Using Word Embeddings for Improving Statistical Machine Translation of Phrasal Verbs

Kostadin Cholakov and Valia Kordoni
Humboldt-Universität zu Berlin, Germany
{kostadin.cholakov,kordonieva}@anglistik.hu-berlin.de

Abstract

We examine the employment of word embeddings for machine translation (MT) of phrasal verbs (PVs), a linguistic phenomenon with challenging semantics. Using word embeddings, we augment the translation model with two features: one modelling distributional semantic properties of the source and target phrase and another modelling the degree of compositionality of PVs. We also obtain paraphrases to increase the amount of relevant training data. Our method leads to improved translation quality for PVs in a case study with English to Bulgarian MT system.

1 Introduction

Phrasal verbs (PVs) are a type of multiword expressions (MWEs) and as such, their semantics is not predictable, or is only partially predictable, from the semantics of their components. In statistical machine translation (SMT) the word-to-word translation of MWEs often results in wrong translations (Piao et al., 2005). Previous work (Ren et al. (2009), Carpuat and Diab (2010), Cholakov and Kordoni (2014)) has shown that dedicated techniques for identification of MWEs and their integration into the translation algorithms improve the quality of SMT. Generally, those techniques are based on categorical representations. MWEs are either treated as a single unit or binary features encoding properties of MWEs are added to the translation table. On the other hand, recent works have successfully applied distributional representations of words and phrases in SMT (Mikolov et al. (2013a), Zhang et al. (2014), Alkhouli et al. (2014)). The idea behind is that similar words and phrases in different languages tend to have similar distributional representations (Mikolov et al., 2013a).

In this paper, we explore the usage of such representations for improving SMT of PVs. We propose three strategies based on word embeddings. First, we employ continuous vectors of phrases learnt using neural networks to provide semantic scoring of aligned phrase pairs containing PVs. The addition of this score to the SMT model is a step toward integration of semantic information about the PVs into the translation process. Second, we use the vectors learnt to find paraphrases of the original phrase pairs and add those to the translation table. This increases the amount of relevant parallel data. Third, we make use of word embeddings to map a PV onto a continuous-valued compositionality score and add this score as a feature in the SMT model. The score indicates the semantic similarity between a PV and the verb forming that PV, i.e. the degree of compositionality of the PV. The meaning of (semi-)compositional PVs can be (partially) derived from the meaning of their lexemes, e.g. *carry in*. Previous work (Cholakov and Kordoni, 2014) treats PVs as either compositional or idiomatic while we handle compositionality as a continuous phenomenon.

We perform a case study with an English to Bulgarian SMT system. An English PV is generally translated to a single Bulgarian verb. This many-to-one mapping poses difficulties for SMT. The combined integration of all three strategies presented above outperforms the results reported in previous work both in automated and manual evaluation. Thus we show that word embeddings help SMT to handle better such a challenging linguistic phenomenon as PVs.

Proceedings of the 12th Workshop on Multiword Expressions, pages 56–60,
Berlin, Germany, August 7-12, 2016.

2 Related Work

Previous work on SMT of MWEs (Lambert and Banchs (2005), Carpuat and Diab (2010), Simova and Kordoni (2013)) suggests training the SMT system on corpora in which each MWE is treated as a single unit, e.g. *call_off*. Ren et al. (2009) treat bilingual MWEs pairs as parallel sentences which are then added to the training data. Other methods (Simova and Kordoni (2013), Cholakov and Kordoni (2014)) perform feature mining and modify directly the translation table. In addition to the standard translational probabilities, those methods add binary features which indicate whether a source phrase contains MWEs and whether an MWE is compositional or idiomatic. Our work modifies both the training data (via the addition of paraphrases) and the translation table. However, the modifications come from the usage of word embeddings assuming that those allow for a better incorporation of semantic information into SMT.

