Most *babies* are *little* and most *problems* are *huge*: Compositional Entailment in Adjective-Nouns

Ellie Pavlick
University of Pennsylvania
epavlick@seas.upenn.edu

Chris Callison-Burch
University of Pennsylvania
ccb@cis.upenn.edu

Abstract

We examine adjective-noun (AN) composition in the task of recognizing textual entailment (RTE). We analyze behavior of ANs in large corpora and show that, despite conventional wisdom, adjectives do not always restrict the denotation of the nouns they modify. We use natural logic to characterize the variety of entailment relations that can result from AN composition. Predicting these relations depends on context and on common-sense knowledge, making AN composition especially challenging for current RTE systems. We demonstrate the inability of current state-of-the-art systems to handle AN composition in a simplified RTE task which involves the insertion of only a single word.

1 Overview

The ability to perform inference over utterances is a necessary component of natural language understanding (NLU). Determining whether one sentence reasonably implies another is a complex task, often requiring a combination of logical deduction and simple common-sense. NLU tasks are made more complicated by the fact that language is compositional: understanding the meaning of a sentence requires understanding not only the meanings of the individual words, but also understanding how those meanings combine.

Adjectival modification is one of the most basic types of composition in natural language. Most existing work in NLU makes a simplifying assumption that adjectives tend to be *restrictive*– i.e. adding an adjective modifier limits the set of things to which the noun phrase can refer. For example, the set of *little dogs* is a subset of the set of *dogs*, and we cannot in general say that *dog* entails *little dog*. This assumption has been exploited by high-performing RTE systems (MacCartney and Manning, 2008; Stern and Dagan, 2012), as well as used as the basis for learning new entailment rules (Baroni et al., 2012; Young et al., 2014).

However, this simplified view of adjectival modification often breaks down in practice. Consider the question of whether *laugh* entails *bitter laugh* in the following sentences:

1. Again his *laugh* echoed in the gorge.
2. Her *laugh* was rather derisive.

In (1), we have no reason to believe the man's laugh is bitter. In (2), however, it seems clear from context that we are dealing with an unpleasant person for whom *laugh* entails *bitter laugh*. Automatic NLU should be capable of similar reasoning, taking both context and common sense into account when making inferences.

This work aims to deepen our understanding of AN composition in relation to automated NLU. The contributions of this paper are as follows:

- We conduct an empirical analysis of ANs and their entailment properties.
- We define a task for directly evaluating a system's ability to predict compositional entailment of ANs in context.
- We benchmark several state-of-the-art RTE systems on this task.

2 Recognizing Textual Entailment

The task of recognizing textual entailment (RTE) (Dagan et al., 2006) is commonly used to evaluate the state-of-the-art of automatic NLU. The RTE task is: given two utterances, a *premise* (p) and a *hypothesis* (h), would a human reading p typically infer that h is

(1)	FraCas	p	No delegate finished the report on time.	Quantifiers
		h	Some Scandinavian delegate finished the report on time.	(*no* $\rightarrow \neg$*some*)
(2)	RTE2	p	Trade between China and India is expected to touch $20 bn this year...	Definitions
		h	There is a profitable trade between China and India.	(*$20 bn* $\rightarrow$ *profitable*)
(3)	NA	p	Some delegates finished the report on time.	Implicature
		h	Not all of the delegates finished the report on time.	(*some* $\rightarrow \neg$*all*)
(4)	SICK	p	A couple of white dogs are running along a beach.	Common Sense
		h	Two dogs are playing on the beach.	(*running* $\rightarrow$ *playing*)

Table 1: Examples of sentence pairs coming from various RTE datasets, and the types of inference highlighted by each. While linguistic phenomena like implicature (3) have yet to be explicitly included in RTE tasks, common-sense inferences like those in (4) (from the SICK dataset) have become a common part of NLU tasks like RTE, question answering, and image labeling.

most likely true? Systems are expected to produce either a binary (YES/NO) or trinary (ENTAILMENT/CONTRADICTION/UNKNOWN) output.

The type of knowledge tested in the RTE task has shifted in recent years. While older datasets mostly captured logical reasoning (Cooper et al., 1996) and lexical knowledge (Giampiccolo et al., 2007) (see Examples (1) and (2) in Table 1), the recent datasets have become increasingly reliant on common-sense knowledge of scenes and events (Marelli et al., 2014). In Example (4) in Table 1, for which the gold label is ENTAILMENT, it is perfectly reasonable to assume the dogs are playing. However, this is not necessarily true that *running* entails *playing*– maybe the dogs are being chased by a bear and are running for their lives! Example (4) is just one of many RTE problems which rely on intuition rather than strict logical inference.

