

# Entailment due to Syntactically Encoded Semantic Relationships

**Elena Akhmatova**

Centre for Language Technology  
Macquarie University  
Sydney, Australia  
elena@ics.mq.edu.au

**Mark Dras**

Centre for Language Technology  
Macquarie University  
Sydney, Australia  
madras@ics.mq.edu.au

## Abstract

The majority of the state-of-the-art approaches to recognizing textual entailment focus on defining a generic approach to RTE. A generic approach never works well for every single entailment pair: there are entailment pairs that are recognized poorly by all the generic systems. Automatic identification of such entailment pairs and applying to them an RTE algorithm that is specific to them could thus increase an overall performance of an entailment engine (that in this case will combine a generic RTE algorithm with a number of RTE algorithms for the problematic entailment pairs). We identify one subtype of entailment pairs and develop a two-part probabilistic model for their classification into true and false entailments and evaluate it relative both to a baseline and to the RTE systems. We show that the model performs better than the baseline and the average of the systems from the RTE2 on both the balanced and unbalanced datasets we have created for evaluation.

## 1 Introduction

Recognizing Textual Entailment (RTE) is a task where, given two text snippets, the goal is to determine whether the meaning of one text snippet can be inferred from the meaning of the other (Dagan et al., 2005). The first of the text snippets in such a pair is referred to as *the text* and the other one as *the hypothesis*. The pair of text and hypothesis is called a *text-hypothesis pair* or *entailment pair*, with the two names considered to be synonymous. The text is usually much longer than the hypothesis. It can be represented by one or more coherent sentences, while the hypothesis is usually one short sen-

tence. It is the meaning of the hypothesis that might or might not be entailed from the text. Thus, given a text-hypothesis pair, we recognize the relation between the meanings of the text and the hypothesis in the pair as *a true entailment* if the meaning of the hypothesis is entailed from the meaning of the text. Otherwise, we recognize the relation between the meanings of the texts as *a false entailment*.

There are several datasets for RTE. They contain text-hypothesis pairs marked *yes* if there is a relation of true entailment in a pair and *no* otherwise. These datasets are manually created annually for the RTE Challenges<sup>1</sup> and are freely available.

Most state-of-the-art approaches to RTE seek a generic approach to the task and do not differentiate between text-hypothesis pairs. However, a possible alternative is to consider subclasses of entailment pairs and build models to handle these specialties. An instance of this idea is proposed in Vanderwende and Dolan (2005), where the complete set of entailment pairs is divided in two: those whose categorization could be accurately predicted based solely on syntactic cues and those where it is not the case. Their subsequent work (Vanderwende et al., 2006) presents an RTE system based on this work.

The broader context of our work is to investigate different ways of subclassifying entailment pairs. In this framework, a generic system would have additional special components that take care of the special subclasses of entailment pairs. Such a component is involved when a pair of its subclass is recognized. Note that we do not envisage classifying all the entailment pairs to give a partitioning of the space, a probably infeasible task. We suggest dividing into classes the entailment pairs that are problematic for all the state-of-the-art generic systems and develop separate RTE algorithms for these par-

<sup>1</sup><http://www.pascal-network.org/Challenges/RTE/>

ticular classes. The broad question that we aim to answer is whether this will improve the overall performance of the RTE engine.

In this paper we are looking at one subtype of entailment pairs where a semantic relation expressed in the hypothesis is implicitly represented by a syntactic construction in the text. There are several reasons to work with this type of entailment pairs. First, it proves possible to recognize them well automatically and distinguish them from other entailment pairs using machine learning. Second, narrowing down the entailment pairs to this subset allows us to draw an analogy with, and develop an algorithm related to, the work by Lapata (2001) that finds the implicit relation between attributes to a head noun in the noun group. That together with a conditional probability model in a parallel with SMT will be taken as the basis of an algorithm for classification of entailment pairs of the chosen type. We evaluate the approach on the RTE2 annotated dataset.

The layout of the paper follows the general flow of the research. Section 2 defines the chosen type of entailment pairs. Section 3 describes an automatic classifier which distinguishes the desired type of the entailment pairs. Section 4 describes an algorithm for recognizing true and false entailments for the entailments of the chosen type, and gives some experimental results comparing our algorithm against a number of baselines. Section 5 presents the evaluation results and section 6 concludes the work.

