Word's Vector Representations meet Machine Translation

Eva Martínez Garcia
Cristina España-Bonet
TALP Research Center
Univesitat Politècnica de Catalunya
emartinez@lsi.upc.edu
cristinae@lsi.upc.edu

Jörg Tiedemann
Uppsala University
Department of Linguistics
and Philology
jorg.tiedemann@lingfil.uu.se

Lluís Màrquez
Qatar Computing Research Institute
Qatar Foundation
lluism@lsi.upc.edu

Abstract

Distributed vector representations of words are useful in various NLP tasks. We briefly review the CBOW approach and propose a bilingual application of this architecture with the aim to improve consistency and coherence of Machine Translation. The primary goal of the bilingual extension is to handle ambiguous words for which the different senses are conflated in the monolingual setup.

1 Introduction

Machine Translation (MT) systems are nowadays achieving a high-quality performance. However, they are typically developed at sentence level using only local information and ignoring the document-level one. Recent work claims that discourse-wide context can help to translate individual words in a way that leads to more coherent translations (Hardmeier et al., 2013; Hardmeier et al., 2012; Gong et al., 2011; Xiao et al., 2011).

Standard SMT systems use n-gram models to represent words in the target language. However, there are other word representation techniques that use vectors of contextual information. Recently, several distributed word representation models have been introduced that have interesting properties regarding to the semantic information that they capture. In particular, we are interested in the *word2vec* package available in (Mikolov et al., 2013a). These models proved to be robust and powerful for predicting semantic relations between words and even across languages. However, they are not able to handle lexical ambiguity as they conflate word senses of polysemous words into one common representation. This limitation is already discussed in (Mikolov et al., 2013b) and in (Wolf et al., 2014), in which bilingual extensions of the word2vec architecture are proposed. In contrast to their approach, we are not interested in monolingual applications but instead like to concentrate directly on the bilingual case in connection with MT.

We built bilingual word representation models based on word-aligned parallel corpora by an application of the Continuous Bag-of-Words (CBOW) algorithm to the bilingual case (Section 2). We made a twofold preliminary evaluation of the acquired word-pair representations on two different tasks (Section 3): predicting semantically related words (3.1) and cross-lingual lexical substitution (3.2). Section 4 draws the conclusions and sets the future work in a direct application of these models to MT.

2 Semantic Models using CBOW

The basic architecture that we use to build our models is CBOW (Mikolov et al., 2013a). The algorithm uses a neural network (NN) to predict a word taking into account its context, but without considering word order. Despite its drawbacks, we chose to use it since we presume that the translation task applies the same strategy as the CBOW architecture, i.e., from a set of context words try to predict a translation of a specific given word.

In the monolingual case, the NN is trained using a monolingual corpus to obtain the corresponding projection matrix that encloses the vector representations of the words. In order to introduce the semantic information in a bilingual scenario, we use a parallel corpus and automatic word alignment to extract a training corpus of word pairs: $(w_{i,S}|w_{i,T})$. This approach is different from (Wolf et al., 2014) who build an independent model for each language. With our method, we try to capture simultaneously the semantic information associated to the source word and the information in the target side of the translation. In this way, we hope to better capture the semantic information that is implicitly given by translating a text.

Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 132–134,
October 25, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

Model	Accuracy	Known words
mono_en	32.47 %	64.67 %
mono_es	10.24 %	44.96 %
bi_en-es	23.68 %	13.74 %

Table 1: Accuracy on the Word Relationship set.

3 Experiments

The semantic models are built using a combination of freely available corpora for English and Spanish (EuropalV7, United Nations and Multilingual United Nations, and Subtitles2012). They can be found in the Opus site (Tiedemann, 2012).We trained vectors to represent word pairs forms using this corpora with the *word2vec* CBOW implementation. We built a training set of almost 600 million words and used 600-dimension vectors in the training. Regarding to the alignments, we only used word-to-word ones to avoid noise.

3.1 Accuracy of the Semantic Model

We first evaluate the quality of the models based on the task of predicting semantically related words. A Spanish native speaker built the bilingual test set similarly to the process done to the training data from a list of $19,544$ questions introduced by (Mikolov et al., 2013c). In our bilingual scenario, the task is to predict a pair of words given two pairs of related words. For instance, given the pair `Athens`|`Atenas` `Greece`|`Grecia` and the question `London`|`Londres`, the task is to predict `England`|`Inglaterra`.