Transformation-based RTE. There have been an enormous range of approaches to automatic RTE– from those based on theorem proving (Bjerva et al., 2014) to those based on vector space models of semantics (Bowman et al., 2015a). Transformation-based RTE systems attempt to solve the RTE problem by identifying a sequence of *atomic edits* (MacCartney, 2009) which can be applied, one by one, in order to transform p into h. Each edit can be associated with some entailment relation. Then, the entailment relation that holds between p and h overall is a function of the entailment relations associated with each atomic edit. This approach is appealing in that it breaks potentially complex p/h pairs into a series of bite-sized pieces. Transformation-based RTE is widely used, not only in rule-based approaches (MacCartney and Manning, 2008; Young et al., 2014), but also in statistical RTE systems (Stern and Dagan, 2012; Padó et al., 2014).

MacCartney (2009) defines an *atomic edit* applied to a linguistic expression as the deletion DEL, insertion INS, or substitution SUB of a subexpression. If x is a linguistic expression and e is an atomic edit, than $e(x)$ is the result of applying the edit e to the expression x. For example:

$$
\begin{aligned}
x &= a_1\ \mathit{girl}_2\ \mathit{in}_3\ a_4\ \mathit{red}_5\ \mathit{dress}_6 \\
e &= \mathrm{DEL}(\mathit{red}, 5) \\
e(x) &= a_1\ \mathit{girl}_2\ \mathit{in}_3\ a_4\ \mathit{dress}_5
\end{aligned}
$$

We say that the entailment relation that holds between x and $e(x)$ is *generated* by the edit e. In the above example, we would say that e generates a *forward entailment* ($\sqsubset$) since *a girl in a red dress* entails *a girl in a dress*.

3 Natural Logic Entailment Relations

Natural logic (MacCartney, 2009) is a formalism that describes entailment relationships between natural language strings, rather than operating over mathematical formulae. Natural logic enables both light-weight representation and robust inference, and is an increasingly popular choice for NLU tasks (Angeli and Manning, 2014; Bowman et al., 2015b; Pavlick et al., 2015). There are seven "basic entailment relations" described by natural logic, five of which we explore here.[1]

These five relations, as they might hold between an AN and the head N, are summarized in Figure 1. The *forward entailment* relation is the restrictive case, in which the AN (*brown dog*) is a subset of (and thus entails) the N (*dog*) but the N

[1]We omit two relationships: *negation* and *cover*. These relations require that the sets denoted by the strings being compared are "exhaustive." In this work, this requirement would be met when everything in the universe is either an instance of the noun or it is an instance of the adjective-noun (or possibly both). This is a hard constraint to meet, and we believe that the interesting relations that result from AN composition are adequately captured by the remaining 5 relations.

does not entail the AN (*dog* does not entail *brown dog*). The symmetric *reverse entailment* can also occur, in which the N is a subset of the set denoted by the AN. An example of this is the AN *possible solution*: i.e. all actual *solutions* are *possible solutions*, but there are an abundance of *possible solutions* that are not and will never be actual *solutions*. In the *equivalence* relation, AN and N denote the same set (e.g. the *entire universe* is the same as the *universe*), whereas in the *alternation* relation, AN and N denote disjoint sets (e.g. a *former senator* is not a *senator*). In the *independence* relation, the AN has no determinable entailment relationship to the N (e.g. an *alleged criminal* may or may not be a *criminal*).

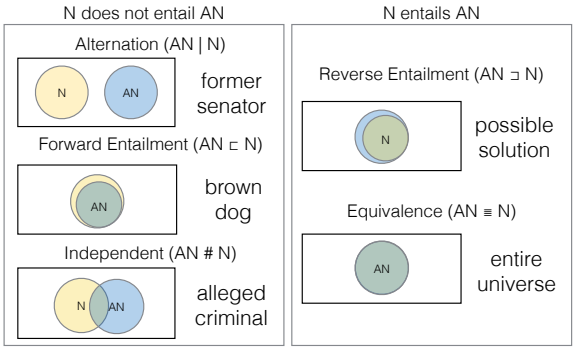

Figure 1: Different entailment relations that can exist between an adjective-noun and the head noun. The best-known case is that of *forward entailment*, in which the AN denotes a subset of the N (e.g. *brown dog*). However, many other relationships may exist, as modeled by natural logic.

4 Simplified RTE Task

The focus of this work is to determine the entailment relation that exists between an AN and its head N in a given context. To do this, we define a simplified entailment task identical to the normal RTE task, with the constraint that p and h differ only by one atomic edit e as defined in Section 2. We look only at insertion INS(A) and deletion DEL(A), where A must be a single adjective.

We use a 3-way entailment classification where the possible labels are ENTAILMENT, CONTRADICTION, and UNKNOWN. This allows us to recover the basic entailment relation from Section 3: by determining the labels associated with the INS operation and the DEL operation, we can uniquely identify each of the five relations (Table 2).

