Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework

Aaron Steven White  
University of Rochester

Pushpendre Rastogi  
Johns Hopkins University

Kevin Duh  
Johns Hopkins University

Benjamin Van Durme  
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; BENTIVOGLI 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 knowledge 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 recast 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:1 (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

---
1 These recasted datasets are made publicly available at http://decomp.net.
(iii) the Definite Pronoun Resolution (DPR) dataset, which contains annotations relevant to complex anaphora resolution (Rahman and Ng, 2012). We use these recast 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; Pennington et al., 2014) 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 task (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007, 2009; Bentivogli et al., 2009, 2010, 2011). With each iteration of this 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). SNLI enabled fully 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; PuriKh 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 targeted datasets that separately test a system’s ability to perform individual bits of interpretation such as paraphrasing, semantic role labeling, and coreference. 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 classifications—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 semanti-
cally 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: 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 main contribution of the current work is to show that such conversion need not be done manually; automatic conversion of some semantic datasets can be done with a high enough quality to create large-scale 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 of a particular meaning-related aspect of the data. Here, we describe how to apply this strategy to the SPR, FN+, and DPR datasets, but there exist many further datasets to which this strategy can be applied.

The SPR dataset  Semantic Proto-Role Labeling (SPRL) is the problem of assigning a likelihood value for a particular proto-role property holding of a particular argument of a particular predicate (Reisinger et al., 2015; White et al., 2016; Teichert et al., 2017). These proto-role properties are inspired by the thematic proto-role theory proposed by Dowty (1991), who argued that, for the purpose of determining the mapping from predicates’ semantic roles to its syntactic arguments, semantic roles should be viewed not as categories, but rather as sets of entailments that arguments must satisfy in the context of an event kind.

For purposes of recasting, we use the SPR1 dataset, which was collected by Reisinger et al. (2015) and contains likelihood judgments for the twelve proto-role properties listed in Table 2.

These judgments were collected by providing the annotator with a sentence in which a predicate and an argument of that predicate were highlighted and asking them to answer, on a five-point scale from 1 (very unlikely) to 5 (very likely), how likely or unlikely each property was to hold of the argument in the context of the predicate.2

For example, given (1), with the antibody as the highlighted argument and killed as the highlighted predicate, the annotator’s job was to answer questions like the one in (2).

(1) The antibody killed the virus.

(2) How likely or unlikely is it that the antibody caused the killing to happen?

For the purposes of SPRL task, Teichert et al. (2017) propose to collapse the five-point scale to a binary variable by mapping response 1–3 to not-entailed and 4–5 to entailed and predicting the resulting binary variable. Collapsing across properties, the current state-of-the-art F1 for the resulting task of predicting this binary variable is reported by Teichert et al. (2017) at 81.7.

Binarized proto-role property judgments can be readily converted to text-hypothesis pairs by simply treating the original sentence as the text and converting the questions listed in Table 2 to statements for use as hypotheses. For example, (1) would be treated as a text, and (3) would be treated as a hypothesis generated from (2).

(3) The antibody caused the killing to happen.

In this case, the annotator gave a 5 (very likely) response to (2), and so in our recasted dataset, the resulting pair is labeled entailed.

---

2 Annotators were also given the option of saying that the question was not applicable (NA). We filter these these responses from our dataset.

<table>
<thead>
<tr>
<th>Role property</th>
<th>How likely or unlikely is it that...</th>
</tr>
</thead>
<tbody>
<tr>
<td>initiation</td>
<td>ARG caused the PRED to happen?</td>
</tr>
<tr>
<td>victimation</td>
<td>ARG chose to be involved in the PRED?</td>
</tr>
<tr>
<td>awareness</td>
<td>ARG was/were aware of being involved in the PRED?</td>
</tr>
<tr>
<td>sentiment</td>
<td>ARG was/were sentient?</td>
</tr>
<tr>
<td>change of location</td>
<td>ARG changed location during the PRED?</td>
</tr>
<tr>
<td>exists as physical</td>
<td>ARG existed as a physical object?</td>
</tr>
<tr>
<td>existed before</td>
<td>ARG existed before the PRED began?</td>
</tr>
<tr>
<td>existed during</td>
<td>ARG existed during the PRED?</td>
</tr>
<tr>
<td>existed after</td>
<td>ARG existed after the PRED stopped?</td>
</tr>
<tr>
<td>change of possession</td>
<td>ARG changed possession during the PRED?</td>
</tr>
<tr>
<td>change of state</td>
<td>ARG was/were altered or somehow changed during or at the end of the PRED?</td>
</tr>
<tr>
<td>stationary</td>
<td>ARG was/were stationary during the PRED?</td>
</tr>
<tr>
<td>location of event</td>
<td>ARG described the location of the PRED?</td>
</tr>
<tr>
<td>physical contact</td>
<td>ARG made physical contact with someone or something else involved in the PRED?</td>
</tr>
<tr>
<td>was used</td>
<td>ARG was/were used in carrying out the PRED?</td>
</tr>
<tr>
<td>pred changed arg</td>
<td>The PRED caused a change in ARG?</td>
</tr>
</tbody>
</table>

