Corpora for Learning the Mutual Relationship between Semantic Relatedness and Textual Entailment

Ngoc Phuoc An Vo[1,2,3], Octavian Popescu[4]
Xerox Research Centre Europe[1], Fondazione Bruno Kessler[2],
University of Trento[3], IBM Research[4]
an.vo@xrce.xerox.com, o.popescu@us.ibm.com

Abstract

In this paper we present the creation of a corpora annotated with both semantic relatedness (SR) scores and textual entailment (TE) judgments. In building this corpus we aimed at discovering, if any, the relationship between these two tasks for the mutual benefit of resolving one of them by relying on the insights gained from the other. We considered a corpora already annotated with TE judgments and we proceed to the manual annotation with SR scores. The RTE 1-4 corpora used in the PASCAL competition fit our need. The annotators worked independently of one each other and they did not have access to the TE judgment during annotation. The intuition that the two annotations are correlated received major support from this experiment and this finding led to a system that uses this information to revise the initial estimates of SR scores. As semantic relatedness is one of the most general and difficult task in natural language processing we expect that future systems will combine different sources of information in order to solve it. Our work suggests that textual entailment plays a quantifiable role in addressing it.

Keywords: Semantic Relatedness, Textual Entailment, RTE Corpora

1. Introduction

In the last years, two tasks that involve meaning processing, namely Semantic Relatedness (SR) (Marelli et al., 2014b) and Textual Entailment (TE) (Dagan et al., 2006), have received particular attention. In various academic competitions, both the TE and SR tasks have been proposed, and useful benchmarks have been created. The SR scores are usually given in the [1-5] interval, while the TE values are discrete signaling an entailment, a contradiction or a non-related relationship between candidate sentences.

Interestingly, the core approach seems to be similar for both of these tasks (Agirre et al., 2012; Dagan et al., 2013). Yet, till 2014, no corpus annotated with both SR and TE labels was available. In the last SemEval 2014, the SICK corpus containing both SR and TE annotations for the same pairs have been released (Marelli et al., 2014a). This corpus allows a direct comparison between the systems that addresses both tasks.

We have developed a technique that uses both SR and TE scores in order to improve the accuracy on each of these tasks. We show that when we take into account the estimates of one task in resolving the other better results are obtained. The strategy we propose is based on the correlation between SR scores and TE values. The main idea is that, in the first phase, we determine an estimation of the SR scores which are taken into account in deciding TE values at a second phase, which are used to re-adjust the SR scores at a third phase. The main intuition is that high/low SR scores signal ENTAILMENT/NEUTRAL TE values, and from correct TE values, we can readjust the initial SR scores, by adding or subtracting a quantity, such as to minimize the SR errors of over/underestimating in those cases correlated with TE values.

We applied the above strategy on SICK SR scores and TE values, and we observed a significant improvement of the results in both tasks. In order to verify further this hypothesis we considered the RTE 1-4 corpora proposed for Pascal Recognizing Textual Entailment Challenge by PASCAL and NIST between 2006 and 2009 (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Giampiccolo et al., 2009). These corpora are already annotated with TE values and we independently annotated each corpus with SR scores. We call the new corpora RTE+SR 1-4 in order to distinguish it from the original ones. The main reason that we choose to annotate the RTE corpora with semantic relatedness scores, instead of semantic similarity scores is that the RTE corpora contain not only the "is a" relation (occurring for semantic similarity) but also other broader semantic relations such as antonymy and meronymy between two texts (corresponding to NEUTRAL and CONTRADICTION relations in RTE).

In this paper we report on the process of creating RTE+SR corpora and we analyse the correlation between TE values and SR scores on both SICK and RTE+SR corpora. We also investigate and report on the benefits of using the mutual relationship that exists between SR and TE in resolving either of the two tasks.

This paper is organized as follows: we review the relevant literature in the Related Work section. We present the process of annotating the RTE 1-4 corpora with SR scores in the Building the RTE+SR Corpora section. In the next section, we analyze the correlation between SR and TE scores in SICK, and RTE+SR 1-4 corpora. In SR-TE System Architecture section we present an architecture for addressing the SR and TE tasks taking advantage of their mutual relationship. The paper ends with the Conclusion and Further Research section.

2. Related Work

As understanding sentence's meaning is a crucial challenge for any computational semantic system, there are several attempts on creating corpora for evaluating this task, including Recognizing Textual Entailment (RTE) and Semantic

Textual Similarity (STS) or Relatedness (SR).

The RTE task requires the identification of a directional relation between a pair of text fragments, namely a text (T) and a hypothesis (H). The relation (T $\rightarrow$ H) holds if, typically, a human reading T would infer that H is most likely true. For entailment problem, the first dataset created is the Framework for Computational Semantics (FraCas) (Cooper et al., 1996). The dataset contains entailment problems in which a conclusion has to be derived from one or more premises. And it is not necessary that all premises are needed to verify the entailment. Premises and conclusions are simple, artificial (lab-made), English sentences involving semantic phenomena frequently addressed by formal semanticists, such as generalized quantifiers, ellipsis and temporal reference.