Following the work of Mikolov et al. (2013a), Mikolov et al. (2013b), and Alkhouli et al. (2014), we exploit the idea that vector representations of similar words in different languages are related by a linear transformation. However, we focus on exploring this idea on a specific phenomenon with challenging semantics, namely PVs. Finally, there has been significant research on predicting the compositionality of MWEs (e.g., Schulte im Walde et al. (2013), Salehi et al. (2015)) under the assumption that this could be helpful in applications. Here, we go a step further and prove this assumption correct by integrating compositionality into a real-life application such as SMT.

3 English–Bulgarian SMT System

Translation of PVs. In (1) the PV *called off* has to be mapped to the single Bulgarian verb *otmeni*. For more convenience, the Bulgarian sentence is transcribed with Latin letters.

(1) Toj *otmeni* sreshtata.
he cancelled meeting-the
'He *called off* the meeting.'

Another challenge is the mapping of an English PV to a 'da'-construction. Such constructions are very frequent in Bulgarian since they denote complex verb tenses, modal verb constructions, and subordinating conjunctions. Guessing whether to add a 'da' particle or not is problematic for the SMT system.

Language Resources. We employ the SeTimes news corpus[1] which contains parallel articles in English and 9 Balkan languages. The training data consist of approximately 151,000 sentences. Another 2,000 sentences are used for tuning. The test set consists of 800 sentences, 400 of which contain one or more instances of PVs. We manually identified 138 unique PVs with a total of 403 instances. A language model for the target language is created based on a 50 million words subset of the Bulgarian National Reference Corpus (BNRC).[2] Finally, Moses is employed to build a factored phrase-based translation model which operates on lemmas and POS tags due to the rich Bulgarian morphology.

4 Integration of Word Embeddings

In our work, we construct word embeddings of English phrases which contain PVs and of their aligned counterparts in Bulgarian. Then we use those representations to augment the translation table with new features and phrase alignments. The word embeddings are obtained using the *word2vec* toolkit.[3] We used the continuous bag-of-words (CBOW) model. Experiments with the skip-gram model showed very close results and are not reported here.

4.1 Phrase Corpus

When training phrase vectors using neural networks, the network is presented with a *phrase corpus*. The phrase corpus is similar to a word corpus except that some words are joined to make up phrases. For example, Mikolov et al. (2013b) identify phrases using a monolingual point-wise mutual information criterion with discounting. However, since our goal is to generate phrase vectors that are helpful for translation of PVs, we limit the construction of phrases in the training data for *word2vec* only to those English and Bulgarian phrases which: i) are aligned in the phrase table and ii) the English phrase contains PVs. To determine the latter, we use an automatically created lexicon of English PVs (Simova and Kordoni, 2013) and the jMWE library (Kulkarni and Finlayson, 2011) to mark potential PVs in the data.

[1] http://www.setimes.com
[2] http://webclark.org/
[3] https://code.google.com/p/word2vec

We ran this method on the MT test set of 800 sentences in order to examine its performance. It achieved 91% precision and 93% recall.

As training data for *word2vec*, we use the English part of the SeTimes corpus and the English Wikipedia dump from November 2014. Since the phrase table contains lemmas, the Wikipedia corpus was lemmatised using the TreeTagger (Schmid, 1994). For Bulgarian, the SeTimes corpus and the BNRC were employed. *Word2vec* generates a vector of fixed dimensionality d for each phrase in the training corpus. In our experiments, d is set to 300 and the size of the context window is set to 5.

