

Inference is everything: Recasting semantic resources into a unified evaluation framework

Aaron Steven White
Department of Linguistics
University of Rochester

Kevin Duh
HLTCOE
Johns Hopkins University

Pushpendre Rastogi
Department of Computer Science
Johns Hopkins University

Benjamin Van Durme
Department of Computer Science
Johns Hopkins University

Abstract

We propose to unify a variety of existing semantic classification tasks, such as semantic role labeling, anaphora resolution, and paraphrase detection, under the heading of Recognizing Textual Entailment (RTE). We present a general strategy to automatically generate one or more sentential hypotheses based on an input sentence and pre-existing manual semantic annotations. The resulting suite of datasets enables us to probe a statistical RTE model’s performance on different aspects of semantics. We demonstrate the value of this approach by investigating the behavior of a popular neural network RTE model.

1 Introduction

The Recognizing Textual Entailment (RTE) task aims to assess a system’s ability to do textual inference—i.e. derive valid conclusions from textual clues (Dagan et al., 2006, 2013; Bar-Haim et al., 2006; Giampiccolo et al., 2007, 2009; Ben-tivogli et al., 2009, 2010, 2011). In this task, a system judges whether “typically, a human reading [the sentential context, or *text*] T would infer that [the sentential hypothesis] H is most likely true” (Dagan et al., 2006).

Recent efforts in textual inference have focused on the Stanford Natural Language Inference (SNLI) dataset. SNLI is made up of hundreds of thousands of text-hypothesis pairs, wherein the texts are image captions drawn from the Flickr30k corpus (Young et al., 2014) and the hypotheses are elicited from crowdsourcing workers based on those captions (but not the corresponding image). While SNLI has led to significant methodological improvements, its collection protocol does not lend itself to understanding the types of semantic knowl-

edge necessary for properly understanding a particular example. 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 contrast, datasets such as FraCaS (Cooper et al., 1996) are precisely designed to illustrate a range of semantic phenomenon that a text understanding system should handle. But though this careful design enables fine-grained probes into a system’s semantic capabilities, FraCaS-like datasets tend not to be large-scale enough for recent work in data-driven computational semantics. Asking experts, such as those who constructed FraCaS, to author hundreds of thousands of examples is not practical, just as the existing elicitation protocol behind SNLI will not lead to cleanly partitioned sets of examples that focus specifically on certain kinds of semantic inference.

Our proposal is to leverage existing large-scale semantic annotation collections as a source of targeted textual inference examples. This strategy requires only minor effort in developing dataset-specific generation capabilities to convert annotations into a shared universal representation: natural language sentences.

We demonstrate the use of this strategy in two steps. First, we construct three recasted datasets from three existing semantic resources that target three distinct semantic phenomena: (i) the Semantic Proto-Roles v1 (SPR) dataset (Reisinger et al., 2015), which contains likelihood judgments about the semantic *proto-role* properties (Dowty, 1991) of verbal arguments found in PropBank (Palmer et al., 2005), (ii) the FrameNet Plus (FN+) dataset, which contains likelihood judgments about the paraphrase validity of *frame triggers* (Pavlick et al., 2015), and (iii) the Definite Pronoun Resolution (DPR) dataset, which contains annotations relevant to complex

anaphora resolution (Rahman and Ng, 2012). We use these recasted datasets to train a recent neural RTE model (Bowman et al., 2015) and measure its performance. We show that complex anaphora is the most difficult semantic phenomenon for neural RTE models to capture, followed by predicting thematic proto-role properties. Perhaps unsurprisingly, given the nature of the RTE task, paraphrasing seems to be the easiest phenomenon to model.

In the next section (§2), we discuss previous work in RTE, focusing in particular on the development of RTE datasets. We then discuss our data creation process (§3) as well as the results of a small validation (§4). Finally, we report on the setup and results of our three experiments (§5) and then conclude (§6).

2 Background and Prior Work

The current paper touches on both the broad theme of understanding continuous approaches to natural language understanding as well as the more narrow focus on textual entailment. We begin by discussing how the current paper fits within the broader context and then specify its place within textual entailment.

