DirRelCond3: Detecting Textual Entailment Across Languages With Conditions On Directional Text Relatedness Scores

Alpár Perini
'Babeş-Bólyai' University
Cluj-Napoca, Romania
palpar at gmail.com

Abstract

There are relatively few entailment heuristics that exploit the directional nature of the entailment relation. Cross-Lingual Text Entailment (CLTE), besides introducing the extra dimension of cross-linguality, also requires to determine the exact direction of the entailment relation, to provide content synchronization (Negri et al., 2012). Our system uses simple dictionary lookup combined with heuristic conditions to determine the possible directions of entailment between the two texts written in different languages. The key members of the conditions were derived from (Corley and Mihalcea, 2005) formula initially for text similarity, while the entailment condition used as a starting point was that from (Tatar et al., 2009). We show the results obtained by our implementation of this simple and fast approach at the CLTE task from the SemEval-2012 challenge.

1 Introduction

Recognizing textual entailment (TE) is a key task for many natural language processing (NLP) problems. It consists in determining if an entailment relation exists between two texts: the text (T) and the hypothesis (H). The notation $T \rightarrow H$ says that the meaning of H can be inferred from T, in order words, H does not introduce any novel information with respect to T.

Even though RTE challenges lead to many approaches for finding textual entailment, fewer authors exploited the directional character of the entailment relation. Due to the fact that the entailment relation, unlike the equivalence relation, is not symmetric, if $T \rightarrow H$, it is less likely that the reverse $H \rightarrow T$ can also hold (Tatar et al., 2009).

The novel Cross-Lingual Text Entailment (CLTE) approach increases the complexity of the traditional TE task in two way, both of which have been only partially researched and have promise for great potential (Negri et al., 2012):

- the two texts are no longer written in the same language (cross-linguality);
- the entailment needs to be queried in both directions (content synchronization).

Mehdad et al. (2010) presented initial research directions and experiments for the cross-lingual context and explored possible application scenarios.

2 Theoretical Background

The semantic similarity formula from (Corley and Mihalcea, 2005) defines the similarity of a pair of documents differently depending on with respect to which text it is computed. The formula involves only the set of open-class words (nouns, verbs, adjectives and adverbs) from each text.

Based on this text-to-text similarity metric, Tatar et al. (2009) have derived a textual entailment recognition system. The paper demonstrated that in the case when $T \rightarrow H$ holds, the following relation will take place:

$$sim(T,H)_H > sim(T,H)_T \qquad (1)$$

however, the opposite of this statement is not always true, nevertheless it is likely. In (Tatar et al., 2007)

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 710–714,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

a simpler version for the calculus of $sim(T,H)_T$ is used: namely the only case of similarity is the identity (a symmetric relation) and/or the occurrence of a word from a text in the synset of a word in the other text (not symmetric relation).

Perini and Tatar (2009) used the earlier semantic similarity formula (Corley and Mihalcea, 2005) to derive a formula for directional text relatedness score as follows:

$$rel(T,H)_T = \frac{\sum_{pos}\sum_{T_i \in WS^T_{pos}}(maxRel(T_i) \times idf(T_i))}{\sum_{pos}\sum_{T_i \in WS^T_{pos}} idf(T_i)} \quad (2)$$

A mathematically similar formula could be given for $rel(T,H)_H$ (by swapping T for H in the RHS of (2)) which would normally produce a different score. In (2), $maxRel(T_i)$ was defined as the highest *relatedness* between (in this order) word T_i and words from H having the same part of speech as T_i. The relatedness between a pair of words was computed by taking the weight of the highest-ranked WordNet relation that takes place between them. It should be noted that the word order in the pair was strict and that most of the WordNet relations involved in the calculus were not symmetric.

After defining the relatedness of two texts, which depends on the *direction*, Perini and Tatar (2009) introduced a new directional entailment condition, derived from the one in (Tatar et al., 2009):

$$rel(T,H)_T + \sigma > rel(T,H)_H > rel(T,H)_T > \theta \ . \quad (3)$$

3 The DirRelCond3 System

After having presented the necessary theoretical background, in this section we give an overview of our system for CLTE.