Table 1 shows the results, both overall accuracy and accuracy over the known words for the models. Using the first $30,000$ entries of the model (the most frequent ones), we obtain 32% of accuracy for English (mono_en) and 10% for Spanish (mono_es). We chose these parameters for our system to obtain comparable results to the ones in (Mikolov et al., 2013a) for a CBOW architecture but trained with 783 million words (50.4%). Decay for the model in Spanish can be due to the fact that it was built from automatic translations. In the bilingual case (bi_en-es), the accuracy is lower than for English probably due to the noise in translations and word alignment.

3.2 Cross-Lingual Lexical Substitution

Another way to evaluate the semantic models is through the effect they have in translation. We implemented the Cross-Lingual Lexical Substitution task carried out in SemEval-2010 (Task2, 2010) and applied it to a test set of news data from the News Commentary corpus of 2011.

We identify those content words which are translated in more than one way by a baseline translation system (Moses trained with Europarl v7). Given one of these content words, we take the two previous and two following words and look for their vector representations using our bilingual models. We compute a linear combination of these vectors to obtain a context vector. Then, to chose the best translation option, we calculate a score based on the similarity among the vector of every possible translation option seen in the document and the context vector.

In average there are 615 words per document within the test set and 7% are translated in more than one way by the baseline system. Our bilingual models know in average 87.5% of the words and 83.9% of the ambiguous ones, so although there is a good coverage for this test set, still, some of the candidates cannot be retranslated or some of the options cannot be used because they are missing in the models. The accuracy obtained after retranslation of the known ambiguous words is 62.4% and this score is slightly better than the result obtained by using the most frequent translation for ambiguous words (59.8%). Even though this improvement is rather modest, it shows potential benefits of our model in MT.

4 Conclusions

We implemented a new application of word vector representations for MT. The system uses word alignments to build bilingual models with the final aim to improve the lexical selection for words that can be translated in more than one sense.

The models have been evaluated regarding their accuracy when trying to predict related words (Section 3.1) and also regarding its possible effect within a translation system (Section 3.2). In both cases one observes that the quality of the translation and alignments previous to building the semantic models are bottlenecks for the final performance: part of the vocabulary, and therefore translation pairs, are lost in the training process.

Future work includes studying different kinds of alignment heuristics. We plan to develop new features based on the semantic models to use them inside state-of-the-art SMT systems like Moses (Koehn et al., 2007) or discourse-oriented decoders like Docent (Hardmeier et al., 2013).

References

Z. Gong, M. Zhang, and G. Zhou. 2011. Cache-based document-level statistical machine translation. In *Proc. of the 2011 Conference on Empirical Methods in NLP*, pages 909–919, UK.

C. Hardmeier, J. Nivre, and J. Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proc. of the Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning*, pages 1179–1190, Korea.

C. Hardmeier, S. Stymne, J. Tiedemann, and J. Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proc. of the 51st ACL Conference*, pages 193–198, Bulgaria.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proc. of the 45th ACL Conference*, pages 177–180, Czech Republic.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*. http://code.google.com/p/word2vec.

T. Mikolov, Q. V. Le, and I. Sutskever. 2013b. Exploiting similarities among languages for machine translation. In *arXiv*.

T. Mikolov, I. Sutskever, G. Corrado, and J. Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Task2. 2010. Cross-lingual lexical substitution task, semeval-2010. http://semeval2.fbk.eu/semeval2.php?location=tasksT24.

J. Tiedemann. 2009. News from opus - a collection of multilingual parallel corpora with tools and interfaces. In *N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V)*, pages 237–248, Amsterdam/Philadelphia. John Benjamins.

J. Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*. http://opus.lingfil.uu.se/.

L. Wolf, Y. Hanani, K. Bar, and N. Derschowitz. 2014. Joint word2vec networks for bilingual semantic representations. In *Poster sessions at CICLING*.

T. Xiao, J. Zhu, S. Yao, and H. Zhang. 2011. Document-level consistency verification in machine translation. In *Proc. of Machine Translation Summit XIII*, pages 131–138, China.