2.1 Approaches to logical form

All approaches to natural language understanding utilize intermediate logical forms that are interpretable to varying degrees. On one end of the spectrum are approaches that utilize declarative logical forms. In such approaches, semantic parsers first convert a sequence into a meaning representation that expresses the semantics needed for inference. In this case, each individual component of the logical form is clearly interpretable. Tremendous energies within computational linguistics have been spent on building declarative, component-wise-interpretable logical forms such as Hobbsian Logic (Hobbs, 1985), Discourse Representation Theory (Kamp et al., 2011), the Rochester Interactive Planning System (Allen et al., 2007), Minimal Recursion Semantics (Copestake et al., 2005), Episodic Logic (Schubert and Hwang, 2000), Combinatory Categorical Grammar (Steedman, 2000), Semantic Role Labeling (Gildea and Jurafsky, 2002), Framenet Parsing (Fillmore et al., 2003) and Abstract Meaning Representation (Banarescu et al., 2013).

Opposite the above approaches are methods that utilize vector space-based logical forms. Recent

work on word and string embeddings (Mikolov et al., 2013) has produced vector space representations that can be induced from large corpora in an unsupervised manner that have been used to initialize the training of neural networks for tasks as complex as English-to-French machine translation (Sutskever et al., 2014). Vector space-based intermediate forms are not commonly recognized as logical forms but in light of recent work (Bouchard et al., 2015) it seems worthwhile to reconsider this view.

An argument in favor of declarative, interpretable logical forms is that one can directly observe the specific mistakes made by a system in the interpretive process of mapping natural language strings to logical forms—e.g., it is possible to find out whether a prepositional phrase was attached incorrectly, or the wrong sense of a particular word was selected, causing a cascade of downstream errors. Neural systems that use vector space representations for textual inference, instead of logical forms, lack such modularity and interpretability, and therefore it is very difficult to figure out the cause of a particular error in a neural network.

Much prior work has aimed to improve the interpretability of neural networks, focusing in particular on extracting rules from the activations of feed forward networks (Towell and Shavlik, 1993; Thrun, 1993; Fu, 1994; Thrun, 1995). In recent years, this focus has shifted to understanding and visualizing other architectures such as Convolutional Neural Networks and Recurrent Neural Networks (Zeiler and Fergus, 2014; Karpathy et al., 2015), though the guiding principle remains the same: understanding the behavior of neural networks in terms of its activations.

The current paper also presents a strategy for understanding the behavior of neural RTE systems used for solving the task of RTE, but we take a different route. Instead of explaining the behavior of neural networks in terms of its parameters and activations, we benchmark their performance on datasets that each require distinct types of semantic reasoning for high performance. In this sense, our motivation and strategy is similar to the reasoning behind the bAbI dataset for question answering (Weston et al., 2016). Weston et al. argue that, in order to measure the progress towards building dialogue agents, it can be useful to evaluate the ability of systems to perform different kinds of question answering tasks that require specific types

of reasoning. Our strategy for building specific datasets that can probe the ability of machine learning systems to perform specific types of reasoning is similar to theirs; however, instead of constructing completely artificial datasets, we *recast* datasets constructed on top of natural text.

2.2 Approaches to textual entailment

Research in textual entailment, at least in its most recent form, was catalyzed by the RTE shared tasks (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007, 2009; Bentivogli et al., 2009, 2010, 2011). With each iteration of the shared task, manually annotated examples were created for testing competing systems. But even after multiple iterations, the amount of available data for RTE was still small. The Sentences Involving Compositional Knowledge (SICK) corpus was released with the goal of alleviating this problem (Marelli et al., 2014). A significant further contribution was made with the Stanford Natural Language Inference (SNLI) corpus, which uses crowdsourcing to gather two orders of magnitude more examples than all previous datasets (Bowman et al., 2015). This dataset enabled full supervised training of powerful machine learning models like neural networks. A number of researchers have pursued this direction by applying completely supervised neural models for sequential data to the problem of textual entailment (Rocktäschel et al., 2015; Mou et al., 2015; Shuohang and Jing, 2015; Liu et al., 2016; Cheng et al., 2016; Parikh et al., 2016; Munkhdalai and Yu, 2016).

But though the state of the art performance of neural sequential models has steadily increased over the past year, it appears that this area has reached a point where the paradigm of training and evaluating on a single general-purpose RTE dataset has become insufficient for reaching the next level of improvements. It is still informative to measure the performance of a new RTE model on the SNLI dataset, but this black-box evaluation does not help us understand the fine-grained aspects of a model’s capability in performing particular types of natural language inference, such as its ability to handle coreference, paraphrasing, or its ability to judge thematic properties of a named entity. The issue of lack of understandability is especially important for neural models, which are notoriously difficult to interpret.