## 2 Entailment types

We looked through the RTE2 test set and partitioned the set into several groups of entailments. Though the entailment pairs are different, for every word in the hypothesis there is often a word in the text from which it is entailed. It is not always so and we focus on the entailment pairs where this is not the case.

### 2.1 Syntactically encoded semantics

The entailment relationship we are focusing on is named *an Entailment due to Syntactically Encoded Semantic Relationships (ESES)*, as a specific syntactic construction in the text encodes a semantic relationship between its elements that is explicitly shown in the hypothesis.

Being more precise, the text-hypothesis pairs of

interest have the following characteristics:

1. The hypothesis is a simple sentence. That is a sentence that consists of a subject, a predicate, and an object, and has no subordinate clauses.
2. Both subject and object of the hypothesis (or their morphological variants) are found in the text.
3. The predicate of the hypothesis has no match with anything in the text that is linked to the matches of the subject and the object of the hypothesis.
4. The matches of the subject and the object in the hypothesis can be linked to each other in the text by any syntactic relationship except depending from the same verb or a derivative of it.

Thus, the predicate of the hypothesis is the semantic relationship between its subject and object that is not explicitly defined in the text but is implicitly presented in the syntactic relationship between the matches of the subject and object of the hypothesis in the text.

The most frequent syntactic relationships between the matches of the subject and the object of the hypothesis in the text in the RTE2 dataset are apposition,<sup>2</sup> a noun group and its prepositional attachment, and attributive relation within a noun group.

Consider the examples of the entailments of the described type:

- (1) *Text*: From Les Combes, in the Italian Alps, yesterday, where the Pope is on vacation, the Vatican's Press Office Director, Joaquin Navarro Valls, responded with a written statement to the accusations made by the Israeli government against Benedict XVI.

*Hypothesis*: Les Combes is located in the Italian Alps.

The location *Les Combes* is in the relation of apposition to *the Italian Alps*. This syntactic relation implicitly encodes the semantic relation represented by the words *is located in* between the noun groups.

<sup>2</sup>We use the definition of Quirk et al. (1985) here.

- (2) *Text*: Lt. Jim Bowell of the Butler Township Fire Department said the 4:45 a.m. accident set fire to about 100 yards of woods.

*Hypothesis*: Jim Bowell is engaged by the Butler Township Fire Department.

*Lt. Jim Bowell* is connected syntactically to the *Butler Township Fire Department* via a preposition. That implicitly encodes a relation between the person and organization, *to be engaged by*.

- (3) *Text*: Japan's Kyodo news agency said the US could be ready to set up a liaison office—the lowest level of diplomatic representation—in Pyongyang if it abandons its nuclear program.

*Hypothesis*: Kyodo news agency is based in Japan.

The attributive relationship between *Kyodo news agency* and *Japan* suggests but does not state explicitly the relationship *is based in* between them. The *Kyodo news agency is based in Japan* is entailed from the attributive relationships between the nouns.

## 2.2 Recognition of the entailment types by RTE2 Challenge participants

The fact that most entailment engines rely on high word overlap, longest common substring and other features<sup>3</sup> implies an assumption that there must be a word in the text for every word in the hypothesis. That in its turn suggests the ESESR entailment pairs may not be recognized well.

The RTE2 results confirm that. The mean recognition of the entailments of this subtype is 61.9% among all the 41 system submissions. This places the type we have defined around the middle: difficult enough to be a challenge, but not so difficult as to be infeasible. The agreement on the recognition of the true entailments is around 86%, and the false entailments are recognized correctly with an accuracy of less than 25%. The features mentioned above tend to guess the true entailment as the matches of the subject and the object of the hypothesis in the text give a good score for word overlap, longest common substring and dependency tree matches. The

<sup>3</sup>See, for example, system descriptions in the proceedings of the RTE1 and RTE2 Challenges at <http://www.cs.biu.ac.il/~glikmao/rte05> and <http://ir-srv.cs.biu.ac.il:64080/RTE2/proceedings/> respectively.

false entailment is not found as the predicate of the hypothesis, important in this case, is not taken into account by these generic features.

