UAlberta at SemEval-2020 Task 2: Using Translations to Predict Cross-Lingual Entailment

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, Grzegorz Kondrak
Alberta Machine Intelligence Institute, Department of Computing Science
University of Alberta, Edmonton, Canada
{bmhauer,amirahmad,yixing1,amallik,gkondrak}@ualberta.ca

Abstract

We investigate the hypothesis that translations can be used to identify cross-lingual lexical entailment. We propose novel methods that leverage parallel corpora, word embeddings, and multilingual lexical resources. Our results demonstrate that the implementation of these ideas leads to improvements in predicting entailment.

1 Introduction

In this paper, we discuss the University of Alberta systems for SemEval-2020 Task 2: Predicting Multilingual and Cross-Lingual Lexical Entailment (Glavaš et al., 2020). We focus on the subtask of cross-lingual binary lexical entailment (LE). Vyas and Carpuat (2016) define this task as "the task of detecting whether the meaning of a word in one language can be inferred from the meaning of a word in another language." They note its potential applications to machine translation, question answering, as well as to cross-lingual inference and entity linking. LE is related to hypernym detection, with the former being more general (Upadhyay et al., 2018).

Our principal objective is to provide evidence for the hypothesis that translations are useful in predicting cross-lingual entailment. It has been observed in prior work on cross-lingual lexical semantics that translations may be broader in meaning than the original text (Bentivogli and Pianta, 2000; Rudnicka et al., 2012). In particular, translations may represent concepts entailed by the translated concept. For example, from the English phrase "you gave me the bottle", and its Italian translation "mi hai dato il contenitore", it can be inferred that *bottle* entails *contenitore* ("container").[1] We are interested in leveraging this phenomenon to perform unsupervised LE prediction.

Our use of bitexts, word embeddings, and multilingual wordnets builds upon prior work. Qiu et al. (2018) observe that similar words share entailments, and so semantic similarity can be used to detect additional entailment pairs. Cross-lingual word embedding similarity has been used by Hauer et al. (2017) to identify translations, and by Hauer et al. (2019) to detect frequent word senses. Mehdad et al. (2011) leverage bitext alignment for textual entailment.

The principal contribution of this paper is the presentation and evaluation of LE prediction methods that leverage: (1) translation mining for entailment classification; (2) monolingual word embeddings for expanding the set of entailment pairs; and (3) multilingual lexical resources for improving translation alignment. To the best of our knowledge, we are the first to apply these ideas to cross-lingual LE prediction, and demonstrate that each of them improves LE prediction performance, especially in a low-resource setting.

2 Methods

In this section, we outline our bitext-based approach to predicting cross-lingual entailment. Section 2.1 describes our base method, BITEXT, which mines entailment pairs from word alignments in a bilingual parallel corpus. Section 2.2 describes an enhanced method, VECTORS, which identifies additional

[1] This is an actual translation from the OpenSubtitles corpus (Lison and Tiedemann, 2016).

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 263–269
Barcelona, Spain (Online), December 12, 2020.

