Idioms in state-of-the-art Croatian-English and English-Croatian SMT systems

Maja Manojlović*, Luka Dajak*, Marija Brkić Bakarić**
*Faculty of Humanities and Social Sciences, Rijeka, Croatia
**Department of Informatics, Rijeka, Croatia
*mmmanojlovic@ffri.hr, ldajak@ffri.hr
**mbrkic@inf.uniri.hr

Abstract - Idioms are well known for posing problems to non-native speakers, let alone machines. A failure to identify idioms often leads to unnatural, even hilarious outputs. This paper investigates the treatment of idioms in state-of-the-art SMT systems involving English and Croatian. First we introduce the concept of idioms. Then we construct three short stories abundant with idioms per each language, and translate them into the other language by two state-of-the-art SMT systems. Next we manually inspect the outputs and present results. For the purpose of conducting analysis, we devise an error taxonomy for handling idioms.

I. INTRODUCTION

The quality of translation depends on the language pair, the type of text being translated, and the system used for translation [1]. Idioms are fixed expressions that do not convey literal meaning of their constituents. Words in their structure lose their original meaning and, together with other words within the idiom, convey new meaning. Although words within an idiom do not lose their meaning completely, they get a new context and background. These words must often be in the same order and they cannot be omitted nor replaced with other words. Idioms are comprised of at least two constituent words that are characterized by their common use, integrity and relatively fixed structure. They do not develop in spoken communication, but are integrated in discourse as a whole. This way, they either become a part of a sentence structure or they function as an independent unit [2].

Idioms will always pose a problem in translating, even to humans. Therefore, it is plausible that machines will come across even greater obstacles.

The aim of the research is to compare the way machines translate texts, more precisely idioms, in English-Croatian and Croatian-English language directions. Two online translation systems are used for the purpose of this research: Google Translate (hereinafter GT) and Asistent [3], both phrase-based statistical machine translation (PBSMT) systems. Croatian-English translations are expected to be of higher quality since English, unlike Croatian, is well resourced [4].

Section two clarifies the specificities of idioms. Section three gives an overview of related work. Section four describes our experimental study. The results are presented in section five. Section six highlights the main findings of our work.

II. IDIOMS

According to their structure, idioms can be divided into three categories. First category is comprised of sequences of words that consist of at least two independent words (e.g. gold mine). Second category are phonetic words. These sets of words consist of only one independent unit (e.g. out of sight). Finally, third category is comprised of idioms that acquired a form of a sentence (e.g. the place is getting too hot).

Idioms can also be categorized in accordance to their origin. First category, thus, would consist of biblical idioms. Considering the importance of the Bible and its tremendous influence since the beginning of the Middle ages, it is not surprising that biblical idioms found their equivalents in various languages. Examples of this category of idioms include: Judas kiss, the lost sheep, to cast pearls before swine. Second category consists of literature idioms. These are idioms that have origin in literature or mythology and are commonly used in language (e.g. to be or not to be, Sisyphian task, Tantalus’ agony). There are also idioms that originated from various human professions (e.g. mathematics, music, theatre, nautics) and are used in everyday communication. Some of the examples include: bring sth under the common denominator, have an ear for sth, behind the scenes.

In addition to previously mentioned categories, there are also historical idioms. These idioms are related to historical people and events which left a strong impact in the society where they are still used in everyday communication. Some of the examples include: die is cast, like sheep without a shepherd, meet your Waterloo [5]. The last idiom has a national background and is unknown among other cultures. The translation cannot be found outside the culture in which this idiom has occurred. Even if one tried to translate it, the message would never be truly transmitted.
Consequently, idioms can also be differentiated according to the time and place of their emergence. Phraseology divides its constituents on national and international idioms. While former are almost impossible to translate, the latter can be transferred between languages [5].

Some idioms have both identical meaning and identical components (e.g. swallow the bait which is translated into Croatian as “progutati mamac”). In some instances, it is possible to come across idioms that have the same meaning but slightly different form (e.g. get out of the bed on the wrong side which is translated as “ustati na lijevu nogu”). Sometimes it is possible to find an idiom in the target language that has equivalent meaning in the source language, but a completely different form (e.g. piece of cake which is translated as “mačji kašalj”). Example translations listed are in Croatian.

One of the biggest mistakes when dealing with idioms is translating its components literally, word by word. Translator must be familiar with culture and language of both source and target language in order to recognize idioms and transfer them to the target language appropriately [5].

III. RELATED WORK

A study in [6] illustrates differing patterns between human and machine translations, but also between two different machine translation (MT) systems. Error analysis, as [7] puts it, gives a qualitative view on the MT system and should be an integral part of MT development. It can point to strengths and problem areas for a certain MT system, which is not possible using automatic evaluation metrics [8]. Automatic metrics, as well as some forms of human evaluation such as fluency and adequacy scoring or system ranking, provide quantitative system evaluation [7]. Research community would like to get answers to what kind of errors the system makes most often, whether a particular modification improves some aspect of translation, although the overall score is intact, whether one system is superior in all aspects of translation or just in some, etc. [9]. Idioms are usually explicitly tackled in MT error taxonomies, i.e. as expression in [10, 11], and as idioms in [12, 13, 14].

Standard statistical machine translation (SMT) systems do not model idioms explicitly [15]. The term phrase in PBSMT does not refer to a linguistic unit, but to a sequence of words. Although phrasal translations might indirectly capture multiword expressions (MWE), which is a more general concept, they are not distinguished from any other n-gram [15]. A rising interest has been detected in explicitly modelling MWEs within the SMT framework. The authors in [16] note that highly fixed expressions (e.g. by and large) can be represented as words-with-spaces in Natural Language Processing (NLP) applications. This, however, does not hold for semi-fixed and syntactically flexible expressions. Semi-fixed expressions adhere to strict constraints on word order and composition, but undergo some degree of lexical variation, e.g. inflection, variation in reflexive form, or determiner selection [16].

The authors in [17] show a simple approach to extract domain bilingual multiword expressions (MWE) and three methods to integrate them to Moses, the state-of-the-art PBSMT. The first method adds MWEs to the training corpus, the second adds one feature which has the value of 1 if the source language phrase contains a MWE and the target language phrase contains its translation, or the value of 0 otherwise, and the third method includes an additional phrase table containing automatically extracted MWEs.

A study in [15] proposes two different integration strategies for MWEs in SMT. In the first strategy they identify MWEs and turn them into a single unit. Therefore, from the perspective of SMT, all MWEs are considered frozen. In their second strategy they add a count feature which represents the number of MWEs in the input language phrase and integrate MWE knowledge as a feature in the translation lexicon. In that way the system is biased towards using phrases that do not break MWEs. This is sort of a generalization of a binary strategy from [17].

MWEs are integrated into the phrase table in two different ways in [18]. After identