An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese

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

This paper describes an experiment to evaluate the impact of idioms on Statistical Machine Translation (SMT) process using the language pair English/Brazilian-Portuguese. Our results show that on sentences containing idioms a standard SMT system achieves about half the BLEU score of the same system when applied to sentences that do not contain idioms. We also provide a short error analysis and outline our planned work to overcome this limitation.

1 Introduction and Motivation

An idiom is an expression whose meaning is not compositional (Xatara, 2001). In other words the meaning of an idiom is not simply the joint meaning of the individual words (Garrao and Dias, 2001). For example, the expression *kick the bucket* has an idiomatic meaning (*to die*) that has nothing to do with the meaning of *kick* or *bucket*.

Idioms are a type of multi-word expressions (MWEs) often used in a large variety of texts and by human speakers and thus appear in all languages (Fazly et al., 2008). Consequently, they pose problems to most Natural Language Processing (NLP) applications (Sag et al., 2002). Nevertheless, they often have been overlooked by researchers in NLP (Fazly et al., 2008).

As a class, idioms exhibit a number of properties that make them difficult to handle for NLP applications. For example, idiomatic expressions vary with respect to how morphosyntatically fixed they are. An idiomatic expression is highly fixed if the replacement of any of its constituents by a, syntactically or semantically, similar word causes the idiomatic meaning of the expression to be lost (Fazly et al., 2008). An example of a highly fixed idiom in English is the expression *by and large*. Idioms that are highly fixed can be represented as words-with-spaces by an NLP system (Sag et al., 2002). If, however, an idiomatic meaning persists across morphosyntactic variations of an expression, the idiom can be described as a low fixed idiom, for example, *hold fire* and its variations *hold one's fire* and *held fire*. The words-with-spaces approach does not work for these "more flexible" example of idioms (Fazly et al., 2008). Another feature of idioms that make them difficult for NLP system to process is that idiomatic expressions have both idiomatic and literal (non-idiomatic) usages. Consequently, NLP systems need to distinguish between these types of usages (Fazly et al., 2008).

One of the most important NLP applications that is negatively affected by idioms is Statistical Machine Translation (SMT) systems. The current state-of-the-art in SMT are phrase-based systems (Collins et al., 2005). Phrase-based SMT systems extend the basic SMT word-by-word approach by splitting the translation process into 3 steps: the input source sentence is segmented into "phrases" or multi-word units; these phrases are translated into the target language; and the translated phrases are reordered if needed (Koehn, 2010).

It is worth highlighting that although the term phrase-based translation seems to imply the system works at a phrasal level, the concept of a phrase to these systems is simply a frequently occurring sequence of words and not necessarily a semantic or grammatical phrase. These systems thus limit themselves to a direct translation of phrases without any syntactic or semantic context. Hence, standard phrase-based SMT systems do not model idioms explicitly (Bouamor et al., 2011). Unfortunately modelling idioms in order to improve SMT is not well studied (Ren et al., 2009) and examples of the difficulties in translating these expressions can be seen in the quality of the resultant output of most Machine Translation

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 36–41,
Gothenburg, Sweden, April 27, 2014.

systems (Vieira and Lima, 2001).

Our long-term research goal is to investigate how the translation of idiomatic expressions may be improved. We will initially focus on the case of English/Brazillian-Portugese but we intend our work to be generalizable to other language pairs. As a first step on this research program we wished to scope the impact of idioms on an SMT system. In order to test this we ran an experiment that compared the BLEU scores of an SMT system when it was tested on three distinct sentence aligned corpora. Two of these test corpora consisted of sentences containing idiomatic (rather than literal) usages of idiomatic expressions and the other corpus consisted of sentences that did not contain any idioms. By comparing the BLEU score of a machine translation system on each of these corpora we hoped to gauge the size of the research problem we are addressing.

The paper is organized as follows: Section 2 describes the design and creation of the corpora used in the experiments; Section 3 presents the experiment's methodology; Section 4 reports the results found; and Section 5 both discusses the results and describes an approach to the problem that we will implement in future work.