To address this issue, we take inspiration from

the FraCaS dataset (Cooper et al., 1996) and construct a suite of semantic datasets that separately test a system’s ability to perform individual bits of interpretation such as paraphrasing, semantic role labeling, and coreference. But in contrast to the original FraCaS data set, which is relatively small and which could not support the training of purely lexical neural RTE classifiers, we pursue the strategy of automatically converting *semantic classifications*—i.e., human judgments about semantic properties—into labeled examples for textual entailment. This strategy allows us to construct textual entailment datasets that are of the same order of magnitude as SNLI—and hence support data-driven training of large neural networks—but that are also focused on specific semantic properties.

With our strategy we can also quantify the types of semantic phenomenon that an existing *semantically undifferentiated* dataset contains. For example, if a neural model trained on the SNLI dataset performs poorly on a test set from another domain that exercises the trained model’s ability to perform anaphora resolution, then it can be inferred that either the original dataset did not contain enough examples of anaphora resolution, or that the statistical model failed to capture that phenomenon.

Finally, we note that the idea of converting question answer pairs into text-hypothesis pairs is not novel since the RTE dataset for the second RTE Shared task was created by manually converting existing Information Extraction, Information Retrieval and QA pairs from manually curated datasets such as ACE, MUC, TREC and CLEF (Bar-Haim et al., 2006). The difference between the earlier work and our work is that we show that automatic conversion of some semantic datasets can be done with a high enough quality to create large scale RTE datasets whereas previous work relied on human annotators to convert the semantic resources into RTE datasets.

3 Data Creation Process

In order to create annotated RTE datasets that can probe specific aspects of understanding, our strategy is to rewrite semantic classifications into the form of textual entailment pairs. As mentioned above, we define a semantic classification dataset to be a text corpus, along with manual annotations about semantic aspects of the data. Here, we describe how to apply this strategy to the SPR, FN+, and DPR datasets, but there exist many such

ID	Dataset	Label	Text	Hypothesis
1	SPR	entailed	The network must refund money to the advertisers and loses considerable revenue and prestige.	The network was altered or somehow changed during or by the end of the losing.
2	SPR	entailed	He turned himself in to authorities in New York earlier this year.	He changes location during the turning.
3	SPR	not-entailed	Later, he marketed glue.	He changes location during the marketing.
4	SPR	not-entailed	So he asked the IRS if the plan would work.	The asking caused a change in the IRS.
5	FN+	entailed	An agreement is to be signed in late 10/92.	An agreement is to be inked in late 10/92.
6	FN+	entailed	So our work must continue.	So our labor must continue.
7	FN+	not-entailed	Friday had beautiful weather.	Friday had beautiful forecast.
8	FN+	not-entailed	Your support will help them go to work.	Your support will allow them go to work.
9	DPR	entailed	The bird ate the pie and it died.	The bird ate the pie and the bird died.
10	DPR	not-entailed	The bird ate the pie and it died.	The bird ate the pie and the pie died.
11	DPR	entailed	The bird ate the pie and it was ruined.	The bird ate the pie and the pie was ruined.
12	DPR	not-entailed	The bird ate the pie and it was ruined.	The bird ate the pie and the bird was ruined.

Table 1: Examples of *text-hypothesis* pairs generated from the SPR, FN+ and DPR datasets.

datasets to which this strategy can be applied.

The SPR dataset The *thematic proto-role* theory was proposed by Dowty (1991) who argued that the *Agent* and *Patient* semantic roles should be viewed not as categories, but rather as an agglomeration of entailments that arguments satisfy in the context of an event. For example, consider the following yes-or-no question, “Did the *argument* have volitional involvement in the *event*?” Here the emphasized *event* and *argument* tokens are place-holders that refer to the sentence’s verbal predicate and its argument respectively. This question template specifically identifies whether the subject, that typically takes upon the role of an Agent, performed the action willfully or not. This question can be unambiguously answered by humans in most contexts and textual inference systems should be able to answer such questions. Note that the answer to such questions depends on world knowledge and that the answer to the question above will be *yes* for the sentence, “The salesman sold her shoes.” but it will be *no* for the sentence “The antibody killed the virus.” The SPR dataset collected by Reisinger et al. (2015) contains likelihood judgments for twelve *proto-role* properties that subjects and objects in a sentence possess (see also White et al. 2016): *instigation*, *volition*, *awareness*, *Sentience*, *movement*, *physical existence*, *existence before/during/after the event*, *change of possession*, *change of state*, and *stationarity*. The example questions above tests the *volition* property, whereas a question like “was the *argument* aware of being involved in the *event*” is used to test *awareness*.