	INS	DEL
Equivalence	ENTAILMENT	ENTAILMENT
Forward Entail.	ENTAILMENT	UNKNOWN
Reverse Entail.	UNKNOWN	ENTAILMENT
Independence	UNKNOWN	UNKNOWN
Alternation	CONTRADICTION	CONTRADICTION

Table 2: Entailment generated by INS(A) or DEL(A) for possible relations holding between AN and N. Both INS and DEL are required to distinguish all five entailment relations.

4.1 Limitations

Modeling denotations of ANs and N. We note that this task design does not directly ask about the relationship between the sets denoted by the AN and by the N (as shown in Figure 1). Rather than asking "Is this instance of AN an instance of N?" we ask "Is this statement that is true of AN also true of N?" While these are not the same question, they are often conflated in NLP, for example, in information extraction, when we use statements about ANs as justification for extracting facts about the head N (Angeli et al., 2015). We focus on the latter question and accept that this prevents us from drawing conclusions about the actual set theoretic relation between the denotation of AN and the denotation of N. However, we are able to draw conclusions about the practical entailment relation between statements about the AN and statements about the N.

Monotonicity. In this simplified RTE task, we assume that the entailment relation that holds overall between p and h is attributable wholly to the atomic edit (i.e. the inserted or deleted adjective). This is an over-simplification. In practice, several factors can cause the entailment relation that holds between the sentences overall to differ from the relation that holds between the AN and the N. For example, quantifiers and other downward-monotone operators can block or reverse entailments (*brown dog* $\rightarrow$ *dog*, but *no brown dog* $\not\rightarrow$ *no dog*). While we make some effort to avoid selecting such sentences for our analysis (Section 5.3), fully identifying and handling such cases is beyond the scope of this paper. We acknowledge that monotone operators and other complicating factors (e.g. multiword expressions) might be present in our data, but we believe, based on manual inspection, that they not frequent enough to substantially effect our analyses.

5 Experimental Design

To build an intuition about the behavior of ANs in practice, we collect human judgments of the entailments generated by inserting and deleting adjectives from sentences drawn from large corpora. In this section, we motivate our design decisions, before carrying out our full analysis in Section 6.

5.1 Human judgments of entailment

People often draw conclusions based on "assumptions that seem plausible, rather than assumptions that are known to be true" (Kadmon, 2001). We therefore collect annotations on a 5-point scale, ranging from 1 (definite contradiction) to 5 (definite entailment), with 2 and 4 capturing likely (but not certain) contradiction/entailment respectively. We recruit annotators on Amazon Mechanical Turk. We tell each annotator to assume that the premise "is true, or describes a real scenario" and then, using their best judgement, to indicate how likely it is, on a scale of 1 to 5, that the hypothesis "is also true, or describes the same scenario." They are given short descriptions and several examples of sentence pairs that constitute each score along the 1 to 5 scale. They are also given the option to say that "the sentence does not make sense," to account for poorly constructed p/h pairs, or errors in our parsing. We use the mean score of the three annotators as the true score for each sentence pair.

Inter-annotator agreement. To ensure that our judgements are reproducible, we re-annotate a random 10% of our pairs, using the same annotation setup but a different set of annotators. We compute the intra-class correlation (ICC) between the scores received on the first round of annotation, and those received in the second pass. ICC is related to Pearson correlation, and is used to measure consistency among annotations when the group of annotators measuring each observation is not fixed, as opposed to metrics like Fleiss's κ which assume a fixed set of annotators. On our data, the ICC is 0.77 (95% CI 0.73 - 0.81) indicating very high agreement. These twice-annotated pairs will become our test set in Section 7.

5.2 Data

Selecting contexts. We first investigate whether, in naturally occurring data, there is a difference between contexts in which the author uses the AN and contexts in which the author uses only the (unmodified) N. In other words, in order to study the effect of an A (e.g. *financial*) on the denotation of an N (e.g. *system*), is it better to look at contexts like (a) below, in which the author originally used the AN *financial system*, or to use contexts like (b), in which the author used only the N *system*?

(a) The TED spread is an indication of investor confidence in the U.S. *financial system*.

(b) Wellers hopes the *system* will be fully operational by 2015.

We will refer to contexts like (a) as *natural* contexts, and those like (b) as *artificial*. We take sample of 500 ANs from the Annotated Gigaword corpus (Napoles et al., 2012), and choose three natural and three artificial contexts for each. We generate p/h pairs by deleting/inserting the A for the natural/artificial contexts, respectively, and collect human judgements on the effect of the INS(A) operation for both cases.

Figure 2 displays the results of this pilot study. In sentences which contain the AN naturally, there is a clear bias toward judgements of "entailment." That is, in contexts when an AN appears, it is often the case that this A is superfluous: the information carried by the A is sufficiently entailed by the context that removing it does not remove information. Sentences (a) and (b) above provide intuition: in the case of sentence (a), trigger phrases like *investor confidence* make it clear that the *system* we are discussing is the *financial system*, whether or not the adjective *financial* actually appears. No such triggers exist in sentence (b).