The next one is the RTE datasets created for the PASCAL RTE (Recognizing Textual Entailment) challenge. The RTE datasets are more steadily developed and challenging for entailment problems. From RTE 1-3 (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007), it was a binary classification problem for only two relations: YES and NO, regarding to entailment and non-entailment. However, until RTE-4 (Giampiccolo et al., 2009), a more fine-grained classification problem with respect to a three-way relation was proposed as: ENTAILMENT, NEUTRAL and CONTRADICTION. The RTE datasets are widely applicable for number of applications, such as Question Answering, Information Retrieval or Information Extraction.

In contrast, the STS/SR task requires to identify the degree of similarity or relatedness that exists between two text fragments (phrases, sentences, paragraphs, etc), where similarity is a broad concept and its value is normally obtained by averaging the opinion of several annotators. While the concept of semantic similarity is more specific and it only includes the "is a" relations between two texts, the semantic relatedness may be broader and may include any relation between two terms such as antonymy and meronymy. In STS/SR, systems are required to quantitatively measure the degree of semantic similarity or relatedness between two given sentences which may be derived from heterogeneous data sources. The first STS task was proposed as a pilot task at SemEval 2012 (Agirre et al., 2012) and then continues to 2016. The STS/SR task could be widely applicable in a number of NLP tasks such as Text Summarization, Machine Translation evaluation, Question Answering, Ranking task in Information Retrieval (IR), etc. The STS datsets are created from different data sources, including newswire, video descriptions, Machine Translation evaluation, WordNet, FrameNet and OntoNotes glosses, newswire headlines, forum posts, tweet-news pairs, student answers/reference answers, answers in Stack Exchange forums, and forum data exhibiting committed belief. Each sentence pair is annotated with the semantic similarity score in the scale 0 (no relevance) to 5 (semantically equivalence).

However, before the Sentences Involving Compositional Knowledge (SICK) corpus arrives, though these two tasks are highly related and there are several datasets for evaluating individual task (either RTE or STS/SR), there was no corpus containing the annotations for both tasks. We believe that the SICK corpus is the first dataset which contains the manual annotation for both tasks Semantic Relatedness (SR) and Textual Entailment (TE) (Marelli et al., 2014a). It was created to evaluate the Compositional Distributional Semantic Models (CDSMs) handling the challenging phenomena, such as contextual synonymy and other lexical variation phenomena, active/passive and other syntactic alternations, impact of negation, quantifiers and other grammatical elements, etc. The SICK includes a large number of sentence pairs (around 10,000 English sentence pairs) that are rich in the lexical, syntactic and semantic phenomena that CDSMs are expected to account for, but it is not required to deal with other aspects of existing datasets (multiword expressions, named entities, numbers,) that are not within the domain of compositional distributional semantics. Each sentence pair in SICK is annotated for semantic relatedness score in the semantic scale [1 - 5] and textual entailment relation in 3-way: ENTAILMENT, NEUTRAL, and CONTRADICTION.

Though SICK is the first corpus having annotation for both related tasks RTE and SR, it is only meant to use for evaluating CDSMs with its typical phenomena rather than traditional computational semantic models. To the best of our knowledge, our contribution in annotating the semantic relatedness scores for RTE datasets is the first attempt to bridge these two related tasks in consideration of creating the first corpora that is rich not only in lexical, syntactic, semantic, but also logical, reasoning and other phenomena (multiword expressions, named entities, style, dates, numbers, etc) existing in real-life natural language sentences (different from SICK corpus) for learning the mutual relation between these two tasks to benefit many computational semantic systems in NLP.

3. Building the RTE+SR Corpora

Since we reuse the texts from RTE 1-4 corpora for annotating the SR scores, the texts are already being normalized and preprocessed. We extract the texts (including text and hypothesis) and associated information such as pair ID and entailment relations from the original RTE corpora for analysis and SR annotation. In order to be able to draw an easy and meaningful comparison, we follow the annotation guideline in SICK corpus (Marelli et al., 2014a). For the RTE 1-4 corpora, we use the same SR interval scores and the same type of TE values, that means that a pair of sentences could have a SR score on a 5-point-semantic-scale [1-5] that ranges from 1 (completely unrelated) to 5 (very related) and that there are TE values ENTAILMENT, CONTRADICTION and NEUTRAL. However, as the texts in RTE corpora are usually not equal in size/length between the text (T) and the hypothesis (H), and most of the time, one text is (much) longer than the other one, we need to have a specific rule for this case. We define and apply the rule that annotators do not penalize the shorter text if it is fully and semantically covered by the longer text. It means that the annotators consider it is a semantic equivalence and give highest SR score to the text pair if the longer text semantically covers the shorter one. As in SR annotation, the two texts are considered as same level, hence, we remove the concept of text (T) and hypothesis (H) in

