Better Statistical Machine Translation through Linguistic Treatment of Phrasal Verbs

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

Abstract

This article describes a linguistically informed method for integrating phrasal verbs into statistical machine translation (SMT) systems. In a case study involving English to Bulgarian SMT, we show that our method does not only improve translation quality but also outperforms similar methods previously applied to the same task. We attribute this to the fact that, in contrast to previous work on the subject, we employ detailed linguistic information. We found out that features which describe phrasal verbs as idiomatic or compositional contribute most to the better translation quality achieved by our method.

1 Introduction

Phrasal verbs are a type of multiword expressions (MWEs) and as such, their meaning is not derivable, or is only partially derivable, from the semantics of their lexemes. This, together with the high frequency of MWEs in every day communication (see Jackendoff (1997)), calls for a special treatment of such expressions in natural language processing (NLP) applications. Here, we concentrate on statistical machine translation (SMT) where the word-to-word translation of MWEs often results in wrong translations (Piao et al., 2005).

Previous work has shown that the application of dedicated methods to identify MWEs and then integrate them in some way into the SMT process often improves translation quality. Generally, automatically extracted lexicons of MWEs are employed in the identification step. Further, various integration strategies have been proposed. The so called *static* strategy suggests training the SMT system on corpora in which each MWE is treated as a single unit, e.g. *call_off*. This improves SMT indirectly by improving the alignment between source and target sentences in the training data. Various versions of this strategy are applied in Lambert and Banchs (2005), Carpuat and Diab (2010), and Simova and Kordoni (2013). In all cases there is some improvement in translation quality, caused mainly by the better treatment of separable PVs, such as in *turn the light on*.

Another strategy, which is referred to as *dynamic*, is to modify directly the SMT system. Ren et al. (2009), for example, treat bilingual MWEs pairs as parallel sentences which are then added to training data and subsequently aligned with GIZA++ (Och and Ney, 2003). Other approaches perform feature mining and modify directly the automatically extracted translation table. Ren et al. (2009) and Simova and Kordoni (2013) employ Moses[1] to build and train phrase-based SMT systems and then, in addition to the standard phrasal translational probabilities, they add a binary feature which indicates whether an MWE is present in a given source phrase or not. Carpuat and Diab (2010) employ the same approach but the additional feature indicates the number of MWEs in each phrase. All studies report improvements over a baseline system with no MWE knowledge but these improvements are comparable to those achieved by static methods.

In this article, we further improve the dynamic strategy by adding features which, unlike all previous work, also encode some of the linguistic properties of MWEs. Since it is their peculiar linguistic nature that makes those expressions problematic for SMT, it is our thesis that providing more linguistic information to the translation process will improve it. In particular, we concentrate on a specific type of MWEs, namely phrasal verbs (PVs). We add 4 binary features to the translation table which indicate not only the presence of a PV but also its *transitivity*, *separability*, and *idiomaticity*. We found that PVs are very suitable for this study since we can easily extract the necessary informa-

[1] http://www.statmt.org/moses/

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 196–201,
October 25-29, 2014, Doha, Qatar.

tion from various language resources.

To prove our claim, we perform a case study with an English to Bulgarian SMT system. Bulgarian lacks PVs in the same form they appear in English. It is often the case that an English PV is translated to a single Bulgarian verb. Such many-to-one mappings cause the so called *translation asymmetries* which make the translation of PVs very problematic.

We perform automated and manual evaluations with a number of feature combinations which show that the addition of all 4 features proposed above improves translation quality significantly. Moreover, our method outperforms static and dynamic methods previously applied to the same test data. A notable increase in performance is observed for separable PVs where the verb and the particle(s) were not adjacent in the input English sentence as well as for idiomatic PVs. This clearly demonstrates the importance of linguistic information for the proper treatment of PVs in SMT.

We would like to point out that we view the work presented here as a preliminary study towards a more general linguistically informed method for handling similar types of translation asymmetries. The experiments with a single phenomenon, namely PVs, serve as a case study the purpose of which is to demonstrate the validity of our approach and the crucial role of properly integrated linguistic information into SMT. Our work, however, can be immediately extended to other phenomena, such as collocations and noun compounds.