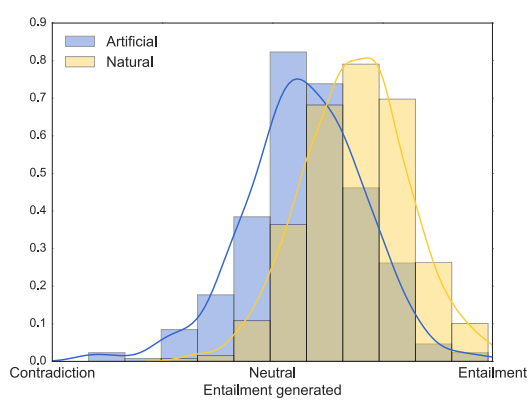

Figure 2: p/h pairs derived from natural contexts result in a notable bias toward judgements of "entailment" for the INS(A) operation, compared to p/h pairs derived from artificial contexts.

Selecting ANs. We next investigate whether the frequency with which an AN is used effects its tendency to entail/be entailed by the head N. Again, we run a small pilot study. We choose 500 ANs stratified across different levels of frequency of occurrence in order to determine if sampling the most frequent ANs introduces bias into our annotation. We see no significant relationship between the frequency with which an AN appears and the entailment judgements we received.

5.3 Final design decisions

As a result of the above pilot experiments, we proceed with our study as follows. First, we use only artificial contexts, as we believe this will result in a greater variety of entailment relations and will avoid systematically biasing our judgements toward entailments. Second, we use the most frequent AN pairs, as these will better represent the types of ANs that NLU systems are likely to encounter in practice.

We look at four different corpora capturing four different genres: Annotated Gigaword (Napoles et al., 2012) (**News**), image captions (Young et al., 2014) (**Image Captions**), the Internet Argument Corpus (Walker et al., 2012) (**Forums**), and the prose fiction subset of GutenTag dataset (Brooke et al., 2015) (**Literature**). From each corpus, we select the 100 nouns which occur with the largest number of unique adjectives. Then, for each noun, we take the 10 adjectives with which the noun occurs most often. For each AN, we choose 3 contexts[2] in which the N appears unmodified, and generate p/h pairs by inserting the A into each.

We collect 3 judgements for each p/h pair. Since this task is subjective, and we want to focus our analysis on clean instances on which human agreement is high, we remove pairs for which one or more of the annotators chose the "does not make sense" option and pairs for which we do not have at least 2 out of 3 agreement (i.e. at least two workers must have chosen the same score on the 5-point scale). In the end, we have a total of 5,560 annotated p/h pairs[3] coming roughly evenly from our 4 genres.

[2]As a heuristic, we skip sentences containing obvious downward-monotone operators, e.g. *not*, *every* (Section 4).

[3]Our data is available at `http://www.seas.upenn.edu/~nlp/resources/AN-composition.tgz`

6 Empirical Analysis

Figure 3 shows how the entailment relations are distributed in each genre. In Image Captions, the vast majority of ANs are in a *forward entailment* (restrictive) relation with their head N. In the other genres, however, a substantial fraction (36% for Forums) are in equivalence relations: i.e. the AN denotes the same set as is denoted by the N alone.

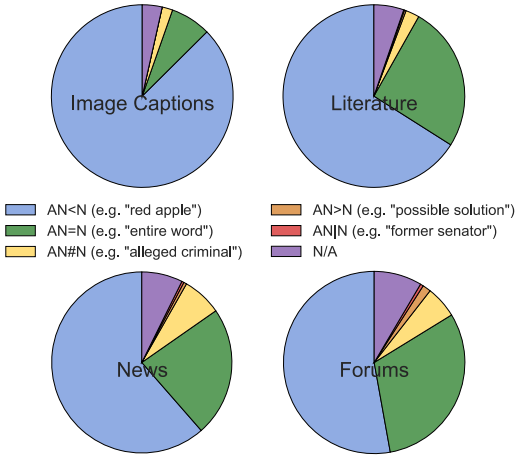

Figure 3: Basic entailment relations assigned to ANs according to the 5,560 p/h pairs our data.