Table 2: Questions posed to SPR annotators.
write hypotheses sentences in response to an im-
formation about evoked frames, their
resource for frame annotated sentences with infor-
matics information from an annotator. Also, note that,
since Reisinger et al. collect annotations for the
twelve types of proto-role properties mentioned
above, the errors made by a neural RTE model
can be automatically subcategorized 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 suggest 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 ac-
curately modeled by statistical RTE models, the
lack of diversity is not an 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 im-
age 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 incen-
tive to spend the least amount of time possible in
constructing their responses. In Table 3, we list a
few such examples.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A woman with glasses and a pink hat rides her bike.</td>
<td>A woman with glasses and a pink hat rides her bike.</td>
</tr>
<tr>
<td>A black dog is playing in water with a green toy.</td>
<td>A black dog is playing in water with a green toy.</td>
</tr>
<tr>
<td>The kid is sliding down a tan plastic slide.</td>
<td>The kid is sliding.</td>
</tr>
<tr>
<td>A woman sings</td>
<td>A woman sings</td>
</tr>
<tr>
<td>Three women enjoying a balloon joyride.</td>
<td>Three women on a balloon ride.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Examples of artifacts in the SNLI dataset that pro-
mote hypothesis sentences to be substrings of the evidence
sentences, especially in case of entailments. The bullet marked
sentences are the evidence sentences and the hypothesis sen-
tences below them.

This is a simple and inexpensive way of cre-
ating entailment pairs, with the benefit that this
annotation scheme probes for fundamental seman-
tic information from an annotator. Also, note that,
since Reisinger et al. collect annotations for the
twelve types of proto-role properties mentioned
above, the errors made by a neural RTE model
can be automatically subcategorized 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 suggest 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 ac-
curately modeled by statistical RTE models, the
lack of diversity is not an 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 im-
age 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 incen-
tive to spend the least amount of time possible in
constructing their responses. In Table 3, we list a
few such examples.

The FN+ dataset In Frame Semantics (Fillmore
et al., 2003), the primary unit of lexical analysis
is the frame, which captures the central proper-
ties of a concept, situation, or event. The largest
resource for frame annotated sentences with infor-
mation about evoked frames, their trigger phrases,
and frame arguments is the FrameNet dataset (Fill-
more 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
data set. 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 annota-
tors to “judge each paraphrase in terms of how
well it preserved the meaning of the original sen-
tence” (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 la-
beled entailed and rated 3.0 and above are usually
grammatically correct, the sentences with an aver-
age rating below 3 and labeled entailed sometimes
contain a grammatical errors, and some are rated
neutral or contradictory. Therefore, we remove
sentences with an average rating less than or equal
to 3.0 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 (4), which is a sentence
from FrameNet.

4) So our work must continue.

The word work triggers a frame and is replaced by
its paraphrase labor by Pavlick et al. to create (5).

5) So our labor must continue.

We consider the first sentence to be the text and the
second sentence to be the hypothesis. The annota-

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sentences</th>
<th>Label Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>FN+</td>
<td>154,605</td>
<td>43.45 56.55</td>
</tr>
<tr>
<td>SPR</td>
<td>154,607</td>
<td>34.80 65.20</td>
</tr>
<tr>
<td>DPR</td>
<td>3,661</td>
<td>49.99 50.01</td>
</tr>
<tr>
<td>Total</td>
<td>312,873</td>
<td>39.13 60.87</td>
</tr>
<tr>
<td>SNLI†</td>
<td>569,033</td>
<td>33.41 66.59</td>
</tr>
</tbody>
</table>

Table 4: Number of text–hypothesis pairs generated from
each dataset along with percentage of entailing v. non-
entailing sentences. SNLI included for comparison.


<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Grammaticality</th>
</tr>
</thead>
<tbody>
<tr>
<td>FN+</td>
<td>85</td>
<td>77</td>
</tr>
<tr>
<td>SPR</td>
<td>94</td>
<td>92</td>
</tr>
<tr>
<td>DPR</td>
<td>98</td>
<td>96</td>
</tr>
<tr>
<td>SNLI</td>
<td>91</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 5: 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 DPR dataset Definite Pronoun Resolution is the problem of identifying the correct antecedent for a definite pronoun—e.g. *him*, *her*, *it*, etc.—in one clause, given two potential antecedents in a preceding clause. Data generation for this task is done manually and relies on the concept of twin sentences. Twin sentences are (minimally) biclausal sentences that share a (linearly) initial clause containing at least two non-pronominal referring expressions but differ on a non-initial clause containing a pronoun that could corefer with either of the two referring expressions in the initial clause but which is biased to corefer with only one.