Judgments about proto-role properties can be readily converted to a pair of text-hypothesis sen-

tences, required for RTE. For example, the question “Was the *argument* sentient?” and its answer *yes* can be converted to the hypothesis “The *argument* was sentient.” and the entailment relation *entailed*. This is a simple and inexpensive way of creating entailment pairs, with the benefit that this annotation scheme probes for fundamental semantic information from an annotator. Also, note that since Reisinger et al. collect annotations for the twelve types of proto-role properties mentioned above, one can automatically subcategorize the errors made by a neural RTE model into these 12 categories, further aiding in interpretation.

One potential criticism of our method is that, because our hypothesis sentences are constructed by filling in templates, they do not have the same syntactic diversity as the free elicitation method used by Bowman et al.. We argue that this is not a problem for two reasons. First, since our goal is to distinguish between the kinds of semantic phenomenon that can be accurately modeled by statistical RTE models, the lack of diversity is no obstacle as long as the particular phenomenon that we wish to probe is being covered properly. Second, even the method used by Bowman et al. of enlisting workers on the Amazon Mechanical Turk (AMT) Platform to write hypotheses sentences in response to an image caption is not without its drawbacks since their method introduces artifacts such as the fact that the *hypotheses* sentences in SNLI are on average half the length of the *text* prompts. We believe that this happens because workers on AMT have an incentive to spend the least amount of time possible in constructing their responses. In Table 2, we list a few such examples.

As an example of the conversion process, given

▶	Three women enjoying a balloon joyride. Three women are on a balloon ride.
▶	A woman sings into a microphone indoors. a women sings
▶	The kid is sliding down a tan plastic slide. The kid is sliding.
▶	A black dog is playing in water with a green toy. the dog has a toy
▶	A woman with glasses and a pink hat rides her bike. A woman with glasses and a pink hat rides her bike

Table 2: Examples of artifacts in the SNLI dataset that promote hypothesis sentences to be substrings of the evidence sentences, specially in case of entailments. The bullet marked sentences are the *evidence* sentences and the *hypothesis* sentences below them.

Dataset	Sentences	Label Percentage	
		Entailed	Not-Entailed
FN+	154,605	43.45	56.55
SPR	154,607	34.80	65.20
DPR	3,661	49.99	50.01
Total	312,873	39.13	60.87
SNLI [†]	569,033	33.41	66.59

Table 3: Number of *text-hypothesis* pairs generated from each dataset along with percentage of entailing v. non-entailing sentences. SNLI included for comparison.

the sentence, *The quake postponed the third and fourth games of the World Series*, Reisinger et al. (2015) highlighted the subject phrase, *the quake* and asked *How likely or unlikely is it that the quake was aware of being involved in the postponing?* The raters answered by choosing a rating on a scale from 1 to 5 with 1 being the least likely and 5 being extremely likely. This particular question probes the property of *awareness*, which was one of Dowty’s *proto-agent* properties, which is correlated with subjecthood—i.e. it tends to be that subjects are aware (though not all subjects are). For example, changing the the subject phrase *the quake* to *the president* would presumably change the likelihood rating from 1 to 5 for this question.

We rewrite the question above as: *The quake was aware of being involved in the postponing* and map a high rating of 4 and above to *entailed* and a rating of below 4 to *not-entailed*. During the conversion process we ensure grammaticality of the generated hypothesis sentences in the following ways. First, we calculate whether the argument being replaced is plural or singular and based on its number category we use *was* or *were* in the hypothesis sentence. Also we discard question-answer pairs rated *NA* since those question-answer pairs tend to be unintelligible. These simple heuristics appeared to be sufficient to ensure the grammaticality of hypothesis sentences.

Data	Accuracy	Grammaticality
FN+	85	77
SPR	94	92
DPR	98	96
SNLI	91	96

Table 4: Accuracy of the labels assigned to the RTE pairs and the grammaticality of the hypothesis sentences. 100 random RTE pairs from each dataset were selected and each pair was assigned a value of 1 if it was correctly labeled/grammatical and 0 otherwise. We report the average score as a percentage in two separate columns for each dataset.