When does N entail AN? If it is possible to insert adjectives into a sentence without adding new information, when does this happen? When is adjectival modification not restrictive? Based on our qualitative analysis, two clear patterns stand out:

1) **When the adjective is prototypical of the noun it modifies.** In general, we see that adding adjectives which are seen as attributes of the "prototypical" instance of the noun tend to generate entailments. E.g. people are generally comfortable concluding that *beach*→*sandy beach*. The same adjective may be prototypical and thus entailed in the context of one noun, but generate a contradiction in the context of another. E.g. if someone has a *baby*, it is probably fine to say they have a *little baby*, but if someone has *control*, it would be a lie to say they have *little control* (Figure 4).[4]

2) **When the adjective invokes a sense of salience or importance.** Nouns are assumed to be salient and relevant. E.g. *answers* are assumed (perhaps naively) to be *correct*, and *problems* are

[4]These curves show the distribution over entailment scores associated with the INS(A) operation. Yellow curves show, for a single N, the distribution over all the As that modify it. Blue curves show, for a single A, the distribution over all the Ns it modifies.

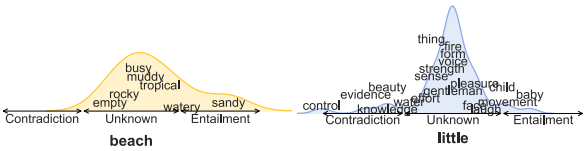

Figure 4: Inserting adjectives that are seen as "prototypical" of the noun tends to generate entailments. E.g., *beach* generally entails *sandy beach*.

assumed (perhaps melodramatically) to be *current* and *huge*. Inserting adjectives like *false* or *empty* tend to generate contradictions (Figure 5).

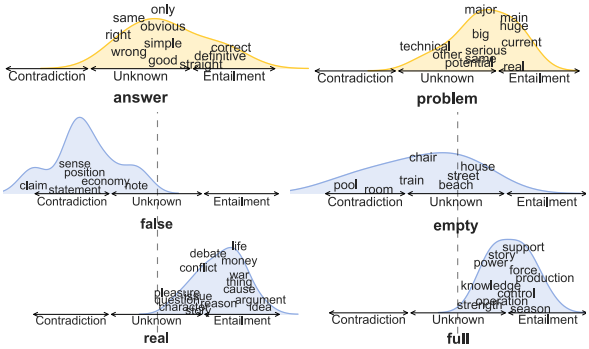

Figure 5: Unless otherwise specified, nouns are considered to be salient and relevant. *Answers* are assumed to be *correct*, and *problems* to be *current*.

What do the different natural logic relations look like in practice? Table 3 shows examples of ANs and contexts exhibiting each of the basic entailment relations. Some entailment inferences depend entirely on contextual information (Example 2a) while others arise from common-sense inference (Example 2b). Many of the most interesting examples fall into the *independence* relation. Recall from Section 3 that independence, in theory, covers ANs such as *alleged criminal*, in which the AN may or may not entail the N. In practice, the cases we observe falling into the independence relation tend to be those which are especially effected by world knowledge. In Example 3, *local economy* is considered to be independent of *economy* when used in the context of *President Obama*: i.e. the assumption that the president would be discussing the national economy is so strong that even when the president says *the local economy is improving*, people do not take this to mean that he has said *the economy is improving*.

Undefined entailment relations. Our annotation methodology– i.e. inferring entailment relations based on the entailments generated by INS and DEL edits– does not enforce that all of the ANs fit into one of the five entailment relations defined by natural logic. Specifically, we observe many instances ($\sim$5% of p/h pairs) in which INS is determined to generate a contradiction, while DEL is said to generate an entailment. In terms of set theory, this is equivalent to the (non-sensical) setting in which "every AN is an instance of N, but no N is an instance of AN." On inspection, these again represent cases in which common-sense assumptions dominate the inference. In Example 6, when given the premise *Bush travels to Michigan to discuss the economy*, annotators are confident enough that *economy* does not entail *Japanese economy* (why on earth would Bush try to discuss the Japanese economy!?) that they label the insertion of *Japanese* as generating a contradiction. However, when presented with the p/h in the opposite direction, annotators agree that the *Japanese economy* does indeed entail the *economy*. These examples highlight the flexibility with which humans perform natural language inference, and the need for automated systems to be equally flexible.

Take aways. Our analysis in this section results in three key conclusions about AN composition. 1) Despite common assumptions, adjectives do not always restrict the denotation of a noun. Rather, adjectival modification can result in a range of entailment relations, including equivalence and contradiction. 2) There are patterns to when the insertion of an adjective is or is not entailment-preserving, but recognizing these patterns requires common-sense and a notion of "prototypical" instances of nouns. 3) The entailment relation that holds between an AN and the head N is highly context dependent. These observations describe sizable obstacles for automatic NLU systems. Common-sense reasoning is still a major challenge for computers, both in terms of how to learn world knowledge and in how to represent it. In addition, context-sensitivity means that entailment properties of ANs cannot be simply stored in a lexicon and looked-up at run time. Such properties make AN composition an important problem on which to focus NLU research.

