Evaluation of a Substitution Method for Idiom Transformation in Statistical Machine Translation

Giancarlo D. Salton and **Robert J. Ross** and **John D. Kelleher**
Applied Intelligence Research Centre
School of Computing
Dublin Institute of Technology
Ireland
giancarlo.salton@mydit.ie {robert.ross,john.d.kelleher}@dit.ie

Abstract

We evaluate a substitution based technique for improving Statistical Machine Translation performance on idiomatic multiword expressions. The method operates by performing substitution on the original idiom with its literal meaning before translation, with a second substitution step replacing literal meanings with idioms following translation. We detail our approach, outline our implementation and provide an evaluation of the method for the language pair English/Brazilian-Portuguese. Our results show improvements in translation accuracy on sentences containing either morphosyntactically constrained or unconstrained idioms. We discuss the consequences of our results and outline potential extensions to this process.

1 Introduction

Idioms are a form of figurative multiword expressions (MWE) that are ubiquitous in speech and written text across a range of discourse types. Idioms are often characterized in terms of their having non-literal and non-compositional meaning whilst occasionally sharing surface realizations with literal language uses (Garrao and Dias, 2001). For example the multiword expression *s/he took the biscuit* can have both a figurative meaning of being (pejoratively) remarkable, and a literal meaning of removing the cookie.

It is notable that idioms are a compact form of language use which allow large fragments of meaning with relatively complex social nuances to be conveyed in a small number of words, i.e., idioms can be seen as a form of compacted regularized language use. This is one reason why idiom use is challenging to second language learners (see, e.g., Cieslicka(2006)).

Another difficulty for second language learners in handling idioms is that idioms can vary in terms of their morphosyntactic constraints or *fixedness* (Fazly et al., 2008). On one hand some idiomatic expressions such as *popped the question* are highly fixed with syntactic and lexical variations considered unacceptable usage. On the other hand idioms such as *hold fire* are less fixed with variations such as *hold one's fire* and *held fire* considered to be acceptable instances of the idiom type.

For reasons such as those outlined above idioms can be challenging to human speakers; but they also pose a great challenge to a range of Natural Language Processing (NLP) applications (Sag et al., 2002). While idiomatic expressions, and more generally multiword expressions, have been widely studied in a number of NLP domains (Acosta et al., 2011; Moreno-Ortiz et al., 2013), their investigation in the context of machine translation has been more limited (Bouamor et al., 2011; Salton et al., 2014).

The broad goal of our work is to advance machine translation by improving the processing of idiomatic expressions. To that end, in this paper we introduce and evaluate our initial approach to the problem. We begin in the next section by giving a brief review of the problem of idiom processing in a Statistical Machine Translation (SMT) context. Following that we outline our substitution based solution to idiom processing in SMT. We then outline a study that we have conducted to evaluate our initial method. This is followed with results and a brief discussion before we draw conclusions and outline future work.

2 Translation & Idiomatic Expressions

The current state-of-the-art in machine translation is phrase-based SMT (Collins et al., 2005). Phrase-based SMT systems extend basic word-by-word SMT by splitting the translation process into 3 steps: the input source sentence is segmented

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 38–42,
Gothenburg, Sweden, 26-27 April 2014.

into "phrases" or multiword units; these phrases are then translated into the target language; and finally the translated phrases are reordered if needed (Koehn, 2010). Although the term phrase-based translation might imply the system works at the semantic or grammatical phrasal level, it is worth noting that the concept of a phrase in SMT is simply a frequently occurring sequence of words. Hence, standard SMT systems do not model idioms explicitly (Bouamor et al., 2011).

Given the above, the question arises as to how SMT systems can best be enhanced to account for idiom usage and other similar multiword expressions. One direct way is to use a translation dictionary to insert the idiomatic MWE along with its appropriate translation into the SMT model phrase table along with an estimated probability. While this approach is conceptually simple, a notable drawback with such a method is that while the MWEs may be translated correctly the word order in the resulting translation is often incorrect (Okuma et al., 2008).

