HDU: Cross-lingual Textual Entailment with SMT Features

Katharina Wäschle and **Sascha Fendrich**
Department of Computational Linguistics
Heidelberg University
Heidelberg, Germany
{waeschle, fendrich}@cl.uni-heidelberg.de

Abstract

We describe the Heidelberg University system for the Cross-lingual Textual Entailment task at SemEval-2012. The system relies on features extracted with statistical machine translation methods and tools, combining monolingual and cross-lingual word alignments as well as standard textual entailment distance and bag-of-words features in a statistical learning framework. We learn separate binary classifiers for each entailment direction and combine them to obtain four entailment relations. Our system yielded the best overall score for three out of four language pairs.

1 Introduction

Cross-lingual textual entailment (CLTE) (Mehdad et al., 2010) is an extension of textual entailment (TE) (Dagan and Glickman, 2004). The task of recognizing entailment is to determine whether a hypothesis H can be semantically inferred from a text T. The CLTE task adds a cross-lingual dimension to the problem by considering sentence pairs, where T and H are in different languages. The SemEval-2012 CLTE task (Negri et al., 2012) asks participants to judge entailment pairs in four language combinations[1], defining four target entailment relations, *forward*, *backward*, *bidirectional* and *no entailment*.

We investigate this problem in a statistical learning framework, which allows us to combine cross-lingual word alignment features as well as common monolingual entailment metrics, such as bag-of-words overlap, edit distance and monolingual alignments on translations of T and H, using standard statistical machine translation (SMT) tools and resources. Our goal is to address this task without deep processing components to make it easily portable across languages. We argue that the cross-lingual entailment task can benefit from direct alignments between T and H, since a large amount of bilingual parallel data is available, which naturally models synonymy and paraphrasing across languages.

2 Related Work

With the yearly Recognizing Textual Entailment (RTE) challenge (Dagan et al., 2006), there has been a lot of work on monolingual TE. We therefore include established monolingual features in our approach, such as alignment scores (MacCartney et al., 2008), edit distance and bag-of-words lexical overlap measures (Kouylekov and Negri, 2010). So far, the only work on CLTE that we are aware of is Mehdad et al. (2010), where the problem is reduced to monolingual entailment using machine translation, and Mehdad et al. (2011), which exploits parallel corpora for generating features based on phrase alignments as input to an SVM. Our approach combines ideas from both, mostly resembling Mehdad et al. (2011). There are, however, several differences; we use word translation probabilities instead of phrase tables and model monolingual and cross-lingual alignment separately. We also include additional similarity measures derived from the MT evaluation metric Meteor, which was used in Volokh and Neumann (2011) for the monolingual TE task. Con-

[1] Spanish-English (**es-en**), Italian-English (**it-en**), French-English (**fr-en**) and German-English (**de-en**).

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 467–471,
Montréal, Canada, June 7-8, 2012.

versely, Padó et al. (2009) showed that textual entailment features can be used for measuring MT quality, indicating a strong relatedness of the two problems.

The CLTE task is also related to the problem of identifying parallel sentence pairs in a non-parallel corpus, so we adapt alignment-based features from Munteanu and Marcu (2005), where a Maximum Entropy classifier was used to judge if two sentences are sufficiently parallel.

Regarding the view on entailment, MacCartney and Manning (2007) proposed the decomposition of top-level entailment, such as *equivalence* (which corresponds to the CLTE *bidirectional* class), into atomic forward and backward entailment predictions, which is mirrored in our multi-label approach with two binary classifiers.

3 SMT Features for CLTE

The SemEval-2012 CLTE task emerges from the monolingual RTE task; however the perception of entailment differs slightly. In CLTE, the sentences T_1 and T_2 are of roughly the same length and the entailment is predicted in both directions. Negri et al. (2011) states that the CLTE pairs were created by paraphrasing an English sentence E and leaving out or adding information to construct a modified sentence E', which was then translated into a different language[2], yielding sentence F and thus creating a bilingual entailment pair. For this reason, we believe that our system should be less inference-oriented than some previous RTE systems and rather should capture

- paraphrases and synonymy to identify semantic equivalence,
- phrases that have no matching correspondent in the other sentence, indicating missing (respectively, additional) information.