4.2 Semantic Scoring Feature

Following the work in Mikolov et al. (2013b) and Alkhouli et al. (2014), we introduce an additional feature in the translation model:

$$(2) \qquad sim(Wx_{\widetilde{f}}, z_{\widetilde{e}})$$

where sim is a similarity function, $x_{\widetilde{f}}$ and $z_{\widetilde{e}}$ are the S-dimensional source and T-dimensional target vectors corresponding to the source (English) phrase $\widetilde{f}$ and target (Bulgarian) phrase $\widetilde{e}$, respectively. W is an $S \times T$ linear projection matrix that maps the source space to the target space. The matrix is estimated by optimizing the following criterion with stochastic gradient descent:

$$(3) \qquad \min_{W} \sum_{i=1}^{N} \|Wx_i - z_i\|^2$$

where the training data consists of the pairs $(x_1, z_1), \ldots, (x_N, z_N)$ corresponding to the source and target vectors. For any given phrase or word and its continuous vector representation x, we can map it to the other language space by computing $z = Wx$. Then we find the word or phrase whose representation is closest to z in the target language space, using cosine similarity as the distance metric.

Since the source and target phrase vectors are learned separately, we do not have an immediate mapping between them. That is why we resort to the phrase table to obtain it. A source and a target vectors are paired if there is a corresponding phrase pair entry in the phrase table.

4.3 Paraphrases

We use the vectors produced for Bulgarian to augment the phrase table with additional entries. Using cosine similarity, we find the top 5 similar phrases and consider them paraphrases of the original Bulgarian phrase. This is done only for entries mapped to a source English phrase containing a PV. The newly generated phrase pair is assigned the same feature values as the pair used to induce it. In order to differentiate the original phrase pair from the induced paraphrases, we introduce an additional feature which indicates the similarity between the Bulgarian phrase and its paraphrase. The value of this feature for the original phrase pair is set to 1. Finally, note that since we are interested in the proper translation of English PVs, we do not paraphrase the source English phrase.

4.4 Compositionality Score

In Cholakov and Kordoni (2014) a binary feature indicates whether a PV is compositional (1) or idiomatic (0). This solution does not reflect the different degrees of compositionality PVs exhibit. We follow the research in Schulte im Walde et al. (2013) and Salehi et al. (2015) and map each PV to a continuously-valued compositionality score which is then added as a feature to the translation model. This score is calculated as:

$$(4) \qquad comp(PV) = sim(PV, V)$$

where PV is the vector associated with the phrase verb in question, V is the vector associated with the verb forming the PV, and sim is a vector similarity function. We use *word2vec* to calculate the similarity sim between the two vectors. The idea behind the score is that the more similar the meaning of the PV is to the meaning of the verb, the more compositional this PV is. Note that in light of the findings reported in Salehi et al. (2014) and Salehi et al. (2015), we do not take into account the vector of the particle.

5 Results

Our work is directly comparable to that in Cholakov and Kordoni (2014) since we used the same datasets and MT system setup. Furthermore, we have successfully reproduced the results reported there.

Automatic Evaluation. Table 1 presents the results from the automatic evaluation, in terms of BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) scores. All results are averages of 3 MERT optimizer runs. Statistical significance is computed using the Approximate Randomization

	with PVs		all	
	bleu	nist	bleu	nist
baseline	0.244	5.97	0.237	6.14
4 binary features	0.267	6.01	0.256	6.16
semantic scoring feature	0.268	6.00	0.258	6.15
paraphrases	0.270	**6.02**	0.261	**6.18**
compositionality feature	0.269	6.01	0.260	6.17
our 3 strategies combined	**0.272**	**6.02**	**0.262**	**6.18**

Table 1: Automatic evaluation of MT quality.

	good	acceptable	incorrect
baseline	0.21	0.41	0.38
4 binary features	0.3	0.5	0.2
semantic scoring feature	0.3	0.54	0.16
paraphrases	**0.31**	0.53	0.16
compositionality feature	0.3	**0.57**	0.13
our 3 strategies combined	**0.31**	**0.57**	**0.12**

Table 2: Manual evaluation of MT quality.

(AR) test. We used the *multeval* toolkit (Clark et al., 2011) for evaluation.