An alternative approach to extending SMT to handle idiomatic and other MWEs is to leave the underlying SMT model alone and instead perform intelligent pre- and post-processing of the translation material. Okuma et al. (2008) is an example of this approach applied to a class of multi- and single word expressions. Specifically, Okuma et al. (2008) proposed a substitution based pre and post processing approach that uses a dictionary of *surrogate words* from the same word class to replace low frequency (or unseen) words in the sentences before the translation with high frequency words from the same word class. Then, following the translation step, the *surrogate words* are replaced with the original terms. Okuma et al.'s direct focus was not on idioms but rather on place names and personal names. For example, given an English sentence containing the relatively infrequent place name ***Cardiff***, Okuma et al.'s approach would: (1) replace this low frequency place name with a high frequency *surrogate* place name, e.g. ***New York***; (2) translate the updated sentence; and (3) replace the *surrogate words* with the correct translation of the original term.

The advantage of this approach is that the word order of the resulting translation has a much higher probability of being correct. While this method was developed for replacing just one word (or a highly fixed name) at a time and those words must be of the same open-class category, we see the basic premise of pre- and post- substitution as also applicable to idiom substitution.

3 Methodology

The hypothesis we base our approach on is that the work-flow that a human translator would have in translating an idiom can be reproduced in an algorithmic fashion. Specifically, we are assuming a work-flow whereby a human translator first identifies an idiomatic expression within a source sentence, then 'mentally' replaces that idiom with its literal meaning. Only after this step would a translator produce the target sentence deciding whether or not to use an idiom on the result. For simplicity we assumed that the human translator should use an idiom in the target language if available. While this work-flow is merely a proposed method, we see it as plausible and have developed a computational method based on this work-flow and the substitution technique employed by (Okuma et al., 2008).

Our idiom translation method can be explained briefly in terms of a reference architecture as depicted in Figure 1. Our method makes use of 3 dictionaries and 2 pieces of software. The first dictionary contains entries for the source language idioms and their literal meaning, and is called the "Source Language Idioms Dictionary". The second dictionary meanwhile contains entries for the target language idioms and their literal meaning, and is called the "Target Language Idioms Dictionary". The third dictionary is a bilingual dictionary containing entries for the idioms in the source language pointing to their translated literal meaning in the target language. This is the "Bilingual Idiom Dictionary".

The two pieces of software are used in the pre- and post-processing steps. The first piece of software analyzes the source sentences, consulting the "Source Language Idioms Dictionary", to identify and replace the source idioms with their literal meaning in the source language. During this first step the partially rewritten source sentences are marked with replacements. Following the subsequent translation step the second piece of software is applied for the post-processing step. The software first looks into the marked sentences to obtain the original idioms. Then, consulting the "Bilingual Idiom Dictionary", the software tries to match a substring with the literal translated mean-

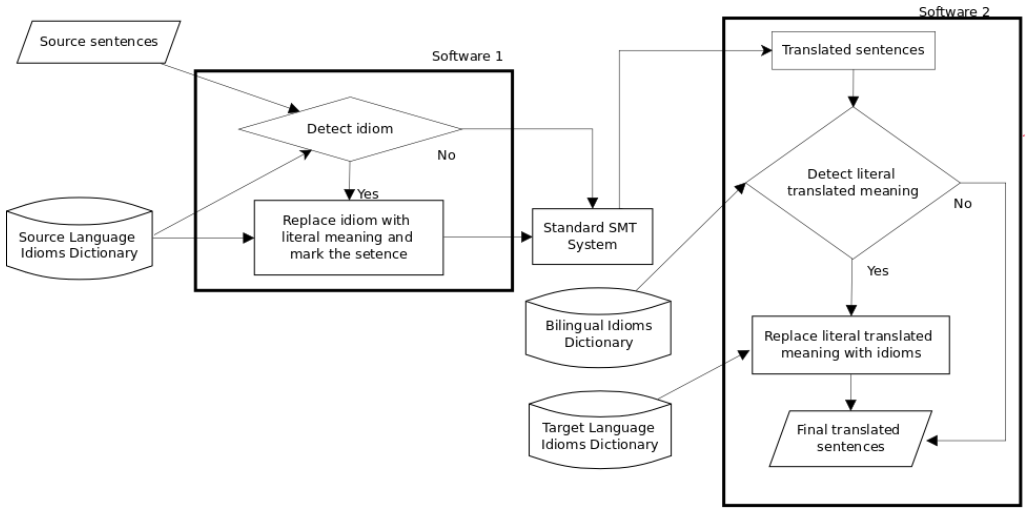

Figure 1: Reference Architecture for Substitution Based Idiom Translation Technique.

ing in the target translation. If the literal meaning is identified, it then checks the "Target Language Idioms Dictionary" for a corresponding idiom for the literal use in the target language. If found, the literal wording in the target translation is then replaced with an idiomatic phrase from the target language. However if in the post-processing step the original idiom substitution is not found, or if there are no corresponding idioms in the target language, then the post-processing software does nothing.