The remainder of the paper is organised as follows. Section 2 describes the asymmetries caused by PVs in English to Bulgarian translation. Section 3 provides details about the resources involved in the experiments. Section 4 describes our method and the experimental setup. Section 5 presents the results and discusses the improvements in translation quality achieved by the method. Sections 6 concludes the paper.

2 Translation Asymmetries

We will first illustrate the main issues which arise when translating English PVs into Bulgarian. For more convenience, the Bulgarian phrases are transcribed with Latin letters.

An English PV is usually mapped to a single Bulgarian verb:

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

In the example above the PV *called off* has to be mapped to the single Bulgarian verb *otmeni*, i.e. there is many-to-one mapping. Other cases require a many-to-many type of mapping. One such case is the mapping of an English PV to a 'da'-construction in Bulgarian. Such constructions are very frequent in Bulgarian every day communication since they denote complex verb tenses, modal verb constructions, and subordinating conjunctions:

(2) Toj trjabva da skasa s neja.
he should break off with her
'He should *break off* with her.'

Here, *da skasa* should be mapped to the PV *break off*. Other such cases include Bulgarian reflexive verb constructions.

Note that such many-to-many mappings in the case of Bulgarian pose an additional challenge for the SMT system because, for a good translation, it needs to guess whether to add a 'da' particle or not which further complicates the treatment of PVs. Also, Bulgarian is a language with rich morphology and often translations with very good semantic quality lack the proper morphological inflection. This affects negatively both automated and manual evaluation of translation quality.

3 Language Resources

We employ the data used in the studies reported in Simova and Kordoni (2013). The authors experimented with both static and dynamic methods for handling PVs in an English to Bulgarian SMT system. This allows us to compare the performance of our linguistically informed approach to that of methods which do not make use of the linguistic properties of PVs.

The data for the experiments are derived from the SeTimes news corpus[2] which contains parallel news articles in English and 9 Balkan languages. The training data consist of approximately 151,000 sentences. Another 2,000 sentences are used for the tuning. The test set consists of 800 sentences, 400 of which contain one or more in-

[2] http://www.setimes.com

stances of PVs. There are 138 unique PVs with a total of 403 instances in the test data. Further, a language model for the target language is created based on a 50 million words subset of the Bulgarian National Reference Corpus.[3] All English data are POS tagged and lemmatised using the TreeTagger (Schmid, 1994). For Bulgarian, these tasks were performed with the BTB-LPP tagger (Savkov et al., 2011).

Simova and Kordoni (2013) create automatically a lexicon containing English PVs. It is employed for the identification of such verbs in the data used in the experiments. The lexicon is constructed from a number of resources: the English Phrasal Verbs section of Wiktionary,[4] the Phrasal Verb Demon dictionary,[5] the CELEX Lexical Database (Baayen et al., 1995), WordNet (Fellbaum, 1998), the COMLEX Syntax dictionary (Macleod et al., 1998), and the gold standard data used for the experiments in McCarthy et al. (2003) and Baldwin (2008). English PVs are identified in the data using the jMWE library (Kulkarni and Finlayson, 2011) as well as a post-processing module implemented in the form of a constrained grammar (Karlsson et al., 1995) which filters out spurious PV candidates. For the identification of PVs, Simova and Kordoni (2013) report 91% precision (375 correct instances found) and a recall score of 93% for the 800 test sentences.

The Moses toolkit is employed to build a factored phrase-based translation model which operates on lemmas and POS tags. Given the rich Bulgarian morphology, the use of lemma information instead of surface word forms allows for a better mapping between source and target translation equivalents. The parallel data are aligned with GIZA++. Further, 2 5-gram language models are built using the SRILM toolkit[6] on the monolingual Bulgarian data to model lemma and POS n-gram information. Note that the Bulgarian POS tags are quite complex, so they can account for a variety of morphological phenomena. Automated translation is performed by mapping English lemmas and POS tags to their Bulgarian equivalents and then generating the proper Bulgarian word form by using lemma and POS tag information.

