CELI: An Experiment with Cross Language Textual Entailment

Milen Kouylekov
Celi S.R.L.
via San Quintino 31
Torino, Italy
kouylekov@celi.it

Luca Dini
Celi S.R.L.
via San Quintino 31
Torino, Italy
dini@celi.it

Alessio Bosca
Celi S.R.L.
via San Quintino 31
Torino, Italy
bosca@celi.it

Marco Trevisan
Celi S.R.L.
via San Quintino 31
Torino, Italy
trevisan@celi.it

Abstract

This paper presents CELI's participation in the SemEval Cross-lingual Textual Entailment for Content Synchronization task.

1 Introduction

The Cross-Lingual Textual Entailment task (CLTE) is a new task that addresses textual entailment (TE) (Bentivogli et. al., 2011), targeting the cross-lingual content synchronization scenario proposed in (Mehdad et. al., 2011) and (Negri et. al., 2011). The task has interesting application scenarios that can be investigated. Some of them are content synchronization and cross language query alignment. The task is defined by the organizers as follows: Given a pair of topically related text fragments (T1 and T2) in different languages, the CLTE task consists of automatically annotating it with one of the following entailment judgments:

- Bidirectional: the two fragments entail each other (semantic equivalence)
- Forward: unidirectional entailment from T1 to T2
- Backward: unidirectional entailment from T2 to T1
- No Entailment: there is no entailment between T1 and T2

In this task, both *T1* and *T2* are assumed to be TRUE statements; hence in the dataset there are no contradictory pairs.

Example for Spanish English pairs:

- **bidirectional**
  Mozart naci en la ciudad de Salzburgo
  Mozart was born in Salzburg.
- **forward**
  Mozart naci en la ciudad de Salzburgo
  Mozart was born on the 27th January 1756 in Salzburg.
- **backward**
  Mozart naci el 27 de enero de 1756 en Salzburgo
  Mozart was born in 1756 in the city of Salzburg
- **no_entailment**
  Mozart naci el 27 de enero de 1756 en Salzburgo
  Mozart was born to Leopold and Anna Maria Pertl Mozart.

2 Our Approach to CLTE

In our participation in the 2012 SemEval Cross-lingual Textual Entailment for Content Synchronization task (Negri et. al., 2012) we have developed an approach based on cross-language text similarity. We have modified our cross-language query similarity system TLike to handle longer texts.

Our approach is based on four main resources:

- A system for Natural Language Processing able to perform for each relevant language basic tasks such as part of speech disambiguation, lemmatization and named entity recognition.
- A set of word based bilingual translation modules.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 696–700,
Montréal, Canada, June 7-8, 2012.

- A semantic component able to associate a semantic vectorial representation to words.

- We use Wikipedia as multilingual corpus.

NLP modules are described in (Bosca and Dini, 2008), and will be no further detailed here.

Word-based translation modules are composed by a bilingual lexicon look-up component coupled with a vector based translation filter, such as the one described in (Curtoni and Dini, 2008). In the context of the present experiments, such a filters has been deactivated, which means that for any input word the component will return the set of all possible translations. For unavailable pairs, we make use of triangular translation (Kraaij, 2003).

As for the semantic component we experimented with a corpus-based distributional approach capable of detecting the interrelation between different terms in a corpus; the strategy we adopted is similar to Latent Semantic Analysis (Deerwester et. al., 1990) although it uses a less expensive computational solution based on the Random Projection algorithm (Lin et. al., 2003) and (Bingham et. al., 2001). Different works debate on similar issues: (Turney, 2001) uses LSA in order to solve synonymy detection questions from the well-known TOEFL test while the method presented by (Inkpen, 2001) or by (Baroni and Bisi, 2001) proposes the use of the Web as a corpus to compute mutual information scores between candidate terms.

More technically, Random Indexing exploits an algebraic model in order to represent the semantics of terms in a Nth dimensional space (a vector of length N); approaches falling into this category, actually create a Terms By Contexts matrix where each cell represents the degree of memberships of a given term to the different contexts. The algorithm assigns a random signature to each context (a highly sparse vector of length N, with few, randomly chosen, non-zero elements) and then generates the vector space model by performing a statistical analysis of the documents in the domain corpus and by accumulating on terms rows all the signatures of the contexts where terms appear.

According to this approach if two different terms have a similar meaning they should appear in similar contexts (within the same documents or surrounded by the same words), resulting into close coordinates in the so generated semantic space.

In our case study semantic vectors have been generated taking as corpus the set of metadata available via the CACAO project (Cacao Project, 2007) federation (about 6 millions records). After processing for each word in the corpus we have:

- A vector of float from 0 to 1 representing its contextual meaning;

- A set of neighbors terms selected among the terms with a higher semantic similarity, calculated as cosine distance among vectors.