The application was implemented in the Java programming language. XML input and output was performed using the DocumentBuilder and the DOM parser from Java.

The first step was to tag both the English and the foreign language sentence using the TreeTagger (Schmid, 1995), which had the advantage that it was fast and it supported all the languages required by this task by providing it with the necessary parameter file, and also had a nice Java wrapper for it (annolab, 2011). The output of the tagger was used to obtain the necessary POS information needed to distinguish the set of open-class words for each sentence. Because the tagset used for each language was different, it was necessary to adapt all the different variants to the four generic classes: noun, verb, adjective and adverb.

The translation step followed for the foreign language sentence, which took words only from these classes and translated them using two dictionaries in some cases. The base dictionary used for word lookup was the FreeDict (FreeDictProject, 2012), for which it was possible to download the language files and use them locally with the help of a server (ktulu, 2006) and a Java client (SourceForge, 2001). The disadvantage of this dictionary was that it had rather few headwords mainly for the Italian and Spanish languages. A later improvement was to use an additional online dictionary as a fall-back, WordReference.com (WordReference.com, 2012), which had a very good headword count for the Italian and French languages, it also provided a very nice JSON API to access it and there was a ready-to-use Java API (SourceForge, 2011) for it that supported caching the results. Although the number of queries per hour was limited, it was very helpful that they approved the caching of the results for the duration of the development. The dictionary lookup process attached to each foreign word that was found the set of English meanings, corresponding to each sense that was found.

The penultimate step was to compute the text relatedness scores with respect to each sentence, $rel(T,H)_T$ and $rel(T,H)_H$, by applying (2). The only modification compared to the original formula was that in the case of the translated word, all the obtained meanings were used and the one producing the maximum relatedness was kept. We have used the following weights (assigned intuitively) for the different WordNet relations in the final *word relatedness score*:

- equals: 1.0;
- same synset: 0.9;
- similar to: 0.85

- hypernyms: 0.8;
- hyponyms: 0.7;
- entailment: 0.7;
- meronyms: 0.5;
- holonyms: 0.5;
- not in WordNet or dictionaries: 0.01.

The final step was to devise a condition based on these two text relatedness scores, similar to (3), but one that would be able to report the entailment vote for both directions:

$$\begin{cases} \text{noentail}, & \text{if } rel(T,H)_T \text{ or } rel(T,H)_H < \theta \\ \text{bidir}, & \text{if } abs(rel(T,H)_T, rel(T,H)_H) < \delta \\ \text{forward}, & \text{if } rel(T,H)_H > rel(T,H)_T + \sigma \\ \text{backwd}, & \text{otherwise} \end{cases} \tag{4}$$

4 Experimental Results

The CLTE task provided researchers with training sets of 500 sentence pairs (one English, one foreign) already annotated with the type of entailment that exists between them ('Forward', 'Backward', 'Bidirectional', 'No entailment'). There was one training set for each French-English, German-English, Italian-English, Spanish-English language combination (Negri et al., 2011). The test set consisted in a similarly structured 500 pairs for each language pair but without annotations. The mentioned entailment judgment types were uniformly distributed, both in the case of the development and the test dataset.

The DirRelCond3 system participated at the CLTE task with four runs for each of the above language combinations. Regarding the results, the accuracies obtained are summarized in table 1 as percentages.

Figures 1, 2, 3, 4 show the precision, recall and F-measure for the 'Forward', 'Backward', 'No entailment' and 'Bidirectional' judgments for each of the language pair combinations in the case of the best run that the DirRelCond3 system has obtained:

System	*Spa-En*	*Ita-En*	*Fra-En*	*Deu-En*
Run 1	30.0	28.0	36.2	33.6
Run 2	30.0	28.4	36.0	33.6
Run 3	30.0	*33.8*	*38.4*	36.4
Run 4	*34.4*	31.6	*38.4*	*37.4*

Table 1: DirRelCond3 accuracies obtained for CLTE task. Best results are with italic.