## 3 Classification

In this section we want to verify that entailment pairs of the ESESR subtype can be recognized. To do this we construct a machine learner. It marks entailment pairs as true if they are of the ESESR type and false otherwise.

To extract the features we build first the word-to-word alignment between the words of the text and hypothesis, based on WordNet.<sup>4</sup> The features for the machine learner are based on the syntactic and semantic relationships between the aligned parts of the text and the hypothesis. We build two sets of features: ones that tell that the entailment is of a given type, and ones that tell that the entailment is not of the given type.

### The syntactic features:

**for:** The syntactic features that are in favour of the ESESR type are the existence of a particular syntactic relationship between the matches of the subject and the object of the hypothesis in text, namely apposition, being within the same noun group, representing a noun group and its prepositional attachment or the combination of the above.

**against:** The syntactic features that indicate that the entailment pair is not of the ESESR type show that the aligned parts of the hypothesis in the text are connected in the text by a predicate or represent the predicate themselves.

**The semantic features:** For the semantic description of the text and the hypothesis we are inter-

<sup>4</sup>Two words are aligned if there is a path between them in WordNet of length  $\leq 3$ . The Cartesian product of the set of the words of the text and the set of the words of the hypothesis yields a set of the candidate word pairs. We used WordNet 2.0 and the C++ API provided by the WordNet developers to look for the paths between the words. We consider the path *travel#v#1* – *walk#v#1* as a path of length 2, where *walk#v#1* is a hyponym of *travel#v#1*, *teakettle#n#1* – *kettle#n#1* – *pot#n#1* is a path of length 3. There can be any WordNet relationships between the nodes in the path except antonyms.

ested in the number of the aligned words, predicates and named entities.

We have 16 features all together. For a more detailed description of the features please refer to Akhmatova and Dras (2007).

The RTE2 test set consists of 800 entailment pairs. Only approximately one tenth of those pairs are ESESR entailments. To build the classifier we have duplicated all the ESESR entailment pairs several times to make the distribution of the entailment pairs equal. (We indeed took care later for the cross-validation that the examples on which we test are not in the training set in this case.) The reason for this is that we are interested in true positives to apply to them an algorithm in section 4. Having only a small proportion of the set being of the ESESR type leads the machine learner to underweight these in the attempt to maximize the overall accuracy and gives a low TP, true positive, rate, which is the one we are interested in. We ran the J48 classifier on the dataset with the one-leaf-out cross validation test mode using the WEKA ML API (Witten and Frank, 1999). The overall accuracy is 75% (see table 1).

## 4 Model

The problem of assigning a value of true or false can be thought of probabilistically, evaluating the conditional probability of the hypothesis  $h$  given the text  $t$ ,  $P(h|t)$ . We can rewrite this using Bayes Rule as

$$P(h|t) = \frac{P(t|h) \times P(h)}{P(t)}$$

An analogy with Statistical MT can be drawn here. As in SMT<sup>5</sup> we divide the calculation of  $P(h|t)$  into two parts, each of which we are able to estimate. One difference is that in SMT we find the argmax of this function to find the best target sentence for the source sentence. This allows us to ignore the denominator. In entailment we must find a threshold that will divide the true entailment pairs from false, so  $P(t)$  will constitute at least a scaling factor. It is true that  $P(t)$  may be different for each text, so whether the common threshold can be found is not obvious. However the related work of Glickman

<sup>5</sup>See, for example, "A Statistical MT Tutorial Workbook," unpublished, August 1999 at <http://www.isi.edu/~knight/>.

et al. (2005) on defining probabilistic textual entailment shows that such a threshold is possible. In this paper we regard it as an empirical question; we discuss it further in Section 4.3.

In SMT  $P(t|h)$  is generally referred to as the translation probability and  $P(h)$  as the language model; but  $P(h)$  is more generally speaking just a prior distribution, the knowledge available in the absence of the more detailed information. In the context of this work, when we know nothing about the extra semantic or syntactic relationships between the words of the text and the hypothesis, the estimation of the probability of the hypothesis sentence is a prior probability of the entailment relation in a pair.