RTE, we only consider the two texts as sentence A and sentence B. The output of our SR annotation is similar to the SICK corpus which consists of five tab-separated columns: *pair_ID, sentence_A, sentence_B, entailment_judgment, relatedness_score*. However, due to the license restriction of redistribution of the RTE 4 corpus, we only can release the annotation with two tab-separated columns *pair_ID, relatedness_score*. Thus, users need to acquire the original RTE-4 corpus by themselves.
We provide some examples of SR annotation on the texts extracted from the RTE corpora as follows:

- *(Green cards are becoming more difficult to obtain.)* VS. *(Green card is now difficult to receive.)* (SR score = 5)
- *(Researchers at the Harvard School of Public Health say that people who drink coffee may be doing a lot more than keeping themselves awake - this kind of consumption apparently also can help reduce the risk of diseases.)* VS. *(Coffee drinking has health benefits.)* (SR score = 4.75)
- *(It rewrites the rules of global trade, established by the General Agreement on Tariffs and Trade, or GATT, in 1947, and modified in multiple rounds of negotiations since then.)* VS. *(GATT was formed in 1947.)* (SR score = 3)
- *(The Dutch, who ruled Indonesia until 1949, called the city of Jakarta Batavia.)* VS. *(Formerly (until 1949) Batavia, Jakarta is largest city and capital of Indonesia.)* (SR score = 2.2)
- *(President Bush returned to the Mountain State to celebrate Independence Day, telling a spirited crowd yesterday that on its 228th birthday the nation is "moving forward with confidence and strength.")* VS. *(Independence Day was a popular movie.)*(SR score = 1)

We split the RTE 1-4 corpora into parts of roughly 400 pair of sentences each. There were three annotators that independently annotated parts of the RTE 1-4 corpora. Each part has been annotated by at least two annotators.
Before the annotators started annotating the RTE 1-4 corpora with SR scores, we carried out a training session for annotators in order to calibrate their judgments. For this purpose we used 600 pairs of sentences extracted from the corpora compiled for the semantic text similarity task in SemEval 2012 - 2015 and 300 pairs of sentences extracted from the SICK corpora. We consider discrete scores, with step 0.5, that is the gold standard annotation was rounded to the nearest integer. The annotators re-annotated this annotation training corpus independently of one another and also independently of the gold standard annotation. The goal was to individually maximize post annotation agreement between each annotator and the golden standard. In order to achieve this goal, we split the annotation training corpus in two, the first part containing two thirds of total number of pairs.
The first part was re-annotated and all the cases in which there was a disagreement of more than one point between annotators or the gold standard were discussed collectively. After that, the second round of re-annotation was carried out on the last third of the corpus. As the agreement between annotators was sufficiently high, except for some debatable cases, we considered the training section over and we started annotating the RTE 1-4 corpora. In Table 1 we present the results of the training phase. A cell represents the Cohen's kappa coefficient computed between an annotator and the golden standard on the two annotation training corpora.
The RTE+SR 1-4 SR scores were given by averaging over the annotators. The Cohen's kappa coefficient of inter-agreement on RTE+SR was above 0.7 in overall. Few cases were discussed, less than 5%, in order to leverage over big differences in annotation. However, the annotation difficulty was not homogeneous among the RTE 1-4 corpora. We present a detailed analysis of the difficulties for each corpus separately in order to have a complete overview of the problems encountered during annotation. But firstly, we look at some common difficulties.
One major problem is that some of the RTE corpora lack the CONTRADICTION category. To begin with, this category is particularly troublesome. Are two contradictory sentences are similar or not? In the SICK corpus this problem seemed not to be formally addressed and one can notice that there is in this corpus a variation of the similarity score for pairs which are in a CONTRADICTION relationship. To add to the complexity of making a decision, hardly was any pair of sentences in RTE corpora in a perfect logical contradictory relationship as in:

- *(The boy is playing the piano.)* VS. *(The boy is not playing the piano.)*

In fact, in many cases the contradiction was rather partial, the main idea being the same but with a slight variation:

- *(Monica Meadows, a 22-year-old model from Atlanta, was shot in the shoulder on a subway car in New York City.)* VS. *(Monica Meadows, 23, was shot in shoulder while riding a subway car in New York City.)*
- *(Doug Lawrence bought the impressionist oil landscape by J. Ottis Adams in the mid-1970s at a Fort Wayne antiques dealer.)* VS. *(Doug Lawrence sold the impressionist oil landscape by J. Ottis Adams.)*
- *(Mitsubishi Motors Corp new vehicle sales in the US fell 46 percent in June.)* VS. *(Mitsubishi sales rose 46 percent.)*
- *(Four people were killed and a US helicopter shot down in Najaf.)* VS. *(Five people were killed and an American helicopter was shot down in Najaf.)*

In the above examples, the similarity between sentences is so high that one can recognize that both of them cannot be true at the same time, which is a valid definition of CONTRADICTION. However, they do not contradict each other in a perfectly recognizable ways and they are not similar in the sense that the information is the same. In fact such examples show the multi-dimensionality of similarity.