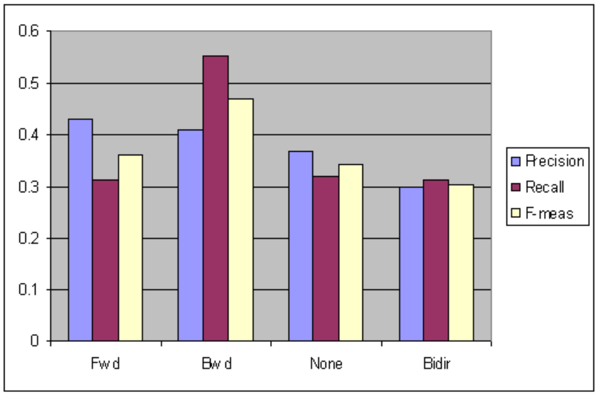

Figure 1: DirRelCond3 German-English pair precision, recall and F-measure values for the different judgments.

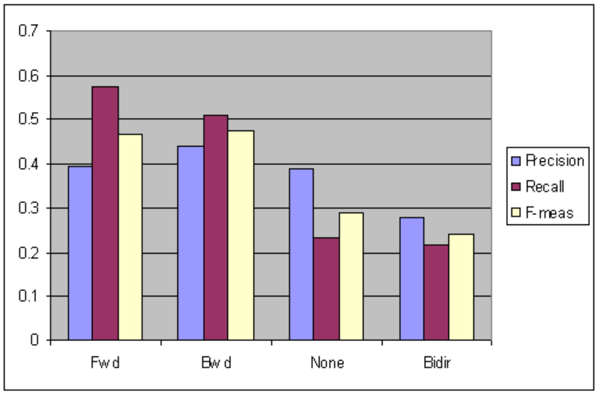

Figure 2: DirRelCond3 French-English pair precision, recall and F-measure values for the different judgments.

The earlier figures pointed out that generally the unidirectional 'Forward' and 'Backward' judgements produced better results than the remaining ones that involved bi-directionality. This is somewhat expected because in this case it is more difficult to correctly judge since there could more possibility for error.

Regarding the individual runs, run 2 added slightly improved dictionary search in addition to run 1, by attempting to look for the lemma form of the word as well, that was available thanks to the

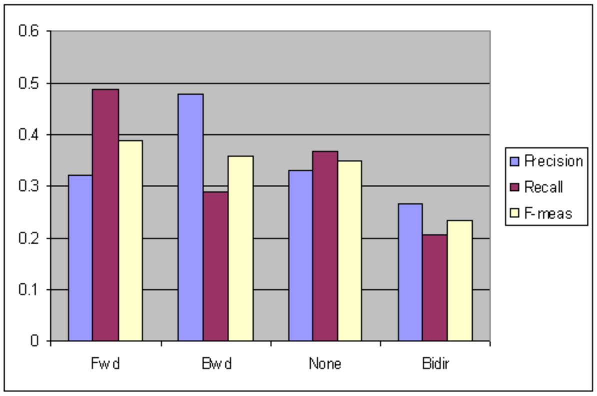

Figure 3: DirRelCond3 Italian-English pair precision, recall and F-measure values for the different judgments.

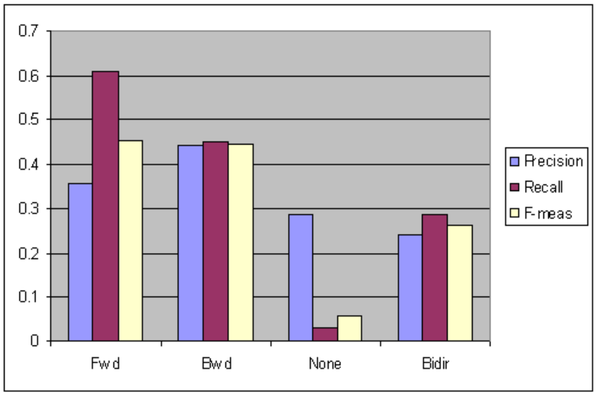

Figure 4: DirRelCond3 Spanish-English pair precision, recall and F-measure values of the different judgments.

TreeTagger tool (Schmid, 1995). In case the word was still not found, but the language was French or Italian and the word contained apostrophe, a lookup was attempted for the part following it.

Run 3 added another slight improvement for German, in case there was still no match for the word, tried to see if the word was a composite containing two parts found in the dictionary, and if so, used the first one.

