Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation

Marine Carpuat Mona Diab
Columbia University
Center for Computational Learning Systems
475 Riverside Drive, New York, NY 10115
{marine,mdiab}@ccls.columbia.edu

Abstract

We conduct a pilot study for task-oriented evaluation of Multiword Expression (MWE) in Statistical Machine Translation (SMT). We propose two different integration strategies for MWE in SMT, which take advantage of different degrees of MWE semantic compositionality and yield complementary improvements in SMT quality on a large-scale translation task.[1]

1 Introduction

A multiword expression (MWE) generally refers to a multiword unit or a collocation of words that co-occur together statistically more than chance. A MWE is a cover term for different types of collocations which vary in their transparency and fixedness. Identifying MWEs and understanding their meaning is considered essential to language understanding, and of crucial importance for any Natural Language Processing (NLP) applications that aim at handling robust language meaning and use. In fact, the seminal paper (Sag et al., 2002) refers to this problem as a key issue for the development of high-quality NLP applications. (Villavicencio et al., 2005) identify Machine Translation as an application of particular interest since " recognition of MWEs is necessary for systems to preserve the meaning and produce appropriate translations and avoid the generation of unnatural or nonsensical sentences in the target language."

However, statistical machine translation (SMT) typically does not model MWEs explicitly. SMT units are typically phrasal translations, defined without any direct syntactic or lexical semantic motivation: they are simply n-grams that are consistently translated in parallel corpora. Phrasal translations might indirectly capture MWEs, but they are not distinguished from any other n-gram.

As a result, the usefulness of explicitly modeling MWEs in the SMT framework has not yet been studied systematically. Previous work has focused on automatically learning and integrating translations of very specific MWE categories, such as, for instance, idiomatic Chinese four character expressions (Bai et al., 2009) or domain specific MWEs (Ren et al., 2009). MWEs have also been defined not from a lexical semantics perspective but from a SMT error reduction perspective, as phrases that are hard to align during SMT training (Lambert and Banchs, 2005). For each of these particular cases, translation quality improved by augmenting the SMT translation lexicon with the learned bilingual MWEs either directly or through improved word alignments.

In this paper, we consider a more general problem: we view SMT as an extrinsic evaluation of the usefulness of monolingual MWEs as used pervasively in natural language regardless of domain, idiomaticity and compositionality. A MWE is compositional if its meaning as a unit can be predicted from the meaning of its component words such as in *make a decision* meaning *to decide*. Some MWEs are more predictable than others, for instance, *kick the bucket*, when used idiomatically to mean *to die*, has nothing in common with the literal meaning of either *kick* or *bucket*, while *make a decision* is very clearly related to *to decide*. These expressions are

[1]The research was partially funded by IBM under the DARPA GALE project.

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 242–245,
Los Angeles, California, June 2010.

both considered MWEs but have varying degrees of compositionality and predictability.

We explore strategies for integrating all MWEs along this continuum in SMT. Given a monolingual MWE lexicon, we propose (1) a **static integration** strategy that segments training and test sentences according to the MWE vocabulary, and (2) a **dynamic integration** strategy that adds a new MWE-based feature in SMT translation lexicons.

In a pilot study of the impact of WordNet MWEs on a large-scale English to Arabic SMT system, we show that static and dynamic strategies both improve translation quality and that their impact is not the same for different types of MWEs. This suggests that the proposed framework would be an interesting testbed for a task-driven evaluation of automatic MWE extraction.

2 Static integration of MWE in SMT

The first strategy for integration can be seen as a generalization of word segmentation for MWEs. Given a MWE lexicon, we identify MWEs in running text and turn them into a single unit by underscoring. We call this integration method **static**, since, once segmented, all MWEs are considered frozen from the perspective of the SMT system. During training and decoding, MWEs are handled as distinct words regardless of their compositionality, and all knowledge of the MWE components is lost.

3 Dynamic integration of MWE in SMT

The second strategy attempts to encourage cohesive translations of MWEs without ignoring their components. Word alignment and phrasal translation extraction are conducted without any MWE knowledge, so that the SMT system can learn word-for-word translations from consistently translated compositional MWEs. MWE knowledge is integrated as a feature in the translation lexicon. For each entry, in addition to the standard phrasal translation probabilities, we define a count feature that represents the number of MWEs in the input language phrase.

We refer to this integration strategy as **dynamic**, because the SMT system decides at decoding time how to segment the input sentence. The MWE feature biases the system towards using phrases that do not break MWEs. This can be seen as a generalization of the binary MWE feature in (Ren et al., 2009), repurposed for monolingual MWEs.

4 Empirical Evaluation

We evaluate the impact of MWEs in SMT on a large-scale English-Arabic translation task.

Using two languages from different families is a challenging testbed for MWEs in SMT. In contrast, very closely related languages such as English and French might present less divergence in lexicalization.

In addition, Arabic-English is a well-studied language pair in SMT, with large amounts of data available. However, we tackle the less common English to Arabic direction in order to take advantage of the rich lexical resources available for English on the input side.

Our test set consists of the 813 newswire sentences of the 2008 NIST Open Machine Translation Evaluation, which is standard evaluation data for Arabic-English translation. The first English reference translation is used as the input to our SMT system, and the single Arabic translation is used as the unique reference[2]. Translation quality is evaluated using two automatic evaluation metrics: (1) BLEUr1n4 (Papineni et al., 2002), which is based on n-gram precisions for $n = 1..4$, and (2) Translation Edit Rate (TER) (Snover et al., 2006), which generalizes edit distance beyond single-word edits.

4.1 SMT system

We use the open-source Moses toolkit (Koehn et al., 2007) to build a standard phrase-based SMT system.