	Gold Standard 2/3	Gold Standard 1/3
Annotator 1	0.67	0.84
Annotator 2	0.72	0.83
Annotator 3	0.68	0.80

Table 1: Annotation Training Corpus with Cohen's coefficient.

We think that an analysis on what is considered CONTRADICTION and what is not considered as such has to tell a lot about similarity in general. Without such analysis any choice would be at least incomplete.

For this experiment, we chose to consider the contradiction as a form of high similarity between two sentences. This choice is motivated by the fact that without similarity sentences are unrelated. The exact importance of the "scrambled" information was left to the annotator, but all of them were instructed to have it bounded below by a high score.

In RTE-1 the main difficulty consists in highly intricate relationships between the sentences. From our point of view, RTE-1 was one of the most difficult corpora for SR problem. In many cases the hypothesis is not readily deductible from the text and the process may involve very different types of information. In RTE-2 the difficulty consists in coping with groups of sentences with highly word overlap rate. The RTE-3 is different from the previous two challenges and much more strategies seem to work on RTE-3 than in RTE-1 and RTE-2. A direct comparison on the systems running with comparable accuracy on these corpora may not be possible. The RTE-4 introduces many to one mapping. However, it is not clear how this could influence the performances of certain systems overall. For example, a bag of word approach may care less about sentence boundaries anyway. However the similarity may become even a fuzzier concept. For this situation we fell that a more elaborate annotation guidelines are necessary. However, one may have the feeling that in provided more guidance it actually restricts the annotator's choices more or less arbitrarily. We think that more analysis is required in order to safely make a jump from one to one similarity to multiple to one similarity.

4. Correlation between the SR and TE Scores

The main question we seek an answer to is whether there is any correlation between the SR and TE scores. Is there a linear dependency among them, or, at least, could it be predicted with a satisfactorily confidence that a certain TE value implies a SR score within a given interval? In a nutshell, the answer is "yes", and this correlation is instrumental in developing a system for enhancing the accuracy for both tasks.

We carried out the analysis of correlation between ST and TE on both train and test corpora considering the SICK corpus. In order to build a system that addresses these two tasks, the correlation is carried out only on the training corpus. We present and discuss here the analysis on training, but there are no significant differences between training and test for any of the RTE+SR corpora.

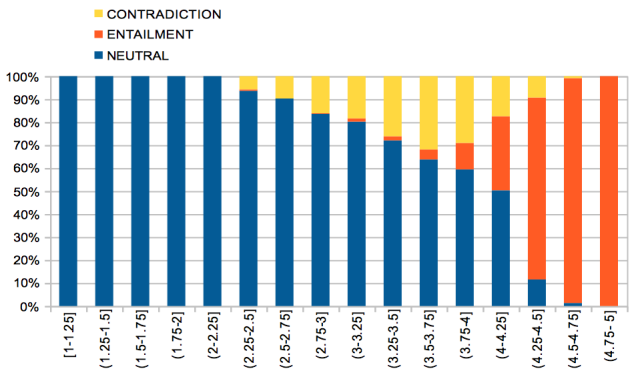

Figure 1: Correlation between SR and TE on SICK Train corpus.

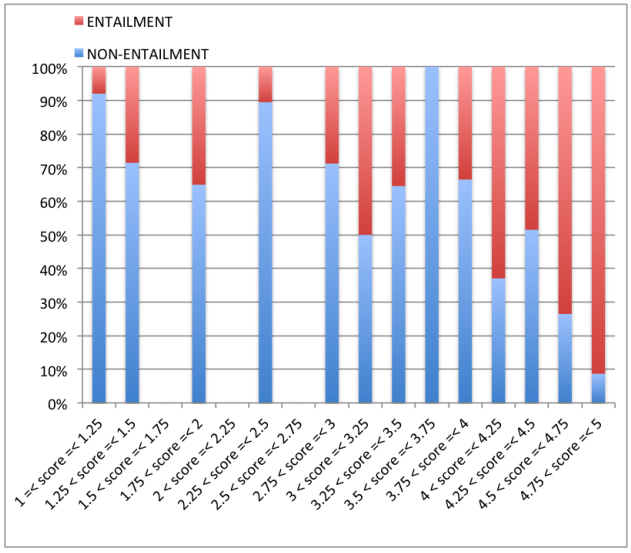

Figure 2: RTE-1 Test - No indications.

The relatedness score is a continuous variable in the range [1-5], while the entailment type is a discrete variable with three possible values: ENTAILMENT, CONTRADICTION and NEUTRAL. In order to conduct a useful analysis we also normalize the values of the relatedness score. We considered intervals of e length, and we let e vary from 1 to 1/10. We noticed the consistency between the relatedness intervals of length e and the entailment type, by measuring the distribution of the entailment type inside each interval. For the SICK training data it turns out that the best trade-off between large interval dimension and high purity is obtained for = 1/4. The distribution of entailment types inside the relatedness interval of this length, i.e. 0.25, is plotted in Figure 1.