For example, if the text sentence contains *Samuel L. Husk, executive director of the Council of Great City Schools, ...* (see example (4)) then it is more likely in the absence of other knowledge to entail that *Samuel L. Husk works for the Council of Great City Schools*, than that *Samuel L. Husk threw a party in the Council of Great City Schools*. Thus, our expectation is that the former sentence is a more probable sentence in the language than the latter, and that it can be supported by corpus statistics.

### 4.1 Model: part I

To calculate a prior probability of the entailment relation,  $P(h)$ , we adapt the work of Lapata (2001). She was interested in disambiguation of a relationship between an adjective and a noun inside a noun group. Using corpus statistics it was estimated that the adjective *fast* and a noun *planes* in a noun group *fast planes* are much more probable to be in a relationship represented by the word *to fly* (*the planes that fly fast*) than in relationships *to break* or *to land* (*the planes that break fast* or *the planes that land fast*). Similar to that, we want to estimate that, if it is not stated otherwise, the most probable relationship between a person *Samuel L. Husk* and a company *the Council of Great City Schools* is *to work for*.

Thus, similar to Lapata (2001), we calculate the probability of the hypothesis sentence as a probability of a triple consisting of a subject of the hypothesis sentence,  $NE_1$ , its predicate,  $R$ , and a direct or indirect object,  $NE_2$ , that is the probability  $P(NE_1, R, NE_2)$ . We had to take named entities instead of the actual subject and object, as firstly, subject and object very often belong to the set of

TP Rate	FP Rate	Precision	Recall	Class
0.87	0.39	0.69	0.87	FALSE
0.61	0.13	0.84	0.61	TRUE

Table 1: The result of the J48 classifier

standard named entities, such as Person, Location, Organization, JobTitle; and secondly, actual subjects and objects will be rare in the corpus, therefore not allowing us to gather reliable statistics about them.

$$\begin{aligned}
P(h) &:= P(NE_1, R, NE_2) \\
&= P(NE_1|R, NE_2) \times P(R, NE_2) \\
&= P(NE_1|R, NE_2) \times P(R) \times P(NE_2|R).
\end{aligned}$$

We will make an approximation assuming that  $NE_1$  is independent of  $NE_2$

$$\begin{aligned}
P(NE_1|R, NE_2) &\approx P(NE_1|R), \text{ thus} \\
P(h) &= P(NE_1|R) \times P(R) \times P(NE_2|R).
\end{aligned}$$

We estimate the individual probabilities by corpus frequency counts ( $C(x)$  represents the counts of  $x$ )

$$\begin{aligned}
P(h) &= \frac{C(NE_1, R)}{C(R)} \times \frac{C(R)}{\sum_{i=1}^n C(R_i)} \times \frac{C(NE_2, R)}{C(R)} \\
&= \frac{C(NE_1, R) \times C(NE_2, R)}{C(R) \times \sum_{i=1}^n C(R_i)}.
\end{aligned}$$

These probabilities have been calculated pairwise for Location, Person, JobTitle and Organization. The corpus was the first 500,000 sentence of the Wikipedia XML corpus (Denoyer and Gallinari, 2006) parsed using the Minipar parser (Lin, 1998) and Annie plug-ing of the GATE development environment (Cunningham et al., 1996). Table 2 shows a selection of the relations found in the RTE2 dataset. So, for example, *Person work(s) in Location* (at rank 93, with a  $-\log_2(P(h))$  of 10.25) is much more frequent than *Person represent(s) Location* (at rank 775, with a  $-\log_2(P(h))$  of 13.60).

## 4.2 Model: part II

Whereas  $P(h)$  is a prior probability looking only at the relationship between subject and object in the hypothesis,  $P(t|h)$  looks at the aspects of the text that might suggest the entailment relationship. Consider example 4 below.