Our training data consists of 2.5M sentence pairs from mostly newswire parallel corpora distributed by the Linguistic Data Consortium. The English side is tokenized using simple punctuation-based rules. The Arabic side is segmented according to the Arabic Treebank v3 tokenization scheme using the MADA+TOKAN morphological analyzer and tokenizer (Habash et al., 2009).

The parallel corpus is word-aligned using GIZA++ in both translation directions, which are

[2]We exclude weblog text since it consists of an informal mix of Modern Standard Arabic and Dialectal Arabic which is suboptimal as a reference translation.

combined by intersection and the grow-diag-final-and heuristic (Koehn et al., 2007). Phrase translations of up to 10 words are extracted in the Moses phrase-table. We use a 5-gram language model with modified Kneser-Ney smoothing. Feature weights are tuned on NIST-MT06.

4.2 English MWE

Our main source of English MWE is the WordNet 3.0 lexical database (Fellbaum, 1998). We use simple rules to augment WordNet entries with morphological variations (e.g., *keep one's eyes peeled* is expanded into *keep her eyes peeled*, etc.). In addition when marking MWEs in text, we allow matches not only with surface forms, but also with lemmatized forms (Schmid, 1994) to account for inflections. This results in a total of about 900 MWE tokens and 500 types in our evaluation test set. MWE identification in running text is performed using a straightforward maximum forward match algorithm.

Second, in order to contrast the impact of MWEs with that of frequent collocations in our dynamic integration strategy, we consider the top 500 most frequent n-grams from the SMT test set, so that the same number of n-gram types and WordNet MWEs are marked in the test set. Unlike WordNet MWEs, these n-gram represent cohesive units, but are not necessarily frozen or even a single concept. We consider n-grams up to length 10 from the phrase-table, and compute their frequency in the English side of the parallel corpus. The top 500 most frequent n-grams and the WordNet MWEs yield two very different lexicons. Only the following 10 entries appear in both: *at the same time, deputy prime minister, for the first time, in the south, in the wake of, international atomic energy agency, islamic resistance movement, on the other hand, osama bin laden, secretary of state*.

5 Static MWE Integration Improves SMT

As seen in Table 1, the static integration of the WordNet MWE lexicon by segmentation of English training and test sentences improves BLEU and TER compared to the SMT baseline. This suggests that WordNet MWEs represent useful units of meaning for alignment and translation into Arabic despite the fact that they are monolingually defined.

MWE	integration	TER	BLEU
Baseline	—	59.43	30.49
Top 500 n-grams	dynamic	59.07	30.98
WordNet MWE	dynamic	58.89	31.07
WordNet MWE	static	58.98	31.27

Table 1: Impact of MWE integration measured on NIST MT08

Consider, for instance, the following input sentence: *the special envoy of the secretary-general will submit an oral report to the international security council rather than a written report*. With static integration, the MWE *written report* is correctly translated as *tqryrA mktwbA*, while the baseline produces the incorrect translation *ktb Altqryr* (*writing the report* or *book of report*).

6 Dynamic MWE Integration Improves SMT

Dynamic integration of the WordNet MWE lexicon and the top 500 n-grams both improve BLEU and TER (Table 1), but WordNet MWEs yield slightly better scores. This confirms the ability of the dynamic integration method to handle compositional MWEs, since the most frequent n-grams are highly compositional by definition.

7 Discussion

At the corpus-level, static integration yields a slightly better BLEU score than dynamic with WordNet MWEs, while the opposite effect is observed on TER. This suggests that the two integration strategies impact translation in different ways. Sentence-level scores indeed reveal that dynamic and static integration strategies have an opposite impact on 27% of the test set (Table 2).

For instance, the dynamic approach fails for phrasal verbs such as *take out*. In *who were then allowed to take out as many unsecured loans as they wanted*, *take out* is realized as *b+ AlHSwl* (*acquire*) with the static approach, while it is entirely dropped from the dynamic translation.

In the static approach, translation quality is often degraded when our simple dictionary matching approach incorrectly detects MWE. For instance, in the sentence *the perpetration of this heinous act on our*

Dynamic integration	helps	hurts
Static integration		
helps	45%	**16%**
hurts	**11%**	28%

Table 2: Percentage of sentences where each integration strategy helps or hurts both BLEU and TER compared to the baseline SMT system.

soil, *act on* is incorrectly identified as a MWE which degrades translation fluency. This suggests that further gains in translation quality could be obtained with a more sophisticated MWE detection method.

8 Conclusion

We have proposed a framework of two complementary integration strategies for MWEs in SMT, which allows extrinsic evaluation of the usefulness of MWEs of varying degree of compositionality. We conducted a pilot study using manually defined WordNet MWE and a dictionary matching approach to MWE detection. This simple model improves English-Arabic translation quality, even on a large SMT system trained on more than 2 Million sentence pairs.

This result suggests that standard SMT phrases do not implicitly capture all useful MWE information. It would therefore be interesting to conduct this study on a larger scale, using more general MWE definitions such as automatically learned collocations (Smadja, 1993) or verb-noun constructions (Diab and Bhutada, 2009).

References

Ming-Hong Bai, Jia-Ming You, Keh-Jiann Chen, and Jason S. Chang. 2009. Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 478–486, Singapore, August.

Mona Diab and Pravin Bhutada. 2009. Verb noun construction MWE token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 17–22, Singapore, August.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt.

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 Herbst. 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.

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

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 of the Association for Computational Linguistics*.

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

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 1–15, London, UK. Springer-Verlag.

Helmut Schmid. 1994. Probabilistic part–of–speech tagging using decision trees. In *Proceedings of the Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Frank A. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231, Boston, MA. Association for Machine Translation in the Americas.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, 19(4):365 – 377. Special issue on Multiword Expression.