[3] http://webclark.org/

[4] http://en.wiktionary.org/wiki/Category:English_phrasal_verbs

[5] http://www.phrasalverbdemon.com/

[6] http://www-speech.sri.com/projects/srilm/

	1	0
feature 1	PV present	no PV
feature 2	transitive	intransitive
feature 3	separable	inseparable
feature 4	idiomatic	(semi-)comp.

Table 1: Values for the 4 new features.

4 Addition of Linguistic Features

The resources from which the PV lexicon is constructed also contain various types of linguistic information. Wiktionary provides the most details since the entries there contain information about the valency of the verb (transitive vs intransitive) and whether a particle can be separated from the PV in particle verb constructions. Consider *fell off his bike* and **fell his bike off* vs *turn the engine on* and *turn on the engine*.

Further, Wiktionary indicates whether a given PV is compositional or idiomatic in nature. The meaning of (semi-)compositional PVs can be (partially) derived from the meaning of their lexemes, e.g. *carry in*. The degree of compositionality affects the productivity with which verbs and particles combine. Verbs with similar semantics often combine with the same particle, e.g. *bring/carry in*. This is not the case for fully idiomatic PVs, e.g. *get/*obtain over*. Therefore, the notion of compositionality plays a very important role in the treatment of PVs and MWEs in general. The dataset described in McCarthy et al. (2003) also indicates whether a PV is idiomatic or not.

We were able to acquire the PV lexicon and we augmented it with the information obtained from the various resources. Then, once the system is trained, we add 4 binary features to each entry in the automatically created translation table. The values those features take are shown in Table 1. If a given property is not specified for some PV in the lexicon, the value of the corresponding feature is 0. Naturally, if no PV is identified in a source phrase, the value of all 4 features is 0. This is different from previous work where only one feature is added, indicating the presence of a PV. By adding those new features, we want to bias the SMT system towards using phrases that do not "split" PVs during decoding.

	with PVs		no PVs		all	
	bleu	nist	bleu	nist	bleu	nist
baseline	0.244	5.97	0.228	5.73	0.237	6.14
static	0.246	**6.02**	0.230	**5.76**	0.239	**6.18**
dynamic-1	0.250	5.92	0.226	5.54	0.244	6.02
dynamic-4	**0.267**	6.01	**0.232**	5.74	**0.256**	6.16

Table 2: Automatic evaluation of translation quality.

	good	acceptable	incorrect
baseline	0.21	0.41	0.38
static	0.25	0.5	0.25
dynamic-1	0.24	0.51	0.25
dynamic-4	**0.3**	0.5	0.2

Table 3: Manual evaluation of translation quality.

5 Results and Discussion

Automatic Evaluation. Table 2 presents the results from the automatic evaluation, in terms of BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) scores, of 4 system setups. The baseline has no MWE knowledge, while the static and the dynamic-1 system setups are reproduced from the experiments described in Simova and Kordoni (2013). Dynamic-1 includes only a single binary feature which indicates the presence of a PV while our method, dynamic-4, includes the 4 features described in Table 1.

Our method outperforms all other setups in terms of BLEU score, thus proving our point that adding features describing the linguistic properties of PVs improves SMT even further. Also, the results for the 400 sentences without PVs show that the 4 new features do not have a negative impact for PV-free contexts.

In terms of NIST the static strategy consistently performs best, followed closely by our method. NIST is a measure which weights the translated n-grams according to their informativeness. Due to the nature of this measure, less frequent correctly translated n-grams are given more weight in the evaluation process because NIST considers them "more informative". Such less frequent n-grams, or in our case PVs, are likely to be captured better by the static setup. Therefore, this setup achieves the highest NIST scores. This fact also suggests that dynamic and static strategies influence the SMT process in different ways, with our method tending to capture more frequent (and thus less informative) n-grams. Interestingly, the other dynamic method, dynamic-1, has the worst performance of all setups in terms of NIST.