2 Related work

The work of Fazly et al. (2008) has provided an inspirational basis for our work. Fazly's work focused on the study of idioms and in particular their identification and analysis in terms of the syntactic and semantic fixedness. Fazly study did not however explore the impact of idioms on SMT.

Some related work in translating idioms can be found in: Garrao and Dias (2001) where the verb+noun combinations and their inclusion in an online automatic translator is explored; Ren et al. (2009) which makes use of a domain constrained bilingual multi-word dictionary to improve the MT results; Bouamor et al. (2011) which explores a hybrid approach for extracting MWEs and their translation in a French-English corpus; and Bungum et al. (2013) which also uses dictionaries to capture MWEs.

None of these works compares the BLEU score of sentences containing and not containing idioms. And also, none of these works address the idioms problem for the English/Brazilian-Portuguese language pair using SMT phrase-based systems.

3 Corpora Design and Collection

The experiment we describe in this paper had two direct targets: (a) we wished to quantify the effect of idioms on the performance of an SMT system; and (b) we wanted to better understand the differences (if any) between high and low fixed idioms with respect to their impact on SMT systems. Consequently, in order to run the experiments four corpora were needed: one initial large sentence-aligned bilingual corpus was needed to build an SMT model for the language pair English/Brazilian-Portuguese; a test corpus containing sentences with "highly fixed" idioms called the "High Idiomatic Corpus"; another test corpus containing sentences with "low fixed" idioms called the "Low Idiomatic Corpus"; and a last corpus with sentences not containing idioms called the "Clean Corpus". In order to make the results comparable the length of each sentence in the three test corpora was kept between 15 to 20 words. All of these corpora were constructed by hand and in the cases of the "High Idiomatic Corpus" and "Low Idiomatic Corpus" care was taken to ensure that all the sentences in these corpora contained idiomatic usages of the relevant idioms.

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 the 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.

Idioms are a heterogeneous class; consequently, in order to better control the experiment we decided to focus on a particular type of idiom - specifically the idiomatic expressions formed from the combination of a verb and a noun as its direct object (verb+noun combinations), for example *hit+road* and *lose+head*. Verb+noun combinations are a subclass of MWE which are notable for their cross-lingual occurrence and high variability, both lexical and semantic (Baldwin and Kim, 2010). Also, it is worth noting that it is possible for a particular verb+noun combination to have both idiomatic and literal usages and these usages must be distinguished if an NLP system is to pro-

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

cess a sentence appropriately.

Fazly et al. (2008) named a dataset of 17 "highly fixed" English verb+noun idioms, for example *cut+figure*, and that list was used to build our "Highly Idiomatic Corpus". This corpus consisted of 170 sentences containing idiomatic usages of these idioms, 10 sentences per idiom in the list. These English sentences were collected from the internet and manually translated into Brazilian-Portuguese. After that these translations were then manually checked and corrected by a second translator.

Fazly et al. (2008) also named a dataset of 11 "low fixed" English verb+noun idioms, for example *get+wind*, and that list was used to build our "Low Idiomatic Corpus". This corpus consisted of 110 sentences containing idiomatic usages of these idioms, 10 sentences per idiom in the list. These English sentences were also collected from the internet and manually translated into Brazilian-Portuguese. After this step these translations were also manually checked and corrected by a second translator. Table 1 presents the English verb+noun combinations used in this experiment and their Brazilian-Portuguese translations.

In order to have a valid comparison between the translation results of sentences containing and not containing idioms the "Clean Corpus" was built. It consisted of 850 sentences with their translations and was created by sampling sentences of the appropriate length (15-20 words) that did not contain idioms from the large bilingual corpus (that we described earlier) which we created to train the SMT system. These sentences were then removed from that corpus. Because the initial corpus was created from the union of corpora from different domains the "Clean Corpus" was randomly split into 5 datasets containing 170 sentences each in order to ensure no specific influence of any of those domains on the BLEU score. We called these "Clean1" to "Clean5". Special care was taken to not have any idioms in any of the sentences in these corpora.

As we wanted to collect 10 sentences for each verb+noun idiomatic combination and due to the limitations of sentence length (15 to 20 words) we were not able to collect the "High Idiomatic Corpus" and the "Low Idiomatic Corpus" from the training corpus. Thus, the samples were collected from the Internet.