In the baseline case Moses is run in a standard configuration, i.e. without any explicit MWE knowledge. Table 1 also shows the best results from Cholakov and Kordoni (2014) where 4 binary features indicate: 1) whether a phrase contains a PV; 2) whether a detected PV is transitive or not; 3) whether the particle in a PV is separable or not; and 4) whether a PV is compositional or not. We evaluated the contribution of each of our 3 strategies based on word embeddings as well as various combinations thereof. Note that, for reasons of space, we do not report on the 400 test sentences without a PV. The results for those are very close for all setups which shows that our modifications do not harm the MT quality for sentences without PVs.

The combination of our three strategies based on word embeddings achieves the best performance in terms of BLEU, with the results being statistically significant compared to all other settings at $p < 0.01$. The semantic scoring feature alone outperforms the baseline but achieves the same performance as the setting with 4 the binary features. On the other hand, the usage of paraphrases or the incorporation of compositionality feature achieve very close results and both are significantly better than the binary features setting. In fact, those settings are almost as good as the best configuration. This shows that: i) paraphrases found using *word2vec* are of good quality and help MT and ii) treating compositionality of PVs as a continuous phenomenon has positive effects on MT and outperforms the binary compositional/idiomatic setting. Last, apart from the baseline, the differences in NIST scores are not significant. We attribute this to the fact that our method improves translation of more frequent and thus less informative for NIST PVs.

Manual Evaluation. A native speaker of Bulgarian was asked to judge the translations of PVs produced by the MT system. A translation was judged as:

- *good* - correct translation of the PV, correct verb inflection
- *acceptable* - correct translation of the PV but wrong inflection, or wrongly built *da-* or reflexive construction
- *incorrect* - wrong translation which changes the meaning of the sentence

Table 2 shows the results. Compared to previous work, all our strategies achieve a significantly higher number of acceptable translations and reduce the number of wrong translations. The improvement in translation comes mostly from better translations of semi-compositional verbs which underlines the importance of better treatment of this phenomenon. Note the good performance of the setting involving the compositionality feature which directly tackles this issue.

6 Conclusion

In this paper we used word embeddings to augment the phrase table of an SMT system with new features and aligned phrase pairs which led to improved SMT of PVs. The new features aim at capturing distributional semantic properties and the degree of compositionality of PVs. In a case study with an English-Bulgarian SMT system, our work clearly outperformed previous research. In future work, we will extend our approach to other types of MWEs.

References

Tamer Alkhouli, Andreas Guta, and Hermann Ney. 2014. Vector space models for phrase-based machine translation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–10, Doha, Qatar.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.*, HLT '10., pages 242–245, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kostadin Cholakov and Valia Kordoni. 2014. Better statistical machine translation through linguistic treatment of phrasal verbs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 196–201, Doha, Qatar, October. Association for Computational Linguistics.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *49th Annual Meeting of the Association for Computational Linguistics:shortpapers*, pages 176–181, Portland, Oregon.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA.

Nidhi Kulkarni and Mark Alan Finlayson. 2011. JMWE – a Java toolkit for detecting multiword expressions. In *Proceedings of the 2011 Workshop on Multiword Expressions*, pages 122–124.

Patrik Lambert and Rafael Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proceedings of the X Machine Translation Summit*, pages 396–403.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA.

Scott Songlin Piao, Paul Rayson, and and Tony McEnery Dawn Archer. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Comuter Speech and Language*, 19(4):378–397.

Zhixiang Ren, Yajuan Lu, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the ACL Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 47–54, Singapore.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 472–481, Gothenburg, Sweden.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Annual Meeting of the North American Chapter of ACL – Human Language Technologies (NAACL HLT)*, Denver, Colorado. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Sabine Schulte im Walde, Stefan Muller, and Stephen Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 255–265, Atlanta, Georgia.

Iliana Simova and Valia Kordoni. 2013. Improving English-Bulgarian statistical machine translation by phrasal verb treatment. In *Proceedings of MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology*, Nice, France.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*.