Umelb: Cross-lingual Textual Entailment with Word Alignment and String Similarity Features

Yvette Graham **Bahar Salehi** **Timothy Baldwin**
Department of Computing and Information Systems
The University of Melbourne
{ygraham,bsalehi,tbaldwin}@unimelb.edu.au

Abstract

This paper describes The University of Melbourne NLP group submission to the Cross-lingual Textual Entailment shared task, our first tentative attempt at the task. The approach involves using parallel corpora and automatic word alignment to align text fragment pairs, and statistics based on unaligned words as features to classify items as forward and backward before a compositional combination into the final four classes, as well as experiments with additional string similarity features.

1 Introduction

Cross-lingual Textual Entailment (CLTE) (Negri et al., 2012) proposes the task of automatically identifying the kind of relation that exists between pairs of semantically-related text fragments written in two distinct languages, a variant of the traditional Recognizing Textual Entailment (RTE) task (Bentivogli et al., 2009; Bentivogli et al., 2010). The task targets the cross-lingual content synchronization scenario proposed in Mehdad et al. (2010, 2011). Compositional classification can be used by training two distinct binary classifiers for forward and backward entailment classification, before combining labels into the four final entailment categories that now include bidirectional and no_entailment labels. The most similar previous work to this work is the cross-lingual approach of the FBK system (Mehdad et al., 2012) from Semeval 2012 (Negri et al., 2012), in which the entailment classification is obtained without translating T1 into T2 for the Spanish–English language pair. We apply the cross-lingual approach to German–English and instead of cross-lingual matching features, we use Giza++ (Och et al., 1999) and Moses (Koehn et al., 2007) to automatically word align text fragment pairs to compute statistics of unaligned words. In addition, we include some additional experiments using string similarity features.

2 Compositional Classification

Given a pair of topically related fragments, T1 (German) and T2 (English), we automatically annotate it with one of the following entailment labels: bidirectional, forward, backward, no_entailment. We take the compositional approach and separately train a forward, as well as a backward binary classifier. Each classifier is run separately on the set of text fragment pairs to produce two binary labels for forward and backward entailment. The two sets of labels are logically combined to produce a final classification for each test pair of forward, backward, bidirectional or no_entailment.

3 Word Alignment Features

The test set of topically-related text fragments, T1 (German) and T2 (English) were added to Europarl German–English parallel text (Koehn, 2005) and Giza++ was used for automatic word alignment in both language directions. Moses (Koehn et al., 2007) was then used for symmetrization with the *grow_diag_final_and* algorithm. This produces a many-to-many alignment between the words of the

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 133–137, Atlanta, Georgia, June 14-15, 2013.

German, T1, and English, T2, with words also remaining unaligned.

The following features are computed for each test pair feature scores for the *forward* classifier:

- A1: count of unaligned words in T2
- A2: count of words comprised soley of digits in T2 not in T1
- A3: count of unaligned words in T2 with low probability of appearing unaligned in Europarl (with threshold p=0.11)

The number of words in T2 (English) that are not aligned with anything in T1 (German) should provide an indication that, for example, the English text fragment contains information not present in the corresponding German text fragment and subsequently evidence against the presence of forward entailment. We there include the feature, A1, that is simply a count of unaligned words in English T2. In addition, we hypothesize that the absence of a number from T2 may be a more significant missing element of T2 from T1. We therefore include as a feature the count of tokens comprised of digits in T2 that are not also present in T1. The final word alignment feature attempts to refine A1, by distinguishing words that are rarely unaligned in German–English translations. Statistics are computed for every lexical item from German–English Europarl translations to produce a lexical unalignment probability, computed for each lexical item based on its relative frequency in the corpus when it is not aligned to any other word.

The *backward* classifier uses the same features but computed for each test pair on counts of unaligned T1 words.

4 Results

Results for several combinations of features are shown in Table 1 when the system is trained on the 500-pair development set training corpus and tested on the 500-pair held-out development test set (DEV), in addition to results for feature combinations when trained on the entire 1000-pair development data and tested on the held-out 500-pair gold standard (TEST) (Negri et al., 2011), when the system is evaluated as two separate binary *forward* and *backward* classifiers (2-CLASS) as well as the final evaluation including all four entailment classes (4-CLASS). The highest accuracy is achieved by the classifier using the single feature of counts of unaligned words, A1, of 34.6%. As two separate binary classifiers, the alignment features, A1+A2+A3, achieve a relatively high accuracy of 74.0% for forward with somewhat less accurate for backward (65.8%) classification (both over the DEV data). When combined to the final four CLTE classes, however, accuracy drops significantly to an overall accuracy of 50% (also over DEV). A main cause is inaccurate labeling of no_entailment gold standard test pairs, as the most severe decline is for recall of test pairs for this label (38.4%).

Accuracy on the development set for the word alignment features, A1+A2+A3, compared to the test set shows a sever decline, from 50% to 32%. On the test data, however, a main cause of inaccuracy is that backward gold standard test pairs, although achieving close accuracy to forward when evaluated as binary classifiers, are inaccurately labeled in the 4-class evaluation, as recall for backward drops to only 18.4% for this label.