The results of the previous analysis show that there is a strong correlation between relatedness and entailment type. Looking at the marginal relatedness scores, such as [1-2] or [4.25-5], one can predict the entailment type on the basis of the relatedness score. And vice versa, which means that we can correct the prediction of one variable by regressing it to the other. We are interested mainly in the correction

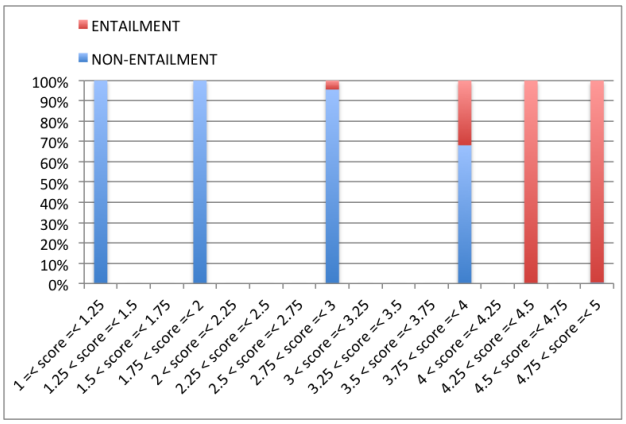

Figure 3: RTE-3 Test - No indications.

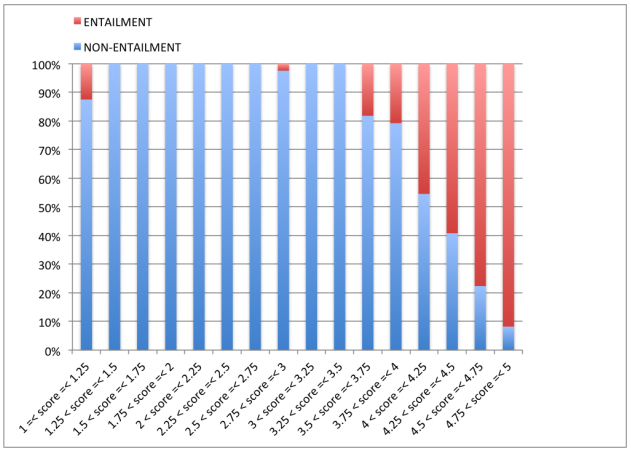

Figure 4: RTE-1 Train - Common guidelines.

of the relatedness score via the entailment decision. To this end, we computed the probability of the median of each relatedness score given the entailment.

The plots for RTE 1-4, training corpus, follow, up to a certain extent, the same shape. There are, though, notable differences from corpus to corpus. The SICK corpus seems to be rather a fortunate case for this type of linear dependency

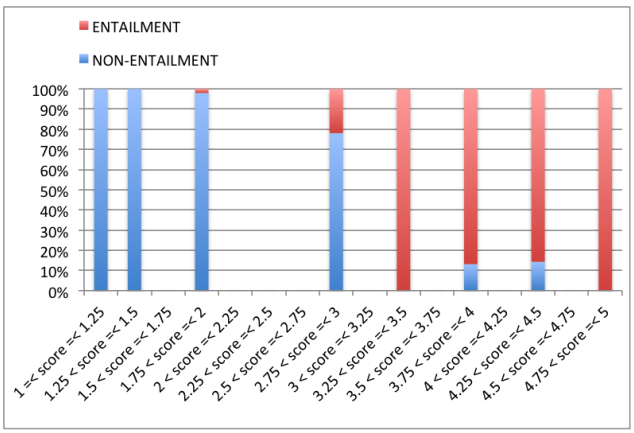

Figure 5: RTE-3 Train - Common guidelines.

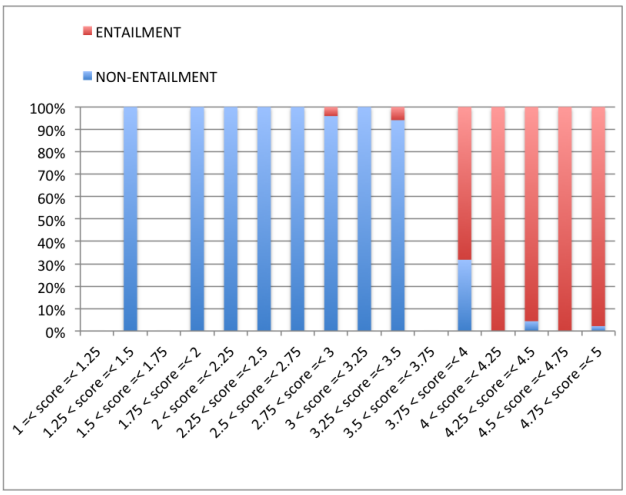

Figure 6: RTE-2 Test.