To this end, we define a number of similarity metrics based on different views on the data pairs, which we combine as features in a statistical learning framework. Our features are both cross- and monolingual. We obtain monolingual pairs by translating the English sentence E into the foreign language, yielding $T(E)$ and vice versa $T(F)$ from F, using Google Translate[3].

3.1 Token ratio features

A first indicator for additional or missing information are simple token ratio features, i.e. the fraction of the number of tokens in T_1 and T_2. We define three token ratio measures:

- English-to-Foreign, $\frac{|E|}{|F|}$
- English-to-English-Translation, $\frac{|E|}{|T(F)|}$
- Foreign-to-Foreign-Translation, $\frac{|T(E)|}{|F|}$

3.2 Bag-of-words and distance features

Typical similarity measures used in monolingual TE are lexical overlap metrics, computed on bag-of-words representations of both sentences. We use the following similarities, computing both $\mathrm{sim}(E, T(F))$ and $\mathrm{sim}(F, T(E))$.

- Jaccard coefficient, $\mathrm{sim}(A, B) = \frac{|A \cap B|}{|A \cup B|}$
- Overlap coefficient, $\mathrm{sim}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$

We also compute the lexical overlap on bigrams and trigrams.

In addition, we include a simple distance measure based on string edit distance ed, summing up over all distances between every token a in A and its most similar token b in B, where we assume that the corresponding token is the one with the smallest edit distance:

- $\mathrm{dist}(A, B) = \log \sum_{a \in A} \min_{b \in B} \mathrm{ed}(a, b)$

3.3 Meteor features

The Meteor scoring tool (Denkowski and Lavie, 2011) for evaluating the output of statistical machine translation systems can be used to calculate the similarity of two sentences in the same language. Meteor uses stemming, paraphrase tables and synonym collections to align words between the two sentences and scores the resulting alignment. We include the overall weighted Meteor score both for $(E, T(F))$

[2] We refer to the non-English language sentence as F.

[3] http://translate.google.com/

and $(F, T(E))$[4] as well as separate alignment precision, recall and fragmentation scores for $(E, T(F))$.

3.4 Monolingual alignment features

We use the alignments output by the Meteor-1.3 scorer for $(E, T(F))$[5] to calculate the following metrics:

- number of unaligned words
- percentage of aligned words
- length of the longest unaligned subsequence

3.5 Cross-lingual alignment features

We calculate IBM model 1 word alignments (Brown et al., 1993) with GIZA++ (Och and Ney, 2003) on a data set concatenated from Europarl-v6[6] (Koehn, 2005) and a bilingual dictionary obtained from dict.cc[7] for coverage. We then heuristically align each word e in E with the word f in F for which we find the highest word translation probability $p(e|f)$ and vice versa. Words for which no translation is found are considered unaligned. From this alignment a, we derive the following features both for E and F (resulting in a total of eight cross-lingual alignment features):

- number of unaligned words
- percentage of aligned words
- alignment score $\frac{1}{|E|} \sum_{e \in E} p(e|a(e))$
- length of the longest unaligned subsequence

4 Classification

To account for the different data ranges, we normalized all feature value distributions to the normal distribution $\mathcal{N}(0, \frac{1}{3})$, so that 99% of the feature values are in $[-1, 1]$. We employed SVM^{light} (Joachims, 1999) for learning different classifiers to output the four entailment classes. We submitted a second run to evaluate our recently implemented stochastic learning toolkit Sol (Fendrich, 2012), which implements binary, multi-class, and multi-label classification.

$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	**entailment**
1	1	*bidirectional*
1	0	*forward*
0	1	*backward*
0	0	*no entailment*

Table 1: Combination of atomic entailment relations.

For development, we split the training set in two parts, which were alternatingly used as training and test set. We first experimented with a multi-class classifier that learned all four entailment classes at once. However, although the task defines four target entailment relations, those can be broken down into two atomic relations, namely directional entailment from T_1 to T_2 and from T_2 to T_1 (table 1). We therefore learned a binary classifier for each atomic entailment relation and combined the output to obtain the final entailment class. We found this view to be a much better fit for the problem, improving the accuracy score on the development set by more than 10 percentage points (table 2). This two-classifiers approach can also be seen as a variant of multi-label learning, with the two atomic entailment relations as labels. We therefore also trained a direct implementation of multi-label classification. Although it substantially outperformed the multi-class approach, the system yielded considerably lower scores than the version using two binary classifiers.