We use Wikipedia as a corpus for calculating word statistics in different languages. We have indexed using Lucene[1] the English, Italian, French, German, Spanish distributions of the resource.

The basic idea behind our algorithm is to detect the probability for two texts to be one a translation of the other. In the simple case we expect that if all the words in text TS have a translation in text TT and if TS and TT have the same number of terms, then T_S and T_T are entailed. Things are of course more complex than this, due to the following facts:

- The presence of compound words make the constraints on cardinality of search terms not feasible (e.g. the Italian Carta di Credito vs. the German KreditCarte).

- One or more words in T_S could be absent from translation dictionaries.

- One or more words in T_S could be present in the translation dictionaries, but contextually correct translation might be missing.

- There might be items which do not need to be translated, notably Named Entities.

The first point, compounding, is only partially an obstacle. NLP technology developed during CACAO Project, which adopted translation dictionaries, deals with compound words both in terms of identification and translation. Thus the Italian *"Carta di Credito"* would be recognized and correctly translated into *"KreditCarte"*. So, in an ideal

[1] *http://lucene.apache.org*

word, the cardinality principle could be considered strict. In reality, however, there are many compounding phenomena which are not covered by our dictionaries, and this forces us to consider that a mismatch in text term cardinality decrease the probability that the two translations are translation of each other, without necessarily setting it to zero.

Concerning the second aspect, the absence of source (T1) words in translation dictionaries, it is dealt with by accessing the semantic repository described in the previous section. We first obtain the list of neighbor terms for the *untranslatable* source word. This list is likely to contain many words that have one or more translations. For each translation, again, we consult our semantic repository and we obtain its semantic vector.

Finally, we compose all vectors of all available translations and we search in the target text (T2) for the word whose semantic vector best matches the composed one (cosine distance). Of course we cannot assume that the best matching vector is a translation of the original word, but we can use the distance between the two vectors as a further weight for deciding whether the two texts are translations one of the other.

There are of course cases when the source word is correctly missing in the source dictionary. This is typically the case of most named entities, such as geographical and person names. These entities should be appropriately recognized and searched as exact matches in the target text, thus by-passing any dictionary look-up and any semantic based matching. Notice that the recognition of named entities it is not just a matter of generalizing the statement according to which *"if something is not in the dictionaries, it is a named entity"*. Indeed there are well known cases where the named entity is homograph with common words (e.g. the French author *"La Fontaine"*), and in these cases we must detect them in order to avoid the rejection of likely translation pairs. In other words we must avoid that the two texts *"La fontaine fables"* and *"La Fontaine favole"* are rejected as translation pairs, just by virtue of the fact that *"La fontaine"* is treated as a common word, thus generating the Italian translation *"La fontana"*. Fortunately CACAO disposes of a quite accurate subsystem for recognizing named entities in texts, mixing standard NLP technologies with statistical processing and other corpus-oriented heuristics.

We concentrated our work on handling cases where two texts are candidates to be mutual translations, but one or more words receive a translation which is not contained in the target text. Typically these cases are a symptom of a non-optimal quality in translation dictionaries: the lexicographer probably did not consider some translation candidate. To address this problem we have created a solution based on a weighting scheme. For each word of the source language we assign a weight that reflects its importance to the semantic interpretation of the text. We define a $match_{weight}$ of a word using the formula represented in Figure 2.In this formula wi_s is a word from the source text, wk_t is a word from the target text, w is a word in the source language and trans is a boolean function that searches in the dictionary for translations between two words.

The $match_{weight}$ is relevant to the matching of a translation of a word from the source with one of the words of the target. If the system finds a direct correspondence the weight is 0. If the match was made using random indexing the weight is inverse to the cosine similarity between the vectors.

In order to make an approximation of the significance of the word to the meaning of the phrase we have used as its cost the inverse document frequency (IDF) of the word calculated using Wikipedia as a corpus. IDF is a most popular measure (a measure commonly used in Information Retrieval) for calculating the importance of a word to a text. If N is the number of documents in a text collection and N_w is the number of documents of the collection that contain w then the IDF of w is given by the formula:

$$weight(wi_s) = idf(w) = log(\frac{N}{N_w}) \qquad (2)$$

Using the $match_{weight}$ and $weight$ we define the $match_{score}$ of a source target pair as:

$$match_{score}(T_s, T_t) = \frac{\sum match_{weigth}(wi_s)}{\sum weight(wi_s)} \qquad (3)$$

If all the words of the source text have a translation in the target text the score is 0. If none is found the score is 1. We have calculated the scores for each

$$match_{weight}(wi_s) = \begin{cases} 0 & \exists wk_t\ trans(wi_s) = wk_t \\ w * (wi_s) * (1-d) & \exists w\ \&wk_t\ distance(wi_s, w) = d \& trans(w) = wk_t \\ w * (wi_s) & otherwise \end{cases} \quad (1)$$

Figure 1: Match Weight of a Word

pair taking t1 as a source and t2 as a target and vice versa.