7 Benchmarking Current SOTA

We have highlighted why AN composition is an interesting and likely challenging phenomenon for

(1)	AN $\sqsubset$ N	He underwent a [successful] operation on his leg at a Lisbon hospital in December.
(2a)	AN $\equiv$ N	The [deadly] attack killed at least 12 civilians.
(2b)	AN $\equiv$ N	The [entire] bill is now subject to approval by the parliament.
(3)	AN # N	President Obama cited the data as evidence that the [local] economy is improving.
(4)	AN $\sqsupset$ N	The [militant] movement was crushed by the People's Liberation Army.
(5)	AN | N	Red numbers spelled out their [perfect] record: 9-2.
(6)	AN ? N	Bush travels Monday to Michigan to make remarks on the [Japanese] economy.

Table 3: Examples of ANs in context exhibiting each of the different entailment relations. Note that these are "artificial" contexts (Section 5.2), meaning the adjective was not originally a part of the sentence.

automated NLU systems. We now turn our investigation to the performance of state-of-the-art RTE systems, in order to quantify how well AN composition is currently handled.

The Add-One Entailment Task. We define the "Add-One Entailment" task to be identical to the normal RTE task, except with the constraint that the premise p and the hypothesis h differ only by the atomic insertion of an adjective: $h = e(p)$ where e=INS(A) and A is a single adjective. To provide a consistent interface with a range of RTE systems, we use a binary label set: NON-ENTAILMENT (which encompasses both CONTRADICTION and UNKNOWN) and ENTAILMENT. We want to test on only straightforward examples, so as not to punish systems for failing to classify examples which humans themselves find difficult to judge. In our test set, therefore, we label pairs with mean human scores ≤ 3 as NON-ENTAILMENT, pairs with scores ≥ 4 as ENTAILMENT, and throw away the pairs which fall into the ambiguous range in between.[5] Our resulting train, dev, and test sets contain 4,481, 510, and 387 pairs, respectively. These splits cover disjoint sets of ANs– i.e. none of the ANs appearing in test were seen in train. Individual adjectives and/or nouns can appear in both train and test. The dataset consists of roughly 85% NON-ENTAILMENT and 15% ENTAILMENT. Inter-annotator agreement achieves 93% accuracy.

7.1 RTE Systems

We test a variety of state-of-the-art RTE systems, covering several popular approaches to RTE. These systems are described in more detail below.

[5]For our training and dev sets, we include all pairs, considering scores < 3.5 as NON-ENTAILMENT and scores ≥ 3.5 as ENTAILMENT. We tried removing "ambiguous" pairs from the training and dev sets as well, but it did not improve the systems' performances on the test set.

Classifier-based. The Excitement Open RTE platform (Magnini et al., 2014) includes a suite of RTE systems, including baseline systems as well as feature-rich supervised systems which provide state-of-the-art performance on the RTE3 datasets (Giampiccolo et al., 2007). We test two systems from Excitement: the simple Maximum Entropy (**MaxEnt**) model which uses a suite of dense, similarity-based features (e.g. word overlap, cosine similarity), and the more sophisticated Maximum Entropy model (**MaxEnt+LR**) which uses the same similarity-based features but additionally incorporates features from external lexical resources such as WordNet (Miller, 1995) and VerbOcean (Chklovski and Pantel, 2004). We also train a standard unigram model (**BOW**).

Transformation-based. The Excitement platform also includes a transformation-based RTE system called **BIUTEE** (Stern and Dagan, 2012). The BIUTEE system derives a sequence of edits that can be used to transform the premise into the hypothesis. These edits are represented using feature vectors, and the system searches over edit sequences for the lowest cost "proof" of either entailment or non-entailment. The feature weights are set by logistic regression during training.

Deep learning. Bowman et al. (2015a) recently reported very promising results using deep learning architectures and large training data for the RTE task. We test the performance of those same implementations on our Add-One task. Specifically, we test the following models: a basic Sum-of-words model (**Sum**), which represents both p and h as the sum of their word embeddings, an **RNN** model, and an **LSTM** model. We also train a bag-of-vectors model (**BOV**), which is simply a logistic regression whose features are the concatenated averaged word embeddings of p and h.

For the LSTM, in addition to the normal train-

ing setting– i.e. training only on the 5K Add-One training pairs– we test a transfer-learning setting (**Transfer**). In transfer learning, the model trains first on a large general dataset before fine-tuning its parameters on the smaller set of target-domain training data. For our Transfer model, we train first on the 500K pair SNLI dataset (Bowman et al., 2015a) until convergence, and then fine-tune on the 5K Add-One pairs. This setup enabled Bowman et al. (2015a) to train a high-performance LSTM for the SICK dataset, which is of similar size to our Add-One dataset (∼5K training pairs).

7.2 Results

Out of the box performances. To calibrate expectations, we first report the performance of each of the systems on the datasets for which they were originally designed. For the Excitement systems, this is the RTE3 dataset (Table 6a). For the deep learning systems, this is the SNLI dataset (Table 6b). For the deep learning systems, in addition to reporting performance when trained on the SNLI corpus (500K p/h pairs), we report the performance in a reduced training setting in which systems only have access to 5K p/h pairs. This is equivalent to the amount of data we have available for the Add-One task, and is intended to give a sense of the performance improvements we should expect from these systems given the size of the training data.