5 Results

The accuracy scores of our two runs on the SemEval-2012 CLTE test set are presented in table 3. Our system performed best out of ten systems for the language pairs **es-en** and **de-en** and tied in first place for **fr-en**. For **it-en**, our system came in second. Regarding the choice of the learner, our toolkit slightly outperformed SVM^{light} on three of the four language pairs.

To determine the contribution of different feature types for each language combination, we performed ablation tests on the development set, where we switched off groups of features and measured the

[4] Meteor-1.3 supports English, Spanish, French and German. We used the Spanish version for scoring Italian, since those languages are related.

[5] Since the synonymy module is only available for English, we do not use the alignment of $(F, T(E))$.

[6] http://www.statmt.org/europarl/

[7] http://www.dict.cc/

	es-en	it-en	fr-en	de-en
multi-class	0.47	0.456	0.466	0.458
multi-label	0.586	0.526	0.568	0.522
2× binary	0.646	0.614	0.628	0.588

Table 2: Different classifiers on development set.

	es-en	it-en	fr-en	de-en
SVM^{light}	0.630	0.554	0.564	0.558
Sol	0.632	0.562	0.570	0.552

Table 3: Results on test set.

impact on the accuracy score (table 4). We assessed the statistical significance of differences in score with an approximate randomization test[8] (Noreen, 1989), indicating a significant impact in **bold font**. The results show that only in two cases a single feature group significantly impacts the score, namely the Meteor score features for **es-en** and the cross-lingual alignment features for **de-en**. However, no feature group hurts the score either, since negative variations in score are not significant. To ensure that the different feature groups actually express diverse information, we also evaluated our system using only one group of features at a time. The results confirm the most significant feature type for each language pair, but even the best-scoring feature group for each pair always yielded scores 3-6 percentage points lower than the system with all feature groups combined. We therefore conclude that the combination of diverse features is one key aspect of our system.

[8]Using a significance level of 0.05.

6 Conclusions

We have shown that SMT methods can be profitably applied for the problem of CLTE and that combining different feature types improves accuracy. Key to our approach is furthermore the view of the four-class entailment problem as a bidirectional binary or multi-label problem. A possible explanation for the superior performance of the multi-label approach is that the overlap of the bidirectional entailment with forward and backward entailment might confuse the multi-class learner.

Regarding future work, we think that our results can be improved by building on better alignments, i.e. using more data for estimating cross-lingual alignments and larger paraphrase tables. Furthermore, we would like to investigate more thoroughly in what way the representation of the problem in terms of machine learning impacts the system performance on the task – in particular, why the two-classifiers approach substantially outperforms the multi-label implementation.

References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability.

Ido Dagan, Oren Glickman, and Bernado Magnini. 2006. The pascal recognising textual entailment challenge. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3:

	es-en		it-en		fr-en		de-en	
feature group (#/features)	score	impact	score	impact	score	impact	score	impact
Meteor scores (5)	0.616	**0.03**	0.6	0.014	0.618	0.01	0.59	-0.002
distance/bow (10)	0.644	0.002	0.608	0.006	0.62	0.008	0.596	-0.008
token ratio (3)	0.652	-0.006	0.606	0.008	0.62	0.008	0.588	0
cross-lingual alignment (8)	0.638	0.008	0.592	0.022	0.62	0.008	0.526	**0.062**
monolingual alignment (3)	0.648	-0.002	0.624	-0.01	0.59	0.038	0.596	-0.008
all (29)	0.646		0.614		0.628		0.588	

Table 4: Ablation tests on development set.

Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Sascha Fendrich. 2012. Sol – Stochastic Learning Toolkit. Technical report, Department of Computational Linguistics, Heidelberg University.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. *Advances in Kernel Methods Support Vector Learning*, pages 169–184.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Milen Kouylekov and Matteo Negri. 2010. An open-source package for recognizing textual entailment. In *Proceedings of the ACL 2010 System Demonstrations*, pages 42–47.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 802–811.

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.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using bilingual parallel corpora for cross-lingual textual entailment. *Proceedings of ACL-HLT*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

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 Conference on Empirical Methods in Natural Language Processing*, pages 670–679.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo 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)*.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction.* Wiley, New York.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Sebastian Padó, Daniel Cer, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation*, 23(2):181–193.

Alexander Volokh and Günter Neumann. 2011. Using MT-based metrics for RTE. In *The Fourth Text Analysis Conference*. NIST.