Another insight revealed for the alignment features, A1+A2+A3, in the 4-class evaluation is that when run on the development set, the classes forward and backward achieve significantly higher f-scores compared to no_entailment. However, the contrary is observed for the test data, as no_entailment achieve higher results than both unidirectional classes. This appears at first to be a somewhat counter-intuitive result, but in this case, the system is simply better at predicting forward and backward when no entailment exists for a translation pair compared to when a unidirectional entailment is present.

4.1 String Similarity Features

In addition to the word alignment features, subsequent to submitting results to the shared task, we have carried out additional experiments using string similarity features, based on our recent success in apply string similarity to both the estimation of compositionality of MWEs (Salehi and Cook, to appear) and also the estimation of similarity between short

		2-CLASS					4-CLASS			
			Acc.	Prec	Recall	F1	Acc.		Prec	Recall	F1
DEV	A1 + A2 + A3	bwrd	**65.80**	**63.12**	**76.00**	**68.96**	**50.00**	bwrd	**54.80**	**59.20**	**56.90**
		fwrd	**74.00**	**72.22**	**78.00**	**75.00**		fwrd	**54.80**	**45.60**	**49.80**
								none	**50.50**	38.40	**43.60**
								bidir	**42.80**	**56.80**	**48.80**
	S1 + S2 + S3	bwrd	58.20	57.75	61.20	59.42	27.40	bwrd	14.30	0.80	1.50
		fwrd	47.00	47.17	50.00	59.42		fwrd	0.00	0.00	0.00
								none	30.70	**39.70**	39.70
								bidir	25.60	52.80	34.50
TEST	A1	bwrd	57.00	**58.54**	48.00	52.75	**34.60**	bwrd	25.50	19.20	21.90
		fwrd	58.40	58.75	56.40	57.55		fwrd	**34.90**	**36.00**	**35.40**
								none	36.70	48.80	**41.90**
								bidir	**38.70**	34.40	36.40
	A2	bwrd	50.00	0.00	0.00	0.00	33.60	bwrd	24.70	18.40	21.10
		fwrd	51.60	50.85	**95.20**	**66.29**		fwrd	34.70	34.40	34.50
								none	36.90	38.40	37.60
								bidir	35.30	**43.20**	**38.80**
	A3	bwrd	54.80	55.61	47.60	51.29	34.20	bwrd	**32.70**	**26.40**	29.20
		fwrd	**61.20**	61.57	59.60	60.57		fwrd	33.30	34.40	33.90
								none	36.90	46.40	41.10
								bidir	32.70	29.60	31.10
	A1+A2	bwrd	57.60	57.72	**56.80**	57.26	33.60	bwrd	24.70	18.40	21.10
		fwrd	59.80	58.84	65.20	61.86		fwrd	34.70	34.40	34.50
								none	36.90	38.40	37.60
								bidir	35.30	**43.20**	**38.80**
	A1+A3	bwrd	57.20	57.96	52.40	55.04	33.00	bwrd	26.60	20.00	22.80
		fwrd	58.60	58.05	62.00	59.96		fwrd	31.90	34.40	33.10
								none	36.70	40.80	38.60
								bidir	34.80	36.80	35.80
	A2+A3	bwrd	54.80	55.83	46.00	50.44	33.40	bwrd	32.30	25.60	28.60
		fwrd	61.00	**61.70**	58.00	59.79		fwrd	32.80	33.60	33.20
								none	34.90	46.40	39.90
								bidir	32.70	28.00	30.20
	A1 + A2 + A3	bwrd	57.60	57.72	**56.80**	57.26	32.00	bwrd	24.00	18.40	20.80
		fwrd	59.20	58.39	64.00	61.07		fwrd	32.30	32.00	32.10
								none	36.20	37.60	36.90
								bidir	34.70	41.60	37.80
	S1 + S2 + S3	bwrd	53.20	53.77	45.60	49.35	26.00	bwrd	20.00	1.50	**29.50**
		fwrd	48.60	48.36	41.20	44.49		fwrd	16.70	0.80	31.50
								none	28.00	**63.20**	38.80
								bidir	23.70	39.20	29.50
	A1 + A2 + A3 + S1	bwrd	57.40	58.30	52.00	54.97	33.00	bwrd	27.60	19.20	22.60
		fwrd	59.80	58.84	65.20	61.86		fwrd	29.80	33.60	31.60
								none	**38.20**	41.60	39.80
								bidir	34.60	37.60	36.00
	A1 + A2 + A3 + S2	bwrd	57.80	58.52	53.60	55.95	32.60	bwrd	26.70	19.20	22.30
		fwrd	59.60	58.70	64.80	61.60		fwrd	30.70	33.60	32.10
								none	37.30	40.00	38.60
								bidir	33.80	37.60	35.60
	A1 + A2 + A3 +S3	bwrd	**58.20**	58.51	56.40	**57.44**	32.80	bwrd	24.70	19.20	21.60
		fwrd	59.60	58.82	64.00	61.30		fwrd	32.00	32.80	32.40
								none	37.40	39.20	38.30
								bidir	34.70	40.00	37.20

