesis.*

- *The hypothesis must be fully entailed by the text. Judgment would be NO if the hypothesis*

includes parts that cannot be inferred from the text.

- *Cases in which inference is very probable (but not completely certain) are judged as YES. For instance, in pair #387 one could claim that although Shapiro's office is in Century City, he actually never arrives to his office, and works elsewhere. However, this interpretation of t is very unlikely, and so the entailment holds with high probability. On the other hand, annotators were guided to avoid vague examples for which inference has some positive probability which is not clearly very high.*

- *Our definition of entailment allows presupposition of common knowledge, such as: a company has a CEO, a CEO is an employee of the company, an employee is a person, etc. For instance, in pair #294, the entailment depends on knowing that the president of a country is also a citizen of that country.*

(Bar-Haim et al., 2006)

A.3 RTE3 Guidelines

As entailment is a directional relation, the hypothesis must be entailed by the given text, but the text need not be entailed by the hypothesis.

- *The hypothesis must be fully entailed by the text. Judgment must be NO if the hypothesis includes parts that cannot be inferred from the text.*

- *Cases in which inference is very probable (but not completely certain) were judged as YES.*

- *Common world knowledge was assumed, e.g. the capital of a country is situated in that country, the prime minister of a state is also a citizen of that state, and so on.*

(Giampiccolo et al., 2007)

	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 4: 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*.

---

unemployment is at an all-time <u>low</u>
▶ unemployment is at an all-time <u>poor</u>

---

aeoi 's activities and <u>facility</u> have been tied to several universities
▶ aeoi 's activities and <u>local</u> have been tied to several universities

---

jerusalem fell to the ottomans in 1517 , remaining under their <u>control</u> for 400 years
▶ jerusalem fell to the ottomans in 1517 , remaining under their <u>regulate</u> for 400 years

---

usually such parking spots are <u>on</u> the side of the lot
▶ usually such parking spots are <u>dated</u> the side of the lot

---

Table 5: 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. <u>Underline words</u> represent the swapped paraphrases.