The FN+ dataset A *frame* is the primary unit of lexical analysis in *Frame Semantics* that captures the vital properties of a concept, situation, or event (Fillmore et al., 2003). The largest resource for frame annotated sentences with information about evoked frames, their *trigger* phrases, and frame arguments is the FrameNet dataset (Fillmore and Baker, 2001), which despite its scale, still suffers from lexical sparsity. In order to alleviate this problem of lexical sparsity Rastogi and Van Durme (2014) use the Paraphrase Database (Ganitkevitch et al., 2013) to automatically paraphrase *trigger* tokens that evoke frames inside sentences from the FrameNet dataset. These paraphrases are noisy, and their quality is not high enough for our use. However, these paraphrases were subsequently manually rated by Pavlick et al., who asked annotators to “judge each paraphrase in terms of how well it preserved the meaning of the original sentence” (Pavlick et al., 2015). These ratings were collected on a scale from 1 to 5, where 5 meant that the paraphrase retained all of the meaning of the original sentence and 1 meant that paraphrase did not mean anything close to the original phrase. We generate our *text-hypothesis* pairs using the manual judgments of meaning retention on these *sentence-paraphrase* pairs collected by Pavlick et al..

While the *sentence-paraphrase* pairs that are labeled *entailed* and rated 3.0 and above are usually grammatically correct, the sentences that are rated below 3.0 and labeled *entailed* sometimes contain a grammatical errors, and some are rated *neutral* or contradictory. Therefore, we remove sentences rated less than or equal to 3 and greater than 2.5. All of the sentence pairs that were rated less than 2.5 were not valid entailments and they were labeled as *not-entailed*.

As an example, consider the following sentence from FrameNet: *[s]o our work must continue*. The

FN+	12:06 a.m. hrh: i was per the berkeley main library when it hit.
SPR	Me existed as a physical object.
DPR	John could not understand his waiter, because the his waiter spoke Spanish.
SNLI	A person in on concrete.

Table 5: Examples of ungrammatical hypothesis sentences from each of our datasets and SNLI.

word *work* triggers a frame and is replaced by its paraphrase *labor* by Pavlick et al. to create the sentence *so our labor must continue*. We consider the first sentence to be the evidence and the second sentence to be the hypothesis. The annotators on Mechanical Turk gave this pair of sentences a rating of 4, and so we consider this pair of sentences as an instance of the *entailed* relation.

The DPR dataset Definite Pronoun Resolution is the problem of identifying the correct antecedent for a pronoun in *twin* sentences. We explain the concept of twin sentences with the following example: *The bee landed on the flower because it wanted pollen* and *The bee landed on the flower because it had pollen* it resolves to *bee* and *flower* in sentence 1 and 2, respectively. In order to assign the correct antecedent of *it* in both sentences, a computational system would presumably need world-knowledge about bees and flowers that it could not glean from textual co-occurrence data. The DPR dataset is a collection of such problems and their solutions, collected by Rahman and Ng (2012) as a step towards solving the Winograd Schema Challenge (Hector et al., 2012). Each DPR coreference problem-solution pair can be converted to two annotated entailment problems by substituting the co-referent pronoun with the two optional entities that it could refer to. For example, sentence 1 is rewritten to, *The bee landed on the flower because the flower wanted pollen* and *the bee landed on the flower because the flower had pollen*. The second re-write is entailed whereas the first re-write is not-entailed. Table 3 summarizes the constructed datasets.¹

4 Data Validation

In the previous section, we described our procedure for automatically converted human judgments about semantic properties of sentences into labeled examples for textual entailment. Since our data is

¹For the SNLI statistics, we map the two categories of *contradiction* and *neutral* to *not-entailed*.

automatically generated, we performed manual validation to ensure that the generated data was high quality. To conduct this validation, we assessed a small subset of our recasted datasets as well as the SNLI dataset. We randomly sampled 100 RTE pairs from each of the four datasets, and then a single annotator rated those 400 RTE pairs on two criteria of grammaticality and correctness. The results of the manual validation presented in Table 4 show that the data quality of the DPR and SPR datasets is on par with the quality of the RTE pairs in the SNLI datasets. The grammaticality of the hypothesis sentences in the FN+ dataset is worse than the other three datasets, but its accuracy is reasonably high.

Tables 5 and 6 show examples of ungrammatical hypothesis sentences and incorrectly labeled RTE sentence pairs to illustrate the types of errors that we make in comparison to the errors made by mechanical turkers.