This concept is exemplified in (6), where the *bee* and the *flower* are the two referring expressions in the initial clause for both (6a) and (6b), and *it* is the pronoun.

(6) The bee landed on the flower because...
   a. ...it wanted pollen.
   b. ...it had pollen.

In (6a), *it* is biased to corefer with the *bee*, and in (6b), *it* is biased to corefer with the *flower*.

In order to assign the correct antecedent of *it* in both sentences, a computational system would presumably need world-knowledge about bees and flowers. 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). The ranking-based system that Rahman and Ng present obtains an accuracy of 73.1% on their dataset. This result—which, to our knowledge, remains the best posted on this dataset—outperforms a random baseline as well as various systems based on the Stanford resolver (Lee et al., 2011).

Each DPR coreference problem-solution pair can be converted to two annotated entailment problems by substituting the target pronoun with the two expressions that it could corefer with. Thus, two RTE pairs are generated for each DPR pair: one that is entailed and one that is not entailed.

For example, (6a) is rewritten to (7a) and (7b).

(7) a. The bee landed on the flower because the bee wanted pollen.
   b. The bee landed on the flower because the flower wanted pollen.

The two RTE pairs are then (7a)–(8a), which is paired with the output *entailed*, and (7a)–(8b), which is paired with the output *not-entailed*.

Statistics Table 4 summarizes the constructed datasets as well as number of sentences and the class category breakdown of the SNLI dataset.

4 Data Validation

Since our data is 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 5 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 6 and 7 show examples of ungrammatical hypothesis sentences and incorrectly labeled sentences on Mechanical Turk gave this pair of sentences an average rating of around 4, and so we consider this pair of sentences as an instance of the entailed relation.

4 Data Validation

Since our data is 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 5 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 6 and 7 show examples of ungrammatical hypothesis sentences and incorrectly labeled sentences.

3For the SNLI statistics, we map the two categories of contradiction and neutral to not-entailed.
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

We now conduct experiments to measure the variation in performance of neural RTE models trained using the datasets described above. The driving idea is that, by analyzing the variation in accuracy of a neural RTE model trained on different datasets on the same test set, such as the SNLI dataset, we can gain insights into the behavior of the model and potentially reveal interesting information about the SNLI dataset itself.

We first split our three datasets into train, validation, and test sets in the proportion of 80:10:10. Prior to training we convert the SNLI test set to a binary scheme by replacing both neutral and contradiction class labels with not-entailed.

For our model, we use the LSTM-based neural RTE model described by Bowman et al. (2015) which was their best performing individual neural model. This model first embeds the words using 300 dimensional word embeddings created using the Glove method (Pennington et al., 2014). Then, 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 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 8.

These results show 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 suggests that these three tasks are relatively 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 the model trained on all four datasets is equal to chance performance on the DPR dataset. Further, it is consistently lower than other cross-evaluation results shown on the off-diagonals of Table 8. This suggests that complex anaphora resolution is difficult for our model to capture, especially when its training data are not focused on demonstrating correct coreference resolution. And since the performance of the SNLI trained model is the least on the DPR dataset, this may suggest that the phenomenon of anaphora resolution occurs less often than paraphrasing or proto-role resolution in the SNLI dataset.

This is corroborated by a small manual analysis. 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, in this analysis, we found that the only text that might plausibly have been rewritten as a pronoun resolution problem (8a) was not; the hypothesis for (8a) is (8b).

(8) a. A man speaking to a woman in a grocery store as he selects a carton of juice.
   b. A man is complimenting a woman on her jacket.

Finally, we can see that the SNLI trained model achieves 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 sentences in FN+ contain language that is rarely seen. Thus, they contain more subtle
Differences compared to the differences between the text and the hypothesis in the SNLI dataset. This may also be 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 the change of State category caused the highest number of errors. Based on manual inspection of examples, such as sentence 4 in Table 1, we believe that this happens because such entailments are not easily captured using lexical patterns. On the other hand information about stationarity and change of possession may 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 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 analyzing 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 http://decomp.net.

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

Acknowledgments

This research was supported by the JHU HLT-COE, DARPA DEFT, DARPA LORELEI, and the JHU Science of Learning Institute. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.
References


Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa T. Dang, and Danilo Giampiccolo. 2010. The sixth PASCAL recognizing textual entailment challenge. In TAC.


Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In EMNLP. ACL.