<u>Person-Location</u>			<u>Person-Organization</u>		
live	182	11.15	attended	4	6.40
resides	711	13.40	works	609	13.70
represents	775	13.60	related	681	14.03
comes	331	12.00	engaged	714	14.18
worked	93	10.25	is player	493	13.31
<u>Organization-Location</u>			writes	242	11.98
operates	36	10.09	command	1206	16.67
based	130	11.41	is representative	206	11.78
located	7	8.13	is head	258	12.13
attended	543	13.93	heads	115	10.93
published	56	10.51	is member	11	8.20
<u>Organization-Organization</u>			occupied	776	14.40
owns	43	10.41	employed	884	14.90
<u>Location-Location</u>			<u>JobTitle-Organization</u>		
located	5	6.15	attended	31	10.53
situated	33	8.87	works	470	15.21
lies	32	8.87	is player	429	14.95
<u>Location-Organization</u>			writes	762	17.40
is subordinate	162	11.86	is head	9	9.12
			heads	73	11.56
			employed	681	16.62

Table 2: Some relations extracted from the first 500,000 sentences of the Wikipedia XML corpus. The three columns give the relation, its rank in a sorted list, and the value  $-\log_2(P(h))$  respectively.

- (4) *Text*: “Relative size and the power of the purse are certainly key factors,” says Samuel L. Husk, executive director of the Council of Great City Schools.

*Hypothesis*: Samuel L. Husk works for the Council of Great City Schools.

There is a direct syntactic connection between *Samuel L. Husk* and *executive director of the Council of Great City Schools*. By contrast, consider example 5.

- (5) *Text*: Both aftershocks had their epicentre around the Nicobar island group in the south of archipelago that lies close to *Indonesia*, *India’s Meteorological Department* said.

*Hypothesis*: *India’s Meteorological Department* operates from *Indonesia*.

There is no syntactic relationship between *India’s Meteorological Department* and *Indonesia*, suggesting the hypothesis is not a valid entailment.

Our approach to estimating  $P(t|h)$ , then, is to decide whether particular relationships in the text hold. To do this we built a classifier with various classes of features.

**Features 1 and 2** syntactic structure of the text sentence: presence or absence of the syntactic connection between the aligned elements; type of the syntactic relationship, if present.

**Features 3 – 6** alignment: number of non-aligned words between the aligned noun groups, number of the non-aligned head elements of the aligned noun groups.

**Features 7 and 8** syntactic structure of the aligned noun groups.

**Feature 9** paraphrases.

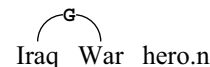
We have already briefly mentioned above the importance of the syntactic dependencies between the matches of the subject and object of the hypothesis in the text.

The alignment features capture the fact that if there are too many missed words in the aligned noun groups then the hypothesis might have acquired different meaning from the one expressed in the

text. Non-aligned head elements of the noun groups greatly increase the possibility of the meaning altering.

In determining the existence of syntactic relationships within the text, we use the Link Grammar Parser (Sleator and Temperley, 1991). To give an example for the features 7 and 8, the link G, for example, connects proper noun words together. For example, *MIT* and *Press* in the *MIT Press Bookstore*, see example (6), as well as *Iraq* and *War* (see example (7)), will be connected by the link G. We would say that the hypothesis is closer to the text if from the noun groups *MIT Press Bookstore* and *the Iraq War hero* the whole parts *MIT Press* and *Iraq War* were present in the hypothesis, rather than just *MIT* or *Iraq*, for example. If it is not the case and one can see only the first parts of the *MIT Press* and *Iraq War* components of the text sentence, then we say that the G link is ‘broken’. A broken G link reduces the probability of the true entailment between the text and the hypothesis.





In the examples (6) and (7) the G relation in the noun groups was broken. *The MIT Press* was substituted with *The MIT*, *Iraq War* with *Iraq*. That led to the hypotheses that the meaning of the text is not entailed correctly.

- (6) *Text*: The MIT Press Bookstore stocks most of the books and journals published by The MIT Press as well as the best of other publishers books in related fields.

*Hypothesis*: The MIT is a book store.

- (7) *Text*: The State Department is making the unusual offer of giving expedited visas to the Cuban sons of Iraq War hero Sgt. Carlos Lazo, so they can visit him in the United States, people familiar with the case said Friday.

*Hypothesis*: Sgt. Carlos Lazo worked in Iraq.

The link GN connects a proper noun to a preceding common noun which introduces it.