5 Experiments and Results

After creating our datasets we conducted experiments to measure the variation in performance of neural RTE models trained using those datasets. Our first hypothesis was that by analysing the variation in accuracy of a neural RTE model trained on different datasets on the same test set, such as the SNLI dataset, one could gain insights into the behavior of the model. Secondly such variations can reveal interesting information about the SNLI dataset itself. For the experiments we first split our three datasets into train, validation and test sets in the proportion of 80 : 10 : 10. We trained our model on the training sets of FN+, SPR, DPR and SNLI. Prior to training we converted the SNLI test set to a binary scheme by replacing both “neutral” and “contradiction” class labels with “not-entailed”. We used the LSTM-based neural RTE model described by Bowman et al. (2015) which was their best performing individual neural model. The model first embeds the words using 300 dimensional word embeddings created using the Glove method (Pennington et al., 2014). After that two LSTM neural networks (Hochreiter and Schmidhuber, 1997) independently encode the text and hypothesis sentences into 100 dimensional vectors. These representations are concatenated and input to a 3-layer deep 200 dimensional neural network classifier. The entire network is trained by maximizing the cross-entropy of the input-output pairs over

<i>Assigned Category</i>	<i>Text and Hypothesis</i>
(FN+) Not Entailed	► The steps passed along the path on the other side of the wall under which i crouched. the steps passed along the path on the other outboard of the wall under which i crouched.
(SPR) Not Entailed	► The machine employs reduced instruction-set computing, or RISC, technology. Reduced instruction-set computing, or RISC, technology existed after the employing stopped.
(DPR) Entailed	► A series of injections are used to battle a type of cancer in patients because they have a special type of drug which counteracts this sickness. A series of injections are used to battle a type of cancer in patients because patients have a special type of drug which counteracts this sickness.
(SNLI) Neural	► two guys playing music with a band The guys play music

Table 6: Examples of incorrectly labeled RTE pairs from each of our datasets and SNLI.

the entire dataset using the AdaDelta (Zeiler, 2012) update rule with L2-regularization and Dropout. We evaluated each of our models on all the test sets to obtain the results in Table 7.

Table 7 shows that when the neural RTE model is trained and tested on the same dataset, the performance on the test set is high (above 80%) for FN+, SPR, and SNLI. This implies that these three tasks are amenable to the application of neural sequence models with the FN+ and SPR dataset being comparable in their difficulty. Moreover we see that the performance of all 4 of the models is equal to chance performance on the DPR dataset and it is consistently lower than other cross-evaluation results shown on the off-diagonals of Table 7. This shows that complex anaphora resolution is more difficult. Also since the performance of the SNLI trained model is the least on the DPR dataset therefore we can conclude that the phenomenon of anaphora resolution occurs less often than paraphrasing or proto-role resolution in the SNLI dataset. In our random sample of 100 sentences from the SNLI dataset we did not find a single example where pronoun coreference resolution was required to predict the label correctly. In fact only the following sentence could have been rewritten as a pronoun resolution problem in the SNLI dataset, “a man speaking to a woman in a grocery store as he selects a carton of juice.” The *hypothesis* corresponding to this *text* was “A man is complimenting a woman on her jacket.” but if the *hypothesis* was “The women selects a carton of juice.” then it would have been a coreference resolution problem albeit not a difficult one. Finally we can see that the SNLI trained model achieved 62.0% on the FN+ test set. While better than a most frequent label baseline (56%), this is still considerably worse than the FN+ model (80.5%), optimized for paraphrastic inference under single word replacement. We believe this is because the sen-

Train Set	FN+	SPR	DPR	SNLI
FN+	80.5	60.0	49.5	62.0
SPR	65.8	80.6	50.7	52.3
DPR	19.2	65.2	49.5	65.7
SNLI	62.0	57.6	48.8	85.3

Table 7: Accuracy under 0-1 loss of predicting the entailment relation. Each cell describes the accuracy of a model trained on the corresponding row’s training set and tested on the corresponding column’s test set

tences in FN+ contain language that is rarely seen and contain more subtle differences compared to the differences between the text and the hypothesis in the SNLI dataset. We believe this is why the model trained on SNLI does not perform well on any of the other datasets.