Manual evaluation. To get a better insight on how the different setups deal with the translation of PVs, we also performed a manual evaluation. A native speaker of Bulgarian was asked to judge the translations of PVs for the 375 test sentences in which such verbs were correctly identified during the identification step. The human subject takes into account the target PV and a limited context around it and judges the translation 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 3 shows the results. Our method dynamic-4 produces more good translations and less incorrect ones than all other setups. This illustrates further the benefits of adding linguistic features to the translation model. The results achieved by the static approach are attributed to the better handling of separable PVs in sentences where the particle was not adjacent to the verb. The dynamic-1 approach and the baseline often interpret the particle literally in such cases which leads to almost twice the amount of wrong translations. Our method, on the other hand, performs slightly lower than the static approach in this respect but still much better than the other 2 setups.

Compared to dynamic-1 and the baseline, the static approach also handles better idiomatic PVs but performs slightly worse for sentences with compositional PVs. However, the addition of a specific feature to encode idiomaticity in the translation model enables our method dynamic-4 to achieve the best performance for idiomatic PVs while still handling successfully many compositional PVs. To summarise, the improved results of our method in comparison to previous work are attributed to the better handling of separable PVs which occur in a split form and even more to the improved ability to differentiate between compositional and idiomatic PVs.

Feature combinations. Our method performs best when all 3 linguistic features described above

are taken into account by the SMT system. However, we also experimented with different combinations of those features in order to get some insight of the way each feature influences the translation quality. Adding only the feature denoting verb transitiveness did not lead to any significant improvement compared to the dynamic-1 setup. Also, the combination which leaves out this feature and uses the remaining ones ranks second, achieving only a slightly worse performance than dynamic-4, the setup in which all features are employed. It seems that the transitiveness feature does not contribute much to the task at hand. Adding only the feature denoting separable vs inseparable PVs and adding only the one denoting idiomaticity led to results slightly higher than those of the dynamic-1 and static setups but still, those results were significantly lower than the ones presented in Tables 2 and 3.

6 Conclusion and Outlook

In this article, we showed that the addition of linguistically informative features to a phrase-based SMT model improves the translation quality of a particular type of MWEs, namely phrasal verbs. In a case study involving SMT from English to Bulgarian, we showed that adding features which encode not only the presence of a PV in a given phrase but also its transitiveness, separability, and idiomaticity led to better translation quality compared to previous work which employs both static and dynamic strategies.

In future research, we will extend our method to other language pairs which exhibit the same type of translation asymmetries when it comes to PVs. Such language pairs include, among others, English-Spanish and English-Portuguese.

Further, we will apply our linguistically informed method to other phenomena which cause similar issues for SMT. Immediate candidate phenomena include other types of MWEs, collocations, and noun compounds. When it comes to MWEs, we will pay special attention to the compositionality aspect since it seems to have contributed most to the good performance achieve by our method in the study presented here.

References

R H Baayen, R Piepenbrock, and L. Gulikers. 1995. The CELEX lexical database (CD-ROM).

Timothy Baldwin. 2008. A resource for evaluating the deep lexical acquisition of english verb-particle constructions. In *Proceedings of the LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions*, pages 1–2, Marakesh, Morocco.

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.

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.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. The MIT press.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Fred Karlsson, Atro Voutilainen, Juha Heikkila, and Arto Anttila. 1995. Constraint grammar: A language-independent system for parsing unrestricted text. *Natural Language Processing*, 4.

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.

Catherine Macleod, Adam Meyers, and Ralph Grishman, 1998. *COMLEX Syntax Reference Manual*. New York University.

Diana McCarthy, B Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, Sapporo, Japan.

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

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.

Aleksandar Savkov, Laska Laskova, Petya Osenova, Kiril Simov, and Stanislava Kancheva. 2011. A web-based morphological tagger for Bulgarian. In *Proceedings of the Sixth International Conference on Natural Language Processing, Multilinguality*, pages 126–137, Bratislava, Slovakia.

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.

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.