4 Methodology

As a first step for this experiment, a SMT model for the English/Brazilian-Portuguese language pair was trained using the Moses toolkit (Koehn et al., 2007) following its "baseline" settings (Koehn et al., 2008). 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.

English	**Brazilian-Portuguese**
blow+top	perder+paciência
blow+trumpet	*"gabar-se"*
cut+figure	causar+impressão
find+foot	*"adaptar-se"*
get+nod	*"obter permissão"*
give+sack	*"ser demitido", "demitir"*
have+word	ter+conversa
hit+road	*"cair na estrada"*
hit+roof	*"ficar zangado"*
kick+heel	*"deixar esperando"*
lose+thread	*"perder o fio da meada"*
make+face*	fazer+careta
make+mark	deixar+marca
pull+plug	*"cancelar algo"*
pull+punch	*"esconder algo"*
pull+weight	*"fazer sua parte"*
take+heart	*"ficar confiante"*
blow+whistle	*"botar a boca no trombone"'*
get+wind	ouvir+murmúrios
hit+wall	*"dar de cara num muro"*
hold+fire	*"conter-se"*
lose+head*	perder+cabeça
make+hay	dar+graças
make+hit	fazer+sucesso
make+pile	fazer+grana
make+scene*	fazer+cena
pull+leg	pegar+pé
see+star*	ver+estrela

Table 1: The English verb+noun combinations used in this experiment and their Brazilian-Portuguese Translations. The idioms marked with an * have direct translations of its constituents resulting in a MWE with the same idiomatic meaning in Brazilian-Portuguese. Also, note that not all translations results in a verb+noun idiom in the target language. Those are presented between double quotes and italics.

In the second step the BLEU scores for the "High Idiomatic Corpus", the "Low Idiomatic Corpus" and the five clean corpora were computed. Then, the average of each evaluation for the clean corpora was calculated.

5 Results and Analysis

Table 2 lists the SMT system BLEU scores for the "High Idiomatic Corpus", "Low Idiomatic Corpus", and the average BLEU score for the clean corpora (i.e, "Clean1" to "Clean5"). The differential between the BLEU scores for the clean corpus and the idiomatic corpora (high and low) indicates that English idiomatic expressions of the verb+noun type pose a significant challenge to standard phrase based SMT.

Corpus	BLEU scores
High Idiomatic	23.12
Low Idiomatic	24.55
Clean (average)	46.28

Table 2: BLEU scores.

The corpora containing idioms achieved only half of the average Clean Corpus score. As noted earlier, some idioms have a direct translation from English to Brazilian-Portuguese and could result in straight forward translations that the basic SMT system (without substitution) can handle correctly. Given this, the BLEU scores for this subset of idioms could be expected to be similar to the clean corpus results. However, it is worth noting that even for idioms that have direct translations, see Table 1, the BLEU score for the sentences containing these idioms is still lower than average BLEU score for the clean corpus. Using the Student's *t*-test, we found a statistical difference between the "Low Idiomatic Corpus" and the "Clean Corpus" ($p \ll 0$), and between the "High Idiomatic Corpus" and the "Clean Corpus" ($p \ll 0$).

The second question that we examined in the experiment was whether there was a difference in performance between the high and low fixed idioms. Table 3 lists the BLEU scores for each of the "highly fixed" verb+noun combinations used in the "High Idiomatic Corpus" and Table 4 lists the BLEU scores for each of the "low fixed" verb+noun combinations from the "Low Idiomatic Corpus". Also, it is important to note that the "High Idiomatic Corpus" and the "Low Idiomatic Corpus" have almost no difference in their BLEU scores. We also found that there are almost no statistical difference ($p = 0.85$) between the "High Idiomatic Corpus" and "Low Idiomatic Corpus" which we believe indicates that both kinds of verb+noun idiomatic combinations pose the same problem to SMT.

"high fixed" verb+noun	BLEU score
blow+top	22.08
blow+trumpet	19.38
cut+figure	20.15
find+foot	24.36
get+nod	22.06
give+sack	23.03
have+word	20.91
hit+road	24.53
hit+roof	21.34
kick+heel	18.85
lose+thread	21.81
make+face	28.62
make+mark	29.46
pull+plug	19.71
pull+punch	28.34
pull+weight	19.94
take+heart	23.41

Table 3: BLEU scores for individual "high fixed" verb+noun idiomatic combinations.