As a second illustration, we analyzed the SNLI trained RTE model’s performance on the SPR test set by dissecting the overall performance of the model by the proto-role properties that the entailment pairs were generated from. Note that all the categories appear equally in the test data. The results, shown in Figure 1, show that entailments belonging to “Change of State” category caused the highest number of errors. Based on looking at examples, such as sentence 4 in Table 1, we believe that this happens because such entailments don’t follow lexical patterns. On the other hand information about, “Stationarity” and “Change of possession” can be captured by neural models because the entailments are tightly coupled to the argument tokens.

6 Conclusion

We argue for constructing a suite of large-scale textual inference datasets that probe specific aspects of semantics, in order to analyze a statistical RTE model’s ability of “understanding” distinct semantic phenomena. To construct such datasets we presented a general strategy of converting semantic

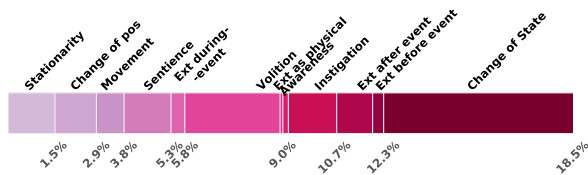


Figure 1: Percentage of errors of RTE model by *proto-role* properties on the SPR test set. The percentage numbers at the bottom are the contribution of the category above to the total errors. Error percentages that are close to each other are omitted for clarity.

classification examples to annotated textual inference pairs that can be used to create large datasets for free on which even neural models for RTE can be trained. Further we used these datasets to gain insights into the behavior of a popular neural RTE model and the SNLI dataset itself. The variation in the performance of that model on the three datasets showed that neural models for natural language understanding recognise lexical variations or paraphrasing much better than anaphora resolution. Recently (Chen et al., 2016) also presented a similar conclusion after manually analysing the errors made by neural systems on a reading comprehension task. Our approach can be thought of as an automatic way of automating the manual error analysis so that it can be used iteratively in a larger system and it can remove the requirement of a human in the loop. Our results also strongly suggested that the SNLI dataset does not contain examples of anaphora resolution which we validated manually. Our datasets and annotations are available at url:anon4review.

In future work, we plan to execute our strategy on labeled data for Word Sense Disambiguation and Prepositional Phrase attachment resolution, among other semantic resources, because we believe that such diverse datasets will require sophisticated RTE models that combine world knowledge with the pattern recognition abilities of neural networks. For example, given the sentence *The dog wagged its tail* and a known sense of the dog the following hypotheses sentences can be generated: *The dog is a domestic dog* and *The dog is a wiener*. The former hypothesis is *entailed* but the latter is not. Disambiguating between word senses and deciding the correct governor of a prepositional phrase requires world knowledge and RTE examples generated from such sources, even though they are generated automatically unlike the FraCaS dataset, will help researchers build robust statistical models for RTE since each semantic classification dataset high-

lights a particular type of semantic phenomenon that a robust system for RTE must model.

References

- James Allen, Mehdi Manshadi, Myroslava Dzikovska, and Mary Swift. 2007. Deep linguistic processing for spoken dialogue systems. In *Workshop on Deep Linguistic Processing*. ACL, pages 49–56.
- Laura Banarescu et al. 2013. [Abstract meaning representation for sembanking](#). In *7th Linguistic Annotation Workshop and Interoperability with Discourse*. ACL, Sofia, Bulgaria, pages 178–186. <http://www.aclweb.org/anthology/W13-2322>.
- Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. In *Second PASCAL challenges workshop on RTE*. volume 6, pages 6–4.
- Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa T. Dang, and Danilo Giampiccolo. 2010. The sixth PASCAL recognizing textual entailment challenge. In *TAC*.
- Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The seventh PASCAL recognizing textual entailment challenge. In *TAC*.
- Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC*. volume 9, pages 14–24.
- Guillaume Bouchard, Sameer Singh, and Théo Trouillon. 2015. On approximate reasoning capabilities of low-rank vector spaces. In *2015 AAAI Spring Symposium Series*.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the ACL*. ACL.
- Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. [Long short-term memory-networks for machine reading](#) <http://arxiv.org/abs/1601.06733v6>.
- Robin Cooper et al. 1996. Using the framework FRACAS a framework for computational semantics. Technical report, The FRACAS consortium.
- Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation* 3(2-3):281–332.