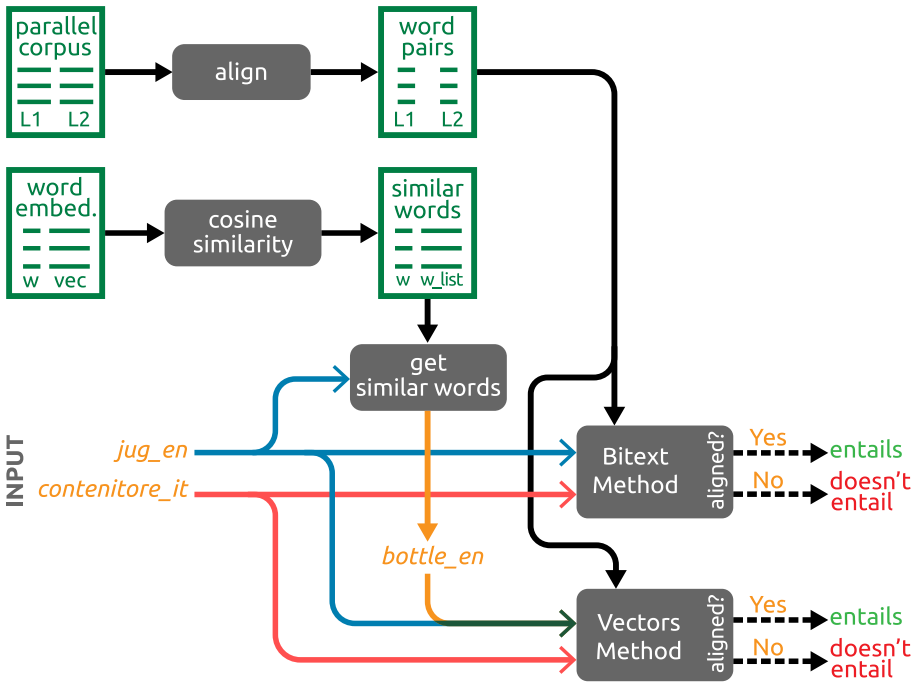

Figure 1: Two methods for identifying cross-lingual entailment.

entailment pairs based on estimates of semantic similarity. Section 2.3 outlines our use of a knowledge-based alignment method, BABALIGN, to further improve entailment coverage.

2.1 Entailment via Alignment

Our basic method, which we call BITEXT, represents the key idea of our approach: lexical translation captures entailment relations. The method uses automatic word alignment of bitexts to mine translation pairs. We make the assumption that a word and its translation either represent the same concept or one entails the other. Therefore, we make a working assumption that a bitext alignment link between a pair of words indicates an entailment relation between them. We expect that in most cases the relation is synonymy or equivalence, but a subset of the word pairs may involve hypernymy instead. How to detect the direction of such entailments is an open question.

Our implementation of the above insight starts by performing word alignment of a sentence-aligned bitext, and then extracts all aligned pairs of words. Given a test instance of LE prediction, we simply check whether the two words are among the aligned pairs to obtain a binary classification. Figure 1 shows the BITEXT methods in action: the English word *jug* is found to entail the Italian word *contenitore* if it is aligned to it in the bitext. Note that the BITEXT methods makes no use of word similarity.

2.2 Semantic Expansion

Our second method, which we refer to as VECTORS, expands upon the BITEXT method outlined in the previous section. The coverage of the latter is limited by the number of translation pairs in the aligned bitext. For example, if *jug* is never translated as *contenitore* in the bitext, the BITEXT method will fail to identify the entailment between the two words. However, the bitext may contain another word, such as *bottle*, that is both aligned to *contenitore* and semantically similar to *jug*. This observation motivates the VECTORS method. The intuition behind this method is that *semantically similar words tend to entail the same set of words*.

In addition to checking whether an input word pair is aligned in the bitext, the VECTORS method searches for aligned words that are semantically similar. Implementation of this method requires an automated way of measuring semantic similarity between words. We use the well-known measure of cosine similarity between monolingual word embeddings. The word embeddings are trained independently

on each side of the bitext. If the cosine similarity of two words is not less than a tunable threshold, the words are deemed to be semantically similar. Note that the search for similar words is performed only with respect to the first of the words in a given instance, which may entail the second word.

Figure 1 shows an example of applying the VECTORS method. Although *jug* is not translated as *contenitore* in the bitext, the method is still able to detect the entailment, based on the semantic similarity between *jug* and *bottle*.