4 Study Design

We have developed an initial implementation of our substitution approach to SMT based idiom translation for the language pair English/Brazillian-Portugese. To evaluate our method we created test corpora where each sentence contained an idiom, and compared the BLEU scores (Papineni et al., 2002) of a baseline SMT system when run on these test corpora with the BLEU scores for the same SMT system when we applied our pre and post processing steps. No sentences with literal uses of the selected idiom form were used in this experiment.

Consequently, three corpora were required for this experiment in addition to the three idiomatic resources introduced in the last section. The first corpus was an initial large sentence-aligned bilingual corpus that was used to build a SMT model for the language pair English/Brazilian-Portuguese. The second corpus was the first of two test corpora. This corpus contained sentences with "highly fixed" idioms and will be referred to as the "High Fixed Corpus". Finally a second test corpus containing sentences with "low fixed" idioms, the "Low Fixed Corpus", was also constructed. In order to make results comparable across test corpora the length of sentences in each of the two test corpora were kept between fifteen and twenty words.

To create the initial large corpus a series of small corpora available on the internet were compiled into one larger corpus which was used to train a SMT system. The resources used in this step were Fapesp-v2 (Aziz and Specia, 2011), the OpenSubtitles2013[1] corpus, the PHP Manual Corpus[2] and the KDE4 localizaton files (v.2)[3]. No special tool was used to clean these corpora and the files were compiled as is.

To create the "High Fixed Corpus" and "Low Fixed Corpus" we built upon the previous work of Fazly et al. (2008) who identified a dataset of 17 "highly fixed" English verb+noun idioms, and 11 "low fixed" English verb+noun idioms. Based on these lists our two test corpora were built by extracting English sentences from the internet which contained instances of each of the high and low fixed idiom types. Each collected sentence was manually translated into Brazilian-Portuguese, before each translations was manually checked and corrected by a second translator. Ten sentences were collected for each idiom type. This resulted in a High Fixed corpus consisting of 170 sentences

[1] http://opus.lingfil.uu.se/OpenSubtitles2013.php
[2] http://opus.lingfil.uu.se/PHP.php
[3] http://opus.lingfil.uu.se/KDE4.php

containing idiomatic usages of those idioms, and a Low-Fixed corpus consisting of 110 sentences containing instances of low-fixed idioms.

As indicated three idiomatic resources were also required for the study. These were: a dictionary of English idioms and their literal meanings; a dictionary of Brazilian-Portuguese idioms and their literal meanings; and a bilingual dictionary from English to Brazilian-Portuguese. The English idioms dictionary contained entries for the idioms pointing to their literal English meanings, along with some morphological variations of those idioms. The Brazilian-Portuguese idioms dictionary similarly contained entries for the idioms pointing to their literal meanings with some morphological variations of those idioms. Finally, the bilingual dictionary contained entries for the same idioms along with morphological variations of the English idioms dictionary but pointing to their literal translated meaning. The Oxford Dictionary of English idioms and the Cambridge Idioms Dictionary were used to collect the literal meanings of the English idioms. Literal meanings were manually translated to Brazilian-Portuguese.

Following resource collection and construction a SMT model for English/Brazilian-Portuguese was trained using the Moses toolkit (Koehn et al., 2007) using its baseline settings. The corpus used for this training consisted of 17,288,109 pairs of sentences (approximately 50% of the initial collected corpus), with another 34,576 pairs of sentences used for the "tuning" process. Following this training and tuning process the baseline translation accuracy, or BLEU scores, were calculated for the two test corpora, i.e., for the "High Fixed Corpus", and the "Low Fixed Corpus".

Having calculated the baseline BLEU scores, the substitution method was then applied to retranslate each of the two test corpora. Specifically both the "High Fixed Corpus" and the "Low Fixed Corpus" were passed through our extended pipeline with new substitution based translations constructed for each of the test corpora. BLEU scores were then calculated for these two output corpora that were built using the substitution method.

5 Results and Discussion

Table 1 presents the results of the evaluation. The BLEU scores presented in the table compare the baseline SMT system against our proposed method for handling English idiomatic MWE of the verb+noun type.

Corpus	Baseline	Substitution
High Idiomatic	23.12	31.72
Low Idiomatic	24.55	26.07

Table 1: Experiment's results.