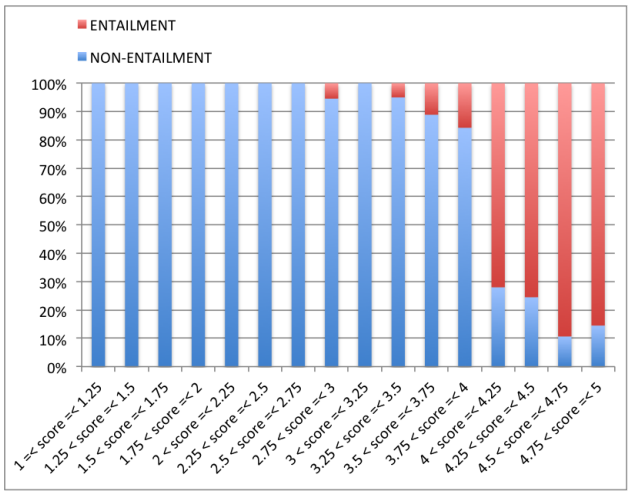

Figure 7: RTE-2 Train.

between TE and SE scores. An unfortunate example of this relationship is probably best represented by RTE-1. However, in all cases, there is a strong correlation between TE and SR scores, especially on the extremities, high SR score vs. ENTAILMENT/CONTRADICTION and low SR and NEUTRAL respectively. We present the plot for each corpora separately (in the final version of the paper). However, only the degree of this correlation may vary, but it is general that ENTAILMENT is associated with high similarity.

What we observed is the dependency of this correlation on the annotation guidelines. For example, one of the annotators, without any guidance produces the chart plotted in Figure 2. We can see in Figure 2 a totally different picture than in Figure 1. This could be a direct consequence of the complexity of the corpus or of a particular view on similarity.

The fact that the same annotator, in the same conditions, has obtained the following charts for RTE-3 suggests that the first, rather than the second might be true.

The plot in Figure 3 is pretty similar to the plot in Figure 1. In both of them a clear threshold for ENTAILMENT

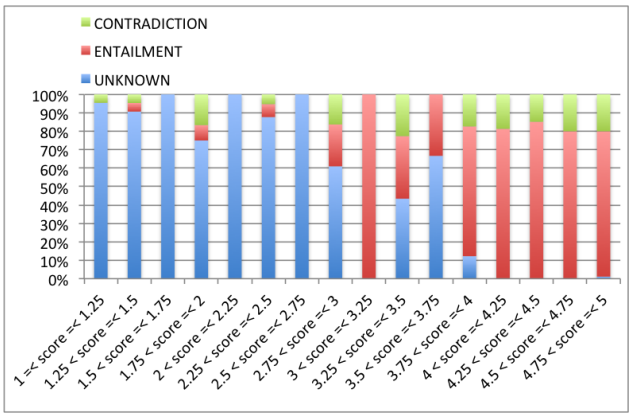

Figure 8: RTE-4.

vs. NEUTRAL can be observed. Also in Figure 2 the linear regression of TE to SR scores can be inferred with high accuracy. This suggests that indeed the dependency of relationship between SR and TE is valid in general, but its exact form may also depend on the annotation guidelines and on the particular type of corpus.

When the annotation guidelines were applied by everyone the differences tend to disappear. This shows that our notion of relatedness in language may not be an objective measure but rather a multidimensional measure which relies on combinations of other measures. The computational analysis of both tasks should be very beneficial for large NLP systems.

Interestingly the plot for RTE-1 data has changed substantially, but the plot for RTE-3 data shows basically the same relationships as in Figure 3. Our analysis suggests that the main factor responsible for the difference between Figure 2 and Figure 5 is the clarification on what "same information" should mean. In the cases where there is entailment, the entailed sentence is just a part of the information in the text. But by itself, the entailed information could have been inferred from different sentences as well (a consequence may have different, unrelated antecedents). Would the "same information" definition include some restrictions on what the antecedents are as well? This may lead to a different type of relationship between SR and TE. However, the bottom line is that the dependency between these two annotations is quantifiable in each case. A system that will take into consideration the conditioned probability of a certain SR score when the entailment relationship is known, or vice-versa, will be presented in the next section.

The plots for RTE-2 data show that there is little variation from the relationship described above.

For RTE-4 there is only one corpus available. For this corpus it was possible to observe the distribution of the CONTRADICTION relationship. Without any guidelines on CONTRADICTION the distribution of this type of entailment across different SR classes is rather uniform. We believe that letting to the individual intuition alone to establish what is the relationship between contradiction and relatedness is not good enough, as this is a multidimensional problem. A strict guidelines set is not good either as they may be hard to follow and incomplete. A solution is to consider a more structural definition of relatedness, which allows the decomposition of a sentence into more or less independent pieces of information.