2.3 Knowledge-Based Alignment

Both of our methods are dependent upon accurate alignment. Therefore, in addition to an off-the-shelf alignment tool, FASTALIGN (Dyer et al., 2013), we also apply a new knowledge-based alignment algorithm, BABALIGN (Luan, 2020). We first obtain all possible translations for all content words in the trial, development, and test sets from a multilingual lexical resource, such as BabelNet (Navigli and Ponzetto, 2012). To bias FASTALIGN, we append the obtained translation pairs to the lemmatized bitext. We then use the sets of translation pairs again to post-process the generated alignment: if FASTALIGN failed to link a source content word to a target content word, we attempt to align the word to its translation. We apply this process in both translation directions.

3 Experiments

In this section, we describe the details of our experiments, including tools, setup, results, and implications. At the end, we discuss an additional experiment in which we use the hypernymy relation in BabelNet as a proxy for lexical entailment.

3.1 Tools and Resources

Our bitexts are from the OpenSubtitles[2] project (Lison and Tiedemann, 2016; Tiedemann, 2016). Table 1 shows the corpus size for each language pair. We lower-case all text, and tokenize by white space and punctuation.

To improve coverage, we experiment with performing lemmatization prior to alignment. We employ TreeTagger (Schmid, 1999; Schmid, 2013) for English, German and Italian, and *reldi-tagger* (Ljubesic et al., 2016) for Croatian. No lemmatization was done for Turkish and Albanian, due to the unavailability of lemmatizers.

Languages	de-en	de-hr	de-it	en-it
lines	22.5M	13.8M	13.6M	35.2M
bytes	2.7G	1.0G	1.1G	2.6G

Table 1: The bitext size in the high-resource setting.

For the purpose of computing word similarity in the VECTORS method, we generate word embeddings using the skip-gram model of word2vec (Mikolov et al., 2013). We set the vector dimensions to 200, the context window size to 10, and run word2vec for 25 iterations. All other parameters affecting the vectors are left at their default values.

3.2 Experimental Setup

We perform experiments in two settings: low-resource (LR) and high-resource (HR). In the LR setting, bitexts are limited to one million randomly selected sentence pairs, and no lemmatization is used. The HR setting is suitable to evaluate the impact of the knowledge-based alignment. For FASTALIGN, we apply lemmatizers after the alignment. For BABALIGN we first extract from BabelNet all possible translations of the tested lemmas, which are then used to guide the alignment process.

The VECTORS method has a tunable parameter: the cosine similarity threshold for deciding semantic similarity. We tune the threshold on the official trial data set of each language pair, if such a set is provided,

[2] http://opus.nlpl.eu/OpenSubtitles2018.php

or the development data otherwise. For surprise language pairs we instead adopt a threshold value of 0.25 based on the tuning results for the other language pairs. All development (including the formulation of the various system configurations) and parameter tuning was performed using only the trial and development data. While the VECTORS method, with BABALIGN, represents our most sophisticated method, we also tested simpler configurations, to facilitate a more comprehensive analysis.

3.3 Test Results

Table 2 shows our low-resource results on the test sets. As can be seen, the incorporation of word embeddings results in an average F-score improvement of 30%.

Method	de-en	de-hr	de-it	de-sq	de-tr	en-hr	en-it	en-sq	en-tr	hr-it	hr-sq	hr-tr	it-sq	it-tr	sq-tr	Average
BITEXT	24.7	11.4	21.4	20.8	17.1	19.2	26.4	20.5	21.6	25.9	25.4	20.1	32.0	24.5	25.0	22.4
VECTORS	63.1	47.6	49.6	43.2	46.4	64.2	67.9	52.0	61.2	55.4	46.4	44.4	50.7	52.2	43.8	52.5

Table 2: F-score on the test sets in the low resource (LR) setting.

Table 3 shows the results in the high-resource setting on a subset of four language pairs. (For comparison, the low-resource results are copied from Table 2.) Our complete system is the VECTORS method combined with our knowledge-based alignment, BABALIGN, which demonstrates the best F-score of 64.5%, averaged over four language pairs. We were unable to obtain results for all language pairs due to time and resource constraints. Moreover, our knowledge-based alignment method, BABALIGN, yields clear improvements over the standard alignment for both BITEXT and VECTORS methods on all language pairs. The VECTORS method is substantially more accurate than the BITEXT method in the high-resource setting as well.