Iraq War hero.n Sgt. Carlos Lazo

MX connects modifying phrases with commas to preceding nouns. Thus, *Sgt. Carlos Lazo* is connected to the *Iraq War hero* in *Iraq War hero Sgt. Carlos Lazo* by the GN link. It is the same for the *Maricopa County Superior Court Judge* and *Lindsay Ellis* in the *Maricopa County Superior Court Judge Lindsay Ellis*, see example (8). In case the *Iraq War hero* and *Sgt. Carlos Lazo* were in the sentence in the relation of apposition, for example, *Sgt. Carlos Lazo, an Iraq War hero*, they would be connected by an MX link. That makes GN and MX links to be equivalent for us here. The parts connected by the links GN and MX are substitutable, *Sgt. Carlos Lazo* is a *hero*, *Lindsay Ellis* is a *judge*. Thus, if the head nouns in *Maricopa County Superior Court Judge* and *Iraq War hero* are not aligned the hypothesis still might be true.

- (8) *Text*: Maricopa County Superior Court Judge Lindsay Ellis also ordered Miss Bickel to pay \$5,000 in restitution to Miss Tomazin’s family and to perform 40 hours per week of community service indefinitely.

*Hypothesis*: Lindsay Ellis occupies a post at the Superior Court.

Feature 9 is the number of paraphrased phrases.

- (9) *Text*: Mahmoud al-Zahar , a Hamas leader in Gaza, said so explicitly, dismissing Mr. Abba’s arguments: History has proven that the rockets have been in the Palestinian interest.

*Hypothesis*: Mahmoud al-Zahar is a member of Hamas.

*Leader* and *member* are not synonyms, but they will be found to be paraphrases of each other by the algorithm proposed in Bannard and Callison-Burch (2005). To acquire the paraphrases we used the PhraseBuilder<sup>6</sup> on English and Dutch corpuses of Europarl.

<sup>6</sup>we have used the PhraseBuilder by Simon Zwarts <http://www.ics.mq.edu.au/~szwarts/Downloads.php>

### 4.2.1 Deriving a Probability

We have selected the  $k$ -nearest neighbours method, which has quite a transparent method of calculating the probability for an instance to belong to a particular class (Mitchell, 1997). We used WEKA API  $k$ -nearest neighbours method implementation for our work.

We then derive a probability from our classifier. In classification, classified instances will fall at varying distances from the boundaries which define the class spaces. This can correspond, for example, to the certainty of classification, and various classification methods have a derived probability of classification. In our case, with classes being true entailment and false entailment, we can use this as an estimate of  $P(t|h)$ .

The accuracy of the machine learner built on these features with  $k = 5$  is not high, 54%, on the one-leaf-out approach. We are interested here though in the probabilities of belonging to a particular class rather than in the classification.  $P(true|instance) = 0.49$  is the same for us here as  $P(true|instance) = 0.51$ . That means that the algorithm is not actually sure to which class the instance belongs. That the  $P(true|instance)$  is greater than, say, 80% would be an important clue in the class prediction.

### 4.3 Combining part I and part II

For calculating our  $P(h|t)$ , as defined at the start of the Section 4, we have estimates of  $P(h)$  and  $P(t|h)$ . We will assume that  $P(t)$  is a constant for all entailment pairs and acts as a normalizing factor. (This may not be true, but we treat it here as an empirical question.)

We want then to find a threshold  $H$  for  $P(h|t)$ , such that where  $P(h|t) \geq H$  the entailment pair is true, and false otherwise. The threshold  $H$  then incorporates the normalizing factor  $P(t)$ .

We have created a balanced corpus of the *true* and *false* examples of the ESESr entailment pairs from the RTE2 dataset. Then, as the one-leaf-out approach suggests, for every instance (that is, for every entailment pair) we created a separate dataset not containing it to build the  $k$ -nearest neighbours classifier. The probability of the instance being a true entailment on this classifier is the outcome of the

baseline	unbalanced dataset performance	balanced dataset performance
41 submissions mean	61.9%	50%
best performing on ESESR system	86%	73%
secondbest system	74%	55%
default “yes”	78%	50%

Table 3: Baselines and their performance on the balanced and unbalanced datasets

classification process, see the section 4.2.1. Then this probability is combined with the probability of the hypothesis  $P(h)$ , described in the section 4.1. This process is repeated for every entailment pair. Thus, as a result, every entailment pair is associated with a value of the probability  $P(h|t)$ .