In Figure 9 we present an architecture that implements this idea. The two main modules, Distributional Module and Structural Module should contain different types of text processing. For first estimation of the SR score may be efficiently obtained by using essentially bag of words techniques (see Figure 10)

5. SR - TE System Architecture

The mutual relationship described in the previous section suggests that a dual leverage architecture may be beneficial in addressing the ST and TE tasks. The main idea is to be able to adjust the TE or SR final decision taking into consideration an initial estimate. In general, good prediction can be obtained for SR by using a distributional strategy. Words are aligned between the two sentences by their similarity and the pair SR score is computed on the basis of individual similarities penalized accordingly to the alignment schema. To decide on the TE values, usually a more local and structural analysis focusing on various linguistics aspects - coordination, negation, semantic roles, etc must be considered as well. However adding this complexity comes with a price on accuracy. Therefore is better to have a more robust way to ensure that the structural analysis does not deviate due to errors.

We propose an architecture in which the first SR score estimates are also used to indicate the entailment. A different module carries out a deep analysis on the sentence structure in order to decide the entailment especially for those border line cases signaled by the SR score, that is scores that are not very high but not very low either The entailment judgment is used to recompute the SR score, according to the following rule described in Figure 9.

However, the Structural Module in Figure 11 should employ a different class of techniques for determining the entailment. As the name of this module shows, a structural analysis of the sentence has to be performed. A technique that uses structural information to infer the entailment relationship was presented in (Popescu et al., 2011; Vo et al., 2014). The structural module should involve more detailed analyses of the semantics of the sentence.

The recent advancements in neural networks show that it may be possible to train a system of a couple of neural networks to decide on entailment with a greater accuracy than before (Rocktäschel et al., 2015). This technique can be used in the Structural Module as well.

The initial SR scores are recomputed after the entailment decision is made. We show in Figure 12 and Figure 13 respectively how the re-computation affects the accuracy. We considered the initial distribution of SR scores in intervals of 0.5 length for the gold standard class [2-3] initially, that is after the distributional module output. You can see a quasi-normal distributions into the whole set of classes (Figure 12). After the re-computation the distribution of scores looked like in Figure 13. A 20% increases in accuracy was obtained.

The approach was able to correct massively errors that were off more than 1 unit. The bias towards the lower extreme

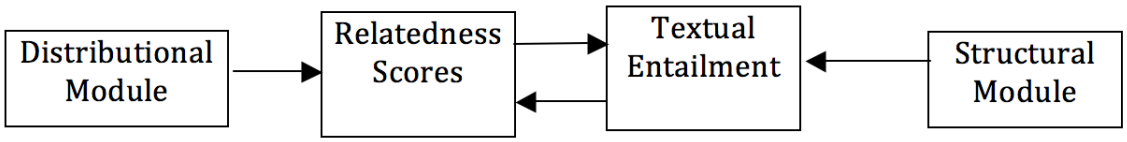

Figure 9: SR revision scores architecture if the entailment test is positive/neutral then a quantity is added/subtracted to the SR score such that the correlation between TE value and SR score is maximized. The above parameters are set at training.

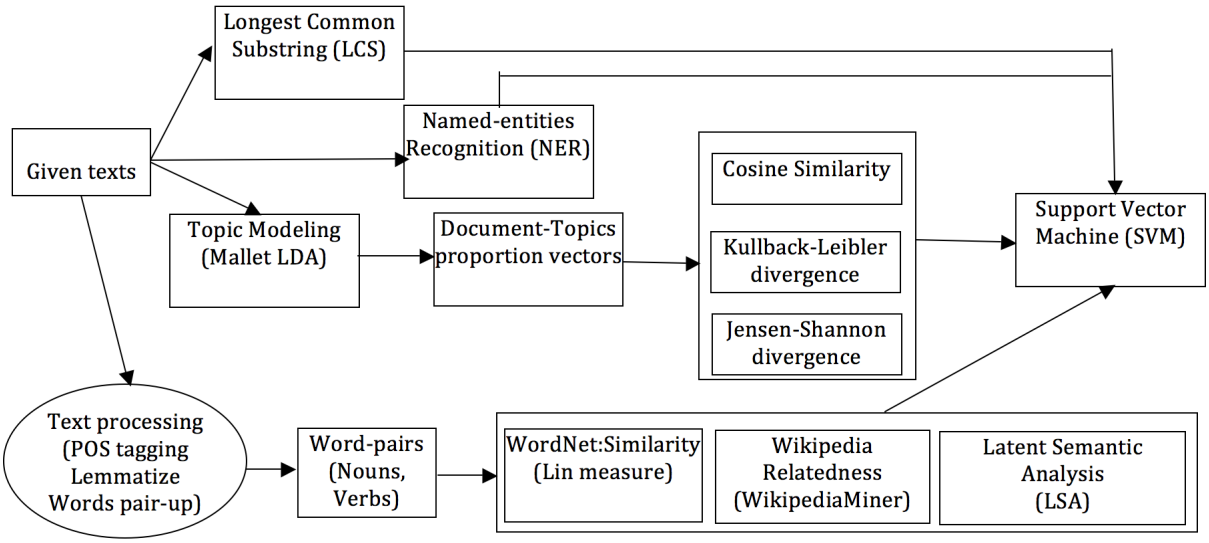

Figure 10: Distributional Module.