Method	Alignment	Setting	de-en	de-hr	de-it	en-it	Average
BITEXT	FASTALIGN	LR	24.7	11.4	21.4	26.4	21.0
	FASTALIGN	HR	31.2	32.6	26.3	60.2	37.6
	BABALIGN	HR	52.4	41.5	40.9	61.5	49.1
VECTORS	FASTALIGN	LR	63.1	47.6	49.6	67.9	57.0
	FASTALIGN	HR	65.0	54.7	51.7	74.3	61.4
	BABALIGN	HR	70.7	55.5	56.6	75.3	64.5

Table 3: F-score on the test sets in both low-resource (LR) and high-resource (HR) settings.

As the shared task only permits the submission of three sets of results, the VECTORS results represent our official three submissions. The BITEXT results are unofficial. The F-scores for the BITEXT methods on the test set were provided to us by the organizers. However, since the precision and recall on the test set were not provided, we present the detailed evaluation results of our best model on the development set in Table 4. Overall, the test results of our complete system in Table 3 and its development results in Table 4 constitute a strong proof-of-concept.

Languages	True Pos.	False Pos.	True Neg.	False Neg.	Precision(%)	Recall(%)	F_1(%)	Accuracy(%)
de-en	167	151	76	24	52.5	87.4	65.6	58.1
de-hr	139	150	131	33	48.1	80.8	60.3	59.6
de-it	122	149	134	41	45.0	74.8	56.2	57.4
en-it	201	99	89	17	67.0	92.2	77.6	71.4

Table 4: Detailed evaluation results on the development sets for our best method of VECTORS using BABALIGN in the HR setting.

3.4 Error Analysis

The VECTORS method is an expansion of the BITEXT method. For any test instance, if BITEXT returns a positive classification, then VECTORS does so as well. Thus the set of entailment relations reported by the former is a subset of the entailment relations reported by the latter. Consequently, VECTORS reduces the number of false negatives, at the cost of a higher number of false positives. Overall, the result is a substantial net gain in LE prediction accuracy. For example, on the English-Italian test set in the low-resource setting, precision drops from 82% to 62%, but the recall increases from 14% to 75%.

For instance, consider the entailment of English *plant* by Italian *rosa* "rose". Since these words are not aligned in the English-Italian bitext, the BITEXT method returns a false negative. However, the VECTORS method identifies *fiore* "flower" as semantically similar to *rosa*, and correctly returns a positive classification because the word *plant* happens to be translated as *fiore* in the bitext.

One weakness of the bitext-based methods is the inability to distinguish the direction of an entailment relation. This can lead to false negatives, an issue which the VECTORS method sometimes exacerbates. For instance, Italian *creatura* "creature" does not entail English *wolf*. However, VECTORS incorrectly predicts otherwise, because *creatura* and *animale* "animal" are semantically similar, and *animale* is found to be aligned with *wolf*.

Another source of errors are non-literal translations. For example, the English phrase *automobile key* is translated in the bitext as *chiave di accensione* "ignition key". This leads to the incorrect conclusion that *automobile* is entailed by *accensione* "ignition" and similar words.

Finally, lemmatization was found to reduce the number of alignment errors. For instance, the Italian word-form *orchidea* is over 10 times less frequent in the bitext than its plural form *orchidee*, which results in the singular (i.e., lemma) form often being misaligned by the cooccurrence-based alignment algorithm. This in turn prevents the BITEXT method from identifying the entailment between *orchidea* and *flower*.