	RTE3
Majority	51.3
BOW	51.0
Edit Dist.	61.9
MaxEnt+LR	63.6
BIUTEE	65.6

(a) Systems from Magnini et al. (2014) on RTE3.

	SNLI 500K / 5K
Majority	65.7
BOV	74.4 / 71.5
RNN	82.1 / 67.0
Sum	85.3 / 69.2
LSTM	86.2 / 68.0

(b) Systems from Bowman et al. (2015a) on SNLI.

Figure 6: Performance of SOTA systems on the datasets for which they were originally developed.

7.3 Performance on Add-One RTE.

Finally, we train each of the systems on the 5,000 Add-One p/h pairs in our dataset and test on our held-out set of 387 pairs. Figure 7 reports the results in terms of accuracy and precision/recall for the ENTAILMENT class. The baseline strategy of predicting the majority class for each adjective, based on the training data, reaches close to human performance (92% accuracy). Given the simplicity of the task (p and h differ by a single word), this baseline strategy should be achievable. However, none of the systems tested come close to this level of performance, suggesting that they fail to learn even the most-likely entailment generated by adjectives (e.g. that INS(*brown*) probably generates NON-ENTAILMENT and INS(*possible*) probably generates ENTAILMENT). The best performing system is the RNN, which achieves 87% accuracy, only two points above the baseline of always guessing NON-ENTAILMENT.

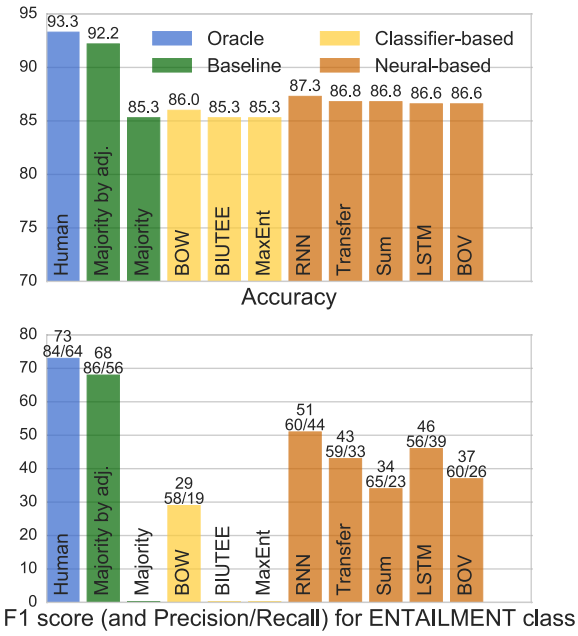

Figure 7: Performances of all systems on AddOne RTE task. The strategy of predicting the majority class for each adjective– based on the training data– reaches near human performance. None of the systems tested come close to human levels, indicating that the systems fail even to memorize the most-likely class for each adjective in training.

8 Related Work

Past work, both in linguistics and in NLP, has explored different classes of adjectives (e.g. privative, intensional) as they relate to entailment (Kamp and Partee, 1995; Partee, 2007; Boleda et al., 2013; Nayak et al., 2014). In general, prior studies have focused on modeling properties of the adjectives alone, ignoring the context-dependent nature of AN/N entailments– i.e. in prior work *little* is always restrictive, whether it is modifying *baby* or *control*. Pustejovsky (2013) offer a preliminary analysis of the contextual complexities surrounding adjective inference, which rein-

forces many of the observations we have made here. Hartung and Frank (2011) analyze adjectives in terms of the properties they modify but don't address them from an entailment perspective. Tien Nguyen et al. (2014) look at the adjectives in the restricted domain of computer vision.

Other past work has employed first-order logic and other formal representations of adjectives in order to provide compositional entailment predictions (Amoia and Gardent, 2006; Amoia and Gardent, 2007; McCrae et al., 2014). Although theoretically appealing, such rigid logics are unlikely to provide the flexibility needed to handle the type of common-sense inferences we have discussed here. Distributional representations provide much greater flexibility in terms of representation (Baroni and Zamparelli, 2010; Guevara, 2010; Boleda et al., 2013). However, work on distributional AN composition has so far remained out-of-context, and has mostly been evaluated in terms of overall "similarity" rather than directly addressing the entailment properties associated with composition.