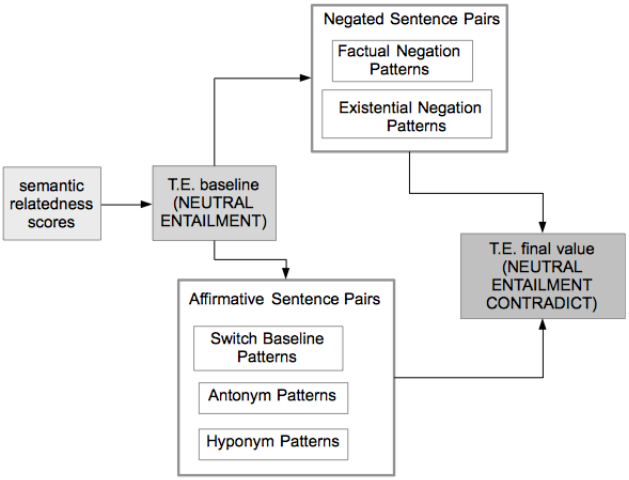

Figure 11: Structural Module.

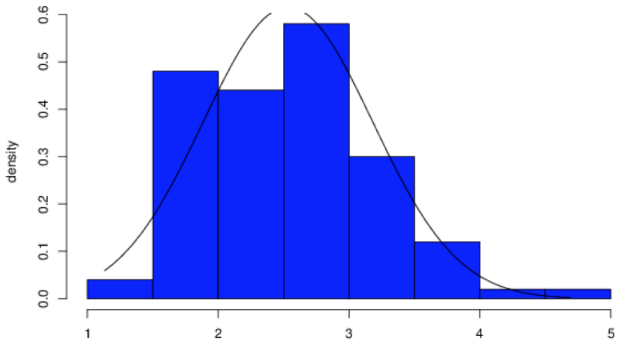

Figure 12: Initial SR scores for [2-3] Class.

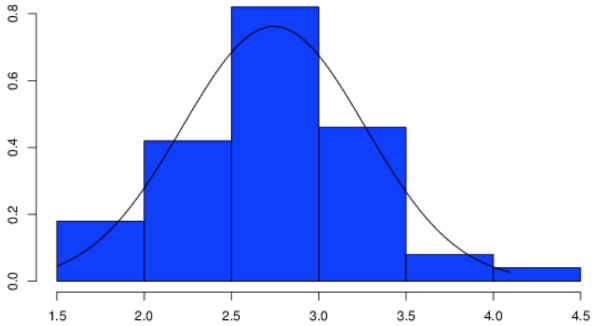

Figure 13: Re-computed SR scores for [2-3] Class.

of the interval [2-3] is due to the training data. This emphasizes the restriction that the train and the test should come from the same distribution, that is, same annotation guidelines, including how the partial contradictions should be judged.

6. Conclusion and Future Work

The aim in this paper was two-fold: (i) to introduce a new resource, and (ii) to analyze the relationship between SR

and TE. The corpus we created by annotating with SR scores the RTE 1-4 corpora, which was already annotated with TE judgments. The analysis carried out on the relationship between TE and SR proves beneficial for building better models for both tasks. In fact, it was revealed that there is a strong dependency on these two annotations.

Another less apparent fact that was revealed regards the dependency relationship between partial contradiction and similarity. This aspect needs to be clarified further.

By employing a powerful textual entailment analyzer which takes into account also the structure of the sentence, in a way that the entailment conditions are revealed a NLP systems can tackle a class of tasks that involve similarity decisions.

It would be interesting to observe the influence of multiple sentences in text on similarity. In order to analyze this relationship, the RTE 5-7 seems like a good starting point. We have planned to work on this for future work.

7. Bibliographical References

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., and Szpektor, I. (2006). The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4.

Cooper, R., Crouch, D., Van Eijck, J., Fox, C., Van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., et al. (1996). Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. Springer.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Giampiccolo, D., Dang, H. T., Magnini, B., Dagan, I., Cabrio, E., and Dolan, B. (2009). The fourth pascal recognizing textual entailment challenge. In *Proceedings of the First Text Analysis Conference (TAC 2008)*. Citeseer.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014a). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014, Reykjavik (Iceland): ELRA*.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014b). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *SemEval-2014*.

Popescu, O., Cabrio, E., and Magnini, B. (2011). Textual entailment using chain calrifying relationships. In *Proceedings of the IJCAI Workshop Learning by Reasoning and its Applications in Intelligent Question-Answering*.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Vo, N. P. A., Popescu, O., and Caselli, T. (2014). Fbk-tr: Svm for semantic relatedness and corpus patterns for rte. *SemEval 2014*, page 289.