One possibility to find a good value of such an  $H$  is to carry out a search over possible values on a development set. As an alternative we used a machine learner again, a decision tree, with the single feature being the combined probability. The top node of the decision tree is the best split of data. Due to the fact that the probabilities  $P(h)$  are quite small numbers, we used as a feature for the decision tree also the product of the logarithms base two of the probabilities. Even though this is not strictly derivable from our model, it is still a ranking and we get a good threshold. The threshold  $H = 3.41$  fits the training set best of all.

## 5 Evaluation

We compare the results of the approach on two datasets, *an unbalanced dataset* consisting of all the ESESR entailments from the RTE2 corpus; and *a balanced dataset*, the set of 50000 random balanced subsets of the unbalanced dataset containing all the false entailments and the same number of randomly chosen true entailments (refer to section 2.2).

We take four baselines as a comparison for our approach:

1. the mean of the accuracy of all the 41 submissions to the RTE2 Challenge;
2. the best performing on the ESESR entailment pairs system;
3. the second best system on ESESR entailment pairs; and

4. the default algorithm that gives “yes” to all the entailments, due to the fact that the majority of the ESESR entailment pairs in the RTE2 test set are true entailments

Refer to the Table 3 to find the evaluation of the performance with respect to the baselines.

We are particularly interested in the balanced dataset, as we do not know the proportion of the true and false entailments of a given type in an arbitrary context.

Our system gets 80% accuracy on the unbalanced dataset and 59% accuracy on the balanced dataset. That means that our method performs noticeably better than the average of the methods from RTE2 Challenge and the “yes” to all baseline on both datasets. It scores about 18% higher than the average and 2% higher than the “yes” to all algorithm on the unbalanced dataset; and 9% higher than these two algorithms on the balanced dataset. Further, our results are higher for all but the best system in the Challenge for this subtype.

## 6 The conclusions and future work

In the current work we have identified a subtype of entailment pairs; presented a machine learner that distinguishes the subtype among the entailment pairs; and presented a probabilistic model that evaluates the conditional probability of the hypothesis given the text. We then evaluated the algorithm against a baseline and two other systems. The result is that the algorithm performs significantly better than the baseline (from 9% up to 18% better) and all but the best system in the Challenge for the type of entailment pairs we are interested in.

We plan to address other subtypes similar to ESESR entailment groups thus contributing more to the recognizing specific types of entailments.



## References

- Elena Akhmatova and Mark Dras. 2007. Syntactically encoded semantic relationships type of entailment pairs. Available from <http://www.ics.mq.edu.au/elena/pub.html>.
- Colin J. Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL*.
- Hamish Cunningham, Yorick Wilks, and Robert J. Gaizauskas. 1996. Gate-a general architecture for text engineering. In *COLING*, pages 1057–1060.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *MLCW*, pages 177–190.
- Ludovic Denoyer and Patrick Gallinari. 2006. The wikipedia xml corpus. *SIGIR Forum*, 40(1):64–69.
- Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. A lexical alignment model for probabilistic textual entailment. In *MLCW*, pages 287–298.
- Maria Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *NAACL*, pages 63–70.
- Dekang Lin. 1998. Dependency-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems*.
- Tom Mitchell. 1997. *Machine Learning*. McGraw Hill.
- Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A grammar of contemporary English*. Longman, Singapore.
- Daniel Sleator and Davy Temperley. 1991. Parsing english with a link grammar. Available at <http://www.link.cs.cmu.edu/link/papers/index.html>.
- Lucy Vanderwende and William B. Dolan. 2005. What syntax can contribute in the entailment task. In *MLCW*, pages 205–216.
- Lucy Vanderwende, Arul Menezes, and Rion Snow. 2006. Microsoft research at rte-2: Syntactic contributions in the entailment task: an implementation. In *2nd PASCAL Challenges Workshop on Recognizing Textual Entailment*.
- Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.