Overall the results are positive. For both the high and low idiomatic corpora we find that applying the pre- and post-processing substitution approach improves the BLEU score of the SMT system. However, it is notable that the High-Fixed idiomatic corpus showed a considerably larger increase in BLEU score than was the case for the Low-Fixedness idiomatic cases, i.e., a positive increase of 8.6 versus 1.52. To investigate further we applied a paired t-test to test for significance in mean difference between baseline and substitution methods for both the high-fixed and low-fixed test corpora. While the results for the "High Idiomatic Corpus" demonstrated a statistically significant difference in BLEU scores ($p \ll 0.05$), the difference between the baseline and substitution method was not statistically significant for the case of the "Low Idiomatic Corpus" ($p \approx 0.7$). We believe the lack of improvement in the case of low fixed idioms may be caused by a higher morphosyntactic variation in the translations of the low fixed idioms. This higher variation makes the post-processing step of our approach (which requires matching a substring in the translated sentence) more difficult for low fixed idioms with the result that our approach is less effective for these idioms.

It is worth noting that the same SMT system (without the substitution extension) achieved a BLEU score of 62.28 on a corpus of sentences from general language; and, achieved an average BLEU score of 46.48 over a set of 5 corpora of sentences that did not contain idioms and were of simlar length to the idiomatic corpora used in this study (15 to 20 words). Both these BLEU scores are higher than the scores we report in Table 1 for our substitution method. This indicates that although our substitution approach does improve BLEU scores when translating idioms there is still a lot of work to be done to solve the problems posed by idioms to SMT.

6 Conclusion

Our results indicate that this substitution approach does improve the performance of the system. However, we are aware that this method is not the entire solution for the MWE problem in SMT. The effectiveness of the approach is dependent on the fixedness of the idiom being translated.

This approach relies on several language resources, including: idiomatic dictionaries in the source and target languages and a bilingual dictionary containing entries for the idioms in the source language aligned with their translated literal meaning in the target language. In future work we investigate techniques that we can use to (semi)automatically address dictionary construction. We will also work on enabling the system to distinguish between idiomatic vs. literal usages of idioms.

Acknowledgments

Giancarlo D. Salton would like to thank CAPES ("Coordenação de Aperfeiçoamento de Pessoal de Nível Superior") for his Science Without Borders scholarship, proc n. 9050-13-2. We would like to thank Acassia Thabata de Souza Salton for her corrections on the Brazilian-Portuguese translation of sentences containing idioms.

References

Otavio Costa Acosta, Aline Villavicencio, and Viviane P. Moreira. 2011. Identification and Treatment of Multiword Expressions applied to Information Retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, pages 101–109.

Wilker Aziz and Lucia Specia. 2011. Fully automatic compilation of a portuguese-english and portuguese-spanish parallel corpus for statistical machine translation. In *STIL 2011*.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2011. Improved Statistical Machine Translation Using MultiWord Expressions. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation)*, pages 15–20.

Anna Cieślicka. 2006. Literal salience in on-line processing of idiomatic expressions by second language learners. 22(2):115–144.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 531–540.

Afsanesh Fazly, Paul Cook, and Suzanne Stevenson. 2008. Unsupervised Type and Token Identification of Idiomatic Expressions. In *Computational Linguistics*, volume 35, pages 61–103.

Milena U. Garrao and Maria C. P. Dias. 2001. Um Estudo de Expressões Cristalizadas do Tipo V+Sn e sua Inclusão em um Tradutor Automático Bilíngüe (Português/Inglês). In *Cadernos de Tradução*, volume 2, pages 165–182.

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 *45th Annual Meeting of the Association for Computational Linguistics*.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York. 2 Ed.

Antonio Moreno-Ortiz, Chantal Pérez-Hernández, and M. Ángeles Del-Olmo. 2013. Managing Multiword Expressions in a Lexicon-Based Sentiment Analysis System for Spanish. In *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013)*, pages 1–10.

Hideo Okuma, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Introducing Translation Dictionary Into Phrase-based SMT. In *IEICE - Transactions on Information and Systems*, number 7, pages 2051–2057.

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 (ACL)*, pages 311–318.

Ivan A. Sag, Thimothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002, Lecture Notes in Computer Science*, volume 2276, pages 1–15.

Giancarlo D. Salton, Robert J. Ross, and John D. Kelleher. 2014. An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese. In *Third Workshop on Hybrid Approaches to Translation (HyTra) at 14th Conference of the European Chapter of the Association for Computational Linguistics*.