9 Conclusion

We have investigated the problem of adjective-noun composition, specifically in relation to the task of RTE. AN composition is capable of producing a range of natural logic entailment relationship, at odds with commonly-used heuristics which treat all adjectives a restrictive. We have shown that predicting these entailment relations is dependent on context and on world knowledge, making it a difficult problem for current NLU technologies. When tested, state-of-the-art RTE systems fail to learn to differentiate entailment-preserving insertions of adjectives from non-entailing ones. This is an important distinction for carrying out human-like reasoning, and our results reveal important weaknesses in the representations and algorithms employed by current NLU systems. The Add-One Entailment task we have introduced will allow ongoing RTE research to better diagnose systems' abilities to capture these subtleties of ANs, which that have practical effects on natural language inference.

Acknowledgments

This research was supported by a Facebook Fellowship, and by gifts from the Alfred P. Sloan Foundation, Google, and Facebook. This material is based in part on research sponsored by the NSF grant under IIS-1249516 and DARPA under number FA8750-13-2-0017 (the DEFT program). 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 and the U.S. Government.

We would like to thank Sam Bowman for helping us to replicate his prior work and ensure a fair comparison. We would also like to thank the anonymous reviewers for thoughtful comments, and the Amazon Mechanical Turk annotators for their contributions.

References

Marilisa Amoia and Claire Gardent. 2006. Adjective based inference. In *Proceedings of the Workshop KRAQ'06 on Knowledge and Reasoning for Language Processing*, pages 20–27. Association for Computational Linguistics.

Marilisa Amoia and Claire Gardent. 2007. A first order semantic approach to adjectival inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 185–192, Prague, June. Association for Computational Linguistics.

Gabor Angeli and Christopher D. Manning. 2014. NaturalLI: Natural logic inference for common sense reasoning. In *Empirical Methods in Natural Language Processing (EMNLP)*, October.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China, July. Association for Computational Linguistics.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA, October. Association for Computational Linguistics.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32, Avignon, France, April. Association for Computational Linguistics.

Johannes Bjerva, Johan Bos, Rob van der Goot, and Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. *SemEval 2014*, page 642.

Gemma Boleda, Marco Baroni, Louise McNally, and Nghia Pham. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS*, pages 35–46.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015b. Learning distributed word representations for natural logic reasoning. In *2015 AAAI Spring Symposium Series*.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47, Denver, Colorado, USA, June. Association for Computational Linguistics.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *EMNLP*, volume 2004, pages 33–40.

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. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognizing textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. Springer.

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, June. Association for Computational Linguistics.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37, Uppsala, Sweden, July. Association for Computational Linguistics.

Matthias Hartung and Anette Frank. 2011. Exploring supervised LDA models for assigning attributes to adjective-noun phrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 540–551, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Nirit Kadmon. 2001. *Formal Pragmatics: Semantics, Pragmatics, Presupposition, and Focus. Willey*. Blackwell. Oxford.

Hans Kamp and Barbara Partee. 1995. Prototype theory and compositionality. *Cognition*, 57(2):129–191.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 521–528.

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

Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Guenter Neumann, Tae-Gil Noh, Sebastian Padó, Asher Stern, and Omer Levy. 2014. The Excitement Open Platform for textual inferences. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 43–48, Baltimore, Maryland, June. Association for Computational Linguistics.

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 *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland, May. ACL Anthology Identifier: L14-1314.

John P. McCrae, Francesca Quattri, Christina Unger, and Philipp Cimiano. 2014. Modelling the semantics of adjectives in the ontology-lexicon interface. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 198–209, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, November.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100.

Neha Nayak, Mark Kowarsky, Gabor Angeli, and Christopher D. Manning. 2014. A dictionary of nonsubsective adjectives. Technical Report CSTR 2014-04, Department of Computer Science, Stanford University, October.

Sebastian Padó, Tae-Gil Noh, Asher Stern, Rui Wang, and Roberto Zanoli. 2014. Design and realization of a modular architecture for textual entailment. *Journal of Natural Language Engineering*.

Barbara Partee. 2007. Compositionality and coercion in semantics: The dynamics of adjective meaning. *Cognitive foundations of interpretation*, pages 145–161.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding semantics to data-driven paraphrasing. In *Association for Computational Linguistics*, Beijing, China, July. Association for Computational Linguistics.

James Pustejovsky. 2013. Inference patterns with intensional adjectives. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 85–89, Potsdam, Germany, March. Association for Computational Linguistics.

Asher Stern and Ido Dagan. 2012. BIUTEE: A modular open-source system for recognizing textual entailment. In *Proceedings of the ACL 2012 System Demonstrations*, pages 73–78, Jeju Island, Korea, July. Association for Computational Linguistics.

Dat Tien Nguyen, Angeliki Lazaridou, and Raffaella Bernardi. 2014. Coloring objects: Adjective-noun visual semantic compositionality. In *Proceedings of the Third Workshop on Vision and Language*, pages 112–114, Dublin, Ireland, August. Dublin City University and the Association for Computational Linguistics.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

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 (TACL)*, 2(Feb):67–78.