The first two runs were only using the FreeDict (FreeDictProject, 2012) dictionary, while starting with run 3, Italian and French language words, in case not found, could also be searched in the WordReference (WordReference.com, 2012) online dictionary.

The first three runs were using entailment conditions common to all language combinations. The values of the parameters were chosen based on the CLTE development dataset (Negri et al., 2011) and were as follows:
$\theta = 0.5, \delta = 0.03, \sigma = 0.0.$
The final run used empirically-tuned conditions for each language pair in the dataset. The θ threshold needed to be lowered for Spanish since many words were not found in FreeDict, which was the only one we had available for use, so the relatedness scores were rather smaller. The values are summarized in table 2 below:

Param	*Spa-En*	*Ita-En*	*Fra-En*	*Deu-En*
θ	0.25	0.55	0.5	0.45
δ	0.03	0.025	0.03	0.04
σ	0.0	0.2	0.0	0.0

Table 2: DirRelCond3 – Run 4 condition parameters.

5 Conclusions and Future Work

In this paper we have presented the DirRelCond3 systems that participated at the CLTE task (Negri et al., 2012) from SemEval-2012. The system was a good example of how an approach for mono-lingual text entailment can be adapted to the new dimension of cross-linguality. It would have been possible to use a MT tool and then do the entailment detection steps all in English as was the original approach, however we expected that that would introduce more possibility for error than translating and comparing words with the same POS.

The overall best result for each language that we have obtained was around the median of all the system runs that were submitted to the CLTE task. The best accuracy obtained by our system was for the French-English pair with 38.4%, but well below the accuracy of the best systems. Generally the results involving German and French were somewhat better than the other two languages. In the case of Spanish this could easily be caused by the significantly smaller dictionary that was available, while for Italian, after relying also on WordReference.com this was no longer the case. A possiblity is that some language particularities were affecting the results (e.g. high usage of apostrophe) but perhaps the entailment heuristic thresholds were not the best either.

Finally, there are several possible improvements.

Firstly, in case the dictionary provides POS information for the translation, that could be used to retain only those senses that have the same POS as the original word. For some languages, particularly for Spanish, it would be helpful to rely on dictionaries with more headwords. Secondly, we can use the inverse document frequency counts for words, obtained either from the CLTE development corpus or from web searches, because currently that was simply one. Thirdly, both the empirically obtained conditions can be further tuned, manually or by means of learning, separately for each language pair. Fourthly, when computing the word relatedness scores, the weights of the WordNet relations could be further adjusted for each language, empirically, or again by learning.

References

annolab. 2011. tt4j – TreeTagger for Java. http://code.google.com/p/tt4j/.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In Ann Arbor, editor, *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18.

FreeDictProject. 2012. FreeDict – free bilingual dictionaries. http://www.freedict.org/en/.

ktulu. 2006. JavaDICT – Java DICT Client. http://ktulu.com.ar/blog/projects/javadictd/.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards cross-lingual textual entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 321–324, Los Angeles, California, June. Association for Computational Linguistics.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 670–679, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

M. Negri, A. Marchetti, Y. Mehdad, L. Bentivogli, and D. Giampiccolo. 2012. Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Alpar Perini and Doina Tatar. 2009. Textual entailment as a directional relation revisited. *Knowledge Engineering: Principles and Techniques*, pages 69–72.

Helmut Schmid. 1995. TreeTagger – a language independent part-of-speech tagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

SourceForge. 2001. JDictClient – JAVA dict server client. http://sourceforge.net/projects/jdictclient/.

SourceForge. 2011. WordReference Java API. http://sourceforge.net/projects/wordrefapi/.

Doina Tatar, Gabriela Serban, and M. Lupea. 2007. Text entailment verification with text similarities. In Babes-Bolyai University, editor, *Knowledge Engineering: Principles and Techniques*, pages 33–40. Cluj University Press.

Doina Tatar, Gabriela Serban, A. Mihis, and Rada Mihalcea. 2009. Textual entailment as a directional relation. *Journal of Research and Practice in Information Technology*, 41(1):17–28.

WordReference.com. 2012. WordReference.com – Online Language Dictionaries. http://www.wordreference.com/.