3 Systems

We have submitted **four** runs in the SemEval CLTE challenge. We used the NaiveBayse algorithm implemented in Mallet[2] to create a classifier that will produce the output for each of the four categories Forward , Backward , Bidirectional and No Entailment.

System 1 As our first system we have created a binary classifier in the classical RTE (Bentivogli et. al., 2011) classification (YES & NO) for each direction Forward and Backward. We assigned the Bidirectional category if both classifiers returned YES. As features the classifiers used only the match scores obtained for the corresponding direction as one and only numeric feature.

System 2 For the second system we trained a classifier using all *four* categories as output. Apart of the scores obtained matching the texts in both directions we have included also a set of eight simple surface measures. Some of these are:

- The length of the two texts.
- The number of common words without translations.
- The cosine similarity between the tokens of the two texts without translation.
- Levenshtein distance between the texts.

System 3 For the third system we trained a classifier using all *four* categories as output. We used as features scores obtained matching the texts in both directions without the surface features used in the System 2.

[2] *http://mallet.cs.umass.edu/*

System 4 In the last system we trained a classifier using all *four* categories as output. We used as features the simple surface measures used in System 2.

The results obtained are shown in Table 1.

4 Analysis

Analyzing the results of our participation we have reached several important conclusions.

The dataset provided by the organizers presented a significant challenge for our system which was adapted from a query similarity approach. The results obtained demonstrate that only a similarity based approach will not provide good results for this task. This fact is also confirmed by the poor performance of the simple similarity measures by themselves (System 4) and by their contribution to the combined run (System 2).

The poor performance of our system can be partially explained also by the small dimensions of the cross-language dictionaries we used. Expanding them with more words and phrases can potentially increase our results.

The classifier with *four* categories clearly outperforms the two directional one (System 1 vs. System 3).

Overall we are not satisfied with our experiment. A radically new approach is needed to address the problem of Cross-Language Textual Entailment, which our similarity based system could not model correctly.

In the future we intend to integrate our approach in our RTE open source system EDITS (Kouylekov et. al., 2011) (Kouylekov and Negri, 2010) available at *http://edits.sf.net*.

Acknowledgments

This work has been partially supported by the ECfunded project Galateas (CIP-ICT PSP-2009-3-250430).

	SPA-ENG	ITA-ENG	FRA-ENG	DEU-ENG
System 1	0.276	0.278	0.278	0.280
System 2	**0.336**	**0.336**	**0.300**	**0.352**
System 3	0.322	0.334	0.298	0.350
System 4	0.268	0.280	0.280	0.274

Table 1: Results obtained.

References

Baroni M., Bisi S. 2004. Using cooccurrence statistics and the web to discover synonyms in technical language In Proceedings of LREC 2004

Bentivogli L., Clark P., Dagan I., Dang H, Giampiccolo D. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge In Proceedings of TAC 2011

Bingham E., Mannila H. 2001. Random projection in dimensionality reduction: Applications to image and text data. In Knowledge Discovery and Data Mining, ACM Press pages 245250

Bosca A., Dini L. 2008. Query expansion via library classification system. In CLEF 2008. Springer Verlag, LNCS

Cacao Project CACAO - project supported by the eContentplus Programme of the European Commission. *http://www.cacaoproject.eu/*

Curtoni P., Dini L. 2006. Celi participation at clef 2006 Cross language delegated search. In CLEF2006 Working notes.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41 391407

Inkpen D. 2007. A statistical model for near-synonym choice. ACM Trans. Speech Language Processing 4(1)

Kraaij W. 2003. Exploring transitive translation methods. In Vries, A.P.D., ed.: Proceedings of DIR 2003.

Kouylekov M., Negri M. An Open-Source Package for Recognizing Textual Entailment. 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010) ,Uppsala, Sweden. July 11-16, 2010

Kouylekov M., Bosca A., Dini L. 2011. EDITS 3.0 at RTE-7. Proceedings of the Seventh Recognizing Textual Entailment Challenge (2011).

Lin J., Gunopulos D. 2003. Dimensionality reduction by random projection and latent semantic indexing. In proceedings of the Text Mining Workshop, at the 3rd SIAM International Conference on Data Mining.

Mehdad Y.,Negri M., Federico M.. 2011. Using Parallel Corpora for Cross-lingual Textual Entailment. In Proceedings of ACL-HLT 2011.

Negri M., Bentivogli L., Mehdad Y., Giampiccolo D., Marchetti A. 2011. Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. In Proceedings of EMNLP 2011.

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

Turney P.D. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In EMCL 01: Proceedings of the 12th European Conference on Machine Learning, London, UK, Springer-Verlag pages 491502