"low fixed" verb+noun	BLEU score
blow+whistle	17.75
get+wind	19.06
hit+wall	16.52
hold+fire	23.26
lose+head	37.40
make+hay	15.87
make+hit	25.48
make+pile	25.31
make+scene	36.93
pull+leg	15.90
see+star	37.86

Table 4: BLEU scores for individual "low fixed" verb+noun idiomatic combinations.

6 Conclusions and Future Work

Certainly, these results are not surprising. BLEU scores are generally dependent on the training and test corpora; that said, it it worthwhile having a quantification of the potential issues that idioms pose for SMT. Due to the fact that BLEU scores

are dependent on the training and test corpora used our results are corpus specific. However, these results are our starting point to develop a hybrid methodology.

As noted earlier, idioms are widely used in every literary genre and new expressions come into existence frequently. Thus, they must be properly handled and translated by a Machine Translation system. Given the results of our experiments it is evident that the problem in translating idioms has not been solved using a standard SMT system. Such evidences and the relatively small amount of current related work on idiomatic expression translation, when compared with the amount of work on other MT aspects, indicates that there is likely not a trivial solution.

To start addressing these problems, we propose a hybrid method inspired by the work developed by Okuma et al. (2008) for translating unseen words using bilingual dictionaries.

Our method, introduced in Salton et al. (2014), work as a pre and post-processing step. We first identify idioms in source sentences using an idiom dictionary. Then, we substitute the idiom in the source sentence with its literal meaning, taken from the dictionary and record the fact that this sentence contained a substituted idiom. For all sentences that are recorded as containing a substitution, after the translation we check if the original idiom that occurred in the source sentence has a corresponding idiom in the target language by consulting a separate bilingual dictionary. If there is a corresponding idiom in the target language then the translation of the literal meaning of the source language idiom is replaced with the target language idiom. If there are no related idioms on the target language, this post-processing step is avoided and the translation is done.

This approach relies on a number of dictionaries being available. Developing these resources is non-trivial and in order to scale our approach to broad coverage a large part of our future work will focus on automating (as much as possible) the development of these language resources. Another problem that we will address in future work is ensuring that we apply substitution appropriately. There are at least two situations where care must be taken. First, a given expression may be used both as an idiom and literally. Consequently, we need to develop mechanisms that will enable our preprocessing step to distinguish between idiomatic and non-idiomatic usages. Second, some idiomatic expressions have direct translations. For these expressions we expect that the substitution method may under-perform the standard SMT system. Ideally, we would like to be able to control the substitution method so that these particular expressions are allowed through the preprocessing and are handled by the standard SMT pipeline. However, for now, considering the proportion of expressions with direct translations in comparison with the overall number of expressions is very low; we hope that this problem will not have too adverse an impact on our approach. Beyond these issues, while we anticipate that our substitution based approached will work reasonably well for "high fixed" idioms, we are aware that the variation in "low fixed" idioms may require us to extend the system in order to handle this variation.

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

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*.

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group.

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.

Lars Bungum, Björn Gambäck, André Lynum, and Erwin Marsi. 2013. Improving Word Translation Disambiguation by Capturing Multiword Expressions with Dictionaries. In *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013)*, pages 21–30.

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, Abhishek Arun, and Hieu Hoang. 2008. Towards better machine translation quality for the German-English language pairs. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142.

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

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.

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 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pages 47–54.

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. Evaluation of a Substitution Method for Idiom Transformation in Statistical Machine Translation. In *The 10th Workshop on Multiword Expressions (MWE 2014) at 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Renata Vieira and Vera Lcia S. Lima. 2001. Linguística Computacional: Princípios e aplicações. In *Ana Teresa Martins & Díbio Leandro Borges (eds.), As Tecnologias da informação e a questão social: anais*.

Cláudia M. Xatara. 2001. O Ensino do Léxico: As Expressões Idiomáticas. In *Trabalhos em Linguística Aplicada*, volume 37, pages 49–59.