3.5 Cross-lingual Entailment in BabelNet

Since lexical entailment is closely related to the hypernymy and hyponymy relations, we decided to investigate the effectiveness of a baseline strategy based on BabelNet. In order to make a LE prediction, we simply check whether two words in a given instance are connected by a chain of hypernymy links in BabelNet. For efficiency, we limit the length of a chain to a maximum of five links.

Table 5 shows the results on the development sets of four language pairs, alongside our best system, the VECTORS method with BABALIGN. The results indicate that BABELNET performs very well, outperforming VECTORS by over 20% on average.

Method	Alignment	Setting	de-en	de-hr	de-it	en-it	Average
VECTORS	BABALIGN	HR	65.6	60.3	56.2	77.6	64.9
BABELNET	n/a	HR	87.4	81.6	81.5	93.9	86.1

Table 5: F-score on the development sets with the hypernymy-based method.

Analysis of incorrect predictions uncovers at least three different sources of errors. First, certain entailment relations involve words that are far apart in BabelNet; for instance, English *animal* is found 7 hypernym links above Italian *pinguino*. This results in a false negative caused by the five-link height limit. Second, some semantic relationships are missing in BabelNet; for example the "planet" sense of *Pluto* has no hypernym, so its entailment of English *planet* remains undetected. In addition to the errors of omission, there are errors of commission. An example is a synset, glossed as "the direction corresponding to the southwestward compass point", which contains both English *north* and Italian *sud* "south". This results in a false positive, because synonyms are considered to entail one another.

The generally high performance of BabelNet-based entailment detection suggests that BabelNet, or a comparable resource, could be used to compensate for the limitations of our method (see Section 3.4). For example, BabelNet could be used to filter out translation pairs which are related by synonymy, rather than entailment, or to verify the direction of an entailment relation.

4 Conclusion

We have demonstrated a strong connection between translations and cross-lingual entailment. The sparsity of the translation data can be alleviated by the use of semantic similarity between word embeddings. Finally, applying knowledge-based word alignment results in substantial improvements in identifying entailment in bitexts.

Acknowledgements

We thank the organizers of the shared task for their effort. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Luisa Bentivogli and Emanuele Pianta. 2000. Looking for lexical gaps. In *Proceedings of the ninth EURALEX International Congress*, pages 8–12. Stuttgart: Universität Stuttgart.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, June.

Goran Glavaš, Ivan Vulić, Anna Korhonen, and Simone Ponzetto. 2020. SemEval-2020 task 2: Predicting multilingual and cross-lingual (graded) lexical entailment. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Bradley Hauer, Garrett Nicolai, and Grzegorz Kondrak. 2017. Bootstrapping unsupervised bilingual lexicon induction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 619–624, April.

Bradley Hauer, Yixing Luan, and Grzegorz Kondrak. 2019. You shall know the most frequent sense by the company it keeps. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 208–215. IEEE.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016*, pages 923–929. European Language Resources Association.

Nikola Ljubesic, Filip Klubicka, Zeljko Agic, and Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of croatian and serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Yixing Luan. 2020. Leveraging translations for word sense disambiguation. Master's thesis, University of Alberta.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-volume 1*, pages 1336–1345. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Wei Qiu, Mosha Chen, Linlin Li, and Luo Si. 2018. NLP_HZ at SemEval-2018 task 9: a nearest neighbor approach. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 909–913, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A strategy of mapping Polish Wordnet onto Princeton Wordnet. In *Proceedings of COLING 2012: Posters*, pages 1039–1048, Mumbai, India, December. The COLING 2012 Organizing Committee.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Helmut Schmid. 2013. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Shyam Upadhyay, Yogarshi Vyas, Marine Carpuat, and Dan Roth. 2018. Robust cross-lingual hypernymy detection using dependency context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 607–618, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2019. Multilingual and cross-lingual graded lexical entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4963–4974, Florence, Italy, July. Association for Computational Linguistics.

Yogarshi Vyas and Marine Carpuat. 2016. Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1187–1197, San Diego, California, June. Association for Computational Linguistics.