Table 1: Cross-lingual Textual Entailment Results for Word alignment Features and String Similarity Measures, A1 = count of unaligned words in T2, A2 = count of unaligned numbers in T2, A3 = count of unaligned words in T2 with unaligned probability < 0.11, S1 = Number of matched words in the aligned sequence given by Smith-Waterman algorithm, S2 = Penalty of aligning sentences using Smith-Waterman algorithm, S3 = Levenshtein distance between the sentences

texts in the *SEM 2013 Shared Task (Gella et al., to appear). Using the alignments, we replace each English word with its corresponding word in German. The resulting German sentence is compared with the actual one using string similarity measures. As the structure of both English and German sentences are usually SVO, we hypothesize that when there is no entailment between the two given sentences, the newly-made German sentence and the original German sentence will differ a lot in word order.

In order to compare the two German sentences, we use the Levenshtein (Levenshtein, 1966) and the Smith-Waterman (Smith and Waterman, 1981) algorithm. The Levenshtein algorithm measures the number of world-level edits to change one sentence into another. The edit operators consist of insertion and deletion. We consider substitution as two edits (combination of insertion and deletion) based on the findings of Baldwin (2009).

We also use Smith-Waterman (SW) algorithm, which was originally developed to find the most similar region between two proteins. The algorithm looks for the longest common substring, except that it permits small numbers of penalized editions consisting of insertion, deletion and substitution. We call the best found substring the 'SW aligned sequence'. In this experiment, we consider the number of matched words and the number of penalties in the SW aligned sequence as features.

Results for the string similarity features are shown in Table 1. Since the string similarity feature scores do not take the entailment direction into account, i.e. there is a single set of feature scores for each text fragment pair as there is no distinction between forward and backward entailment, and they are not suited for standalone use in compositional classification. We do, however, include these scores in Table 1 to illustrate how with the compositional approach using the same set of features for forward and backward ultimately results in a classification of test pairs as either bidirectional or no_entailment.

When individual string similarity features are added to the word alignment features, minor gains in accuracy are achieved over the word alignment features alone, +1% for S1, +0.6% for S2 and +0.8% for S3 (= Levenstein).

5 Possible Additions: Dictionary Features

We hypothesize that when there is no entailment between the two sentences, the aligner may not accurately align words. An on-line dictionary containing lemmatized words, such as Panlex (Baldwin and Colowick, 2010), could be used to avoid errors in such cases. Dictionary-based feature scores based on the presence or absence of alignments in the dictionary could then be applied.

6 Conclusions

This paper describes a compositional cross-lingual approach to CLTE with experiments carried out for the German-English language pair. Our results showed that in the first stages of binary classification as *forward* and *backward*, the word alignment features alone achieved good accuracy but when combined suffer severely. Accuracy of the approach using word alignment features could benefit from a more directional multi-class classification as opposed to the compositional approach we used. In addition, results showed minor increases in accuracy can be achieved using string similarity measures.

Acknowledgments

This work was supported by the Australian Research Council.

References

Timothy Baldwin and Jonathan Pool Susan M. Colowick. 2010. Panlex and lextract: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 37–40.

Timothy Baldwin. 2009. The hare and the tortoise: Speed and reliability in translation retrieval. *Machine Translation*, 23(4):195–240.

L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, and B. Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC 2009 Workshop Proceedings*, Gaithersburg, MD.

L. Bentivogli, P. Clark, I. Dagan, H. T. Dang, and D. Giampiccolo. 2010. The sixth PASCAL recognizing textual entailment challenge. In *TAC 2010 Workshop Proceedings*, Gaithersburg, MD.

Spandana Gella, Bahar Salehi, Marco Lui, Karl Grieser, Paul Cook, and Timothy Baldwin. to appear. Integrating predictions from multiple domains and feature sets

for estimating semantic textual similarity. In *Proceedings of *SEM 2013 Shared Task STS*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan HerbstHieu Hoang. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707.

Y. Mehdad, M. Negri, and M. Federico. 2010. Towards cross-lingual textual entailment. In *Proceedings of NAACL-HLT*.

Y. Mehdad, M. Negri, and M. Federico. 2011. Using parallel corpora for cross-lingual textual entailment. In *Proceedings of ACL-HLT 2011*.

Yashar Mehdad, Matteo Negri, and Jose G. C. de Souza. 2012. Fbk: Cross-lingual textual entailment with-out translation. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval2012)*.

M. Negri, L. Bentivogli, Y. Mehdad, D. Giampiccolo, and A. Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of EMNLP 2011*.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In *First Joint Conference on Lexical and Computational Semantics*, pages 399–407, Montreal, Canada.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, College Park, MD.

Bahar Salehi and Paul Cook. to appear. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*.

Temple F Smith and Michael S Waterman. 1981. The identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197.