- 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 textual Entailment*, Springer, pages 177–190.
- Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on HLT* 6(4):1–220.
- David Dowty. 1991. Thematic proto-roles and argument selection. *language* pages 547–619.
- Charles Fillmore, Christopher Johnson, and Miriam R.L. Petruck. 2003. Background to framenet. *International Journal to Lexicography* 16(3).
- Charles J. Fillmore and Collin F. Baker. 2001. Frame semantics for text understanding. In *WordNet and Other Lexical Resources Workshop*. NAACL.
- LiMin Fu. 1994. [Rule generation from neural networks](#). *IEEE Transactions on Systems, Man, and Cybernetics* 24(8):1114–1124. <https://doi.org/10.1109/21.299696>.
- Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *NAACL-HLT*.
- Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. 2009. The fourth PASCAL recognizing textual entailment challenge. In *TAC*.
- Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *ACL-PASCAL workshop on textual entailment and paraphrasing*. ACL, pages 1–9.
- Daniel Gildea and Daniel Jurafsky. 2002. [Automatic labeling of semantic roles](#). *Computational linguistics* 28(3):245–288. <http://www.cs.cornell.edu/courses/CS775/2002fa/papers/gildea-jurafsky-cl02.pdf>.
- Levesque Hector, Davis Ernest, and Morgenstern Leora. 2012. [The winograd schema challenge](#). In *Knowledge Representation and Reasoning Conference*. <http://www.aaai.org/ocs/index.php/KR/KR12/paper/view/4492/4924>.
- Jerry R. Hobbs. 1985. Ontological promiscuity. In *Proceedings of the ACL*. ACL, pages 60–69.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.
- Hans Kamp, Josef Van Genabith, and Uwe Reyle. 2011. Discourse representation theory. In *Handbook of philosophical logic*, Springer, pages 125–394.
- Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.
- Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. [Learning natural language inference using bidirectional lstm model and inner-attention](#) <http://arxiv.org/abs/1605.09090v1>.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*. pages 216–223.
- Tomas Mikolov, Wen Tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *NAACL-HLT*. pages 746–751.
- Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2015. [Natural language inference by tree-based convolution and heuristic matching](#) <http://arxiv.org/abs/1512.08422v3>.
- Tsendsuren Munkhdalai and Hong Yu. 2016. [Neural tree indexers for text understanding](#) <http://arxiv.org/abs/1607.04492v1>.
- Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics* 31:1.
- Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. [A decomposable attention model for natural language inference](#) <http://arxiv.org/abs/1606.01933v1>.
- Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Drezde, and Benjamin Van Durme. 2015. Framenet+: Fast paraphrastic tripling of framenet. In *ACL*. ACL, Beijing, China.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. ACL.
- Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *Joint Conference on EMNLP and CoNLL*. ACL, pages 777–789.
- Pushpendre Rastogi and Benjamin Van Durme. 2014. [Augmenting framenet via PPDB](#). In *Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. ACL. <http://www.aclweb.org/anthology/W14-2901>.
- Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. [Semantic proto-roles](#). *Transactions of the ACL* 3:475–488. <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/674>.

- Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664* .
- Lenhart K. Schubert and Chung Hee Hwang. 2000. Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding. *NLP And Knowledge Representation* pages 111–174.
- Wang Shuohang and Jiang Jing. 2015. [Learning natural language inference with lstm](http://arxiv.org/abs/1512.08849v1) <http://arxiv.org/abs/1512.08849v1>.
- Mark Steedman. 2000. *The Syntactic Process*, volume 24. MIT Press.
- Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. pages 3104–3112.
- Sebastian B. Thrun. 1993. Extracting provably correct rules from artificial neural networks. Technical report, University of Bonn.
- Sebastian B. Thrun. 1995. [Extracting rules from artificial neural networks with distributed representations](http://papers.nips.cc/paper/3869-conditional-neural-fields.pdf). In *Advances in Neural Information Processing Systems*, pages 505–512. <http://papers.nips.cc/paper/3869-conditional-neural-fields.pdf>.
- Geoffrey G. Towell and Jude W. Shavlik. 1993. [Extracting refined rules from knowledge-based neural networks](https://doi.org/10.1007/BF00993103). *Machine Learning* 13(1):71–101. <https://doi.org/10.1007/BF00993103>.
- Jason Weston, Antoine Bordes, Sumit Chopra, Sasha Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *ICLR*.
- Aaron Steven White, 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*. Association for Computational Linguistics, Austin, TX, pages 1713–1723.
- 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.
- Matthew Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *ECCV*. Springer, pages 818–833.
- Matthew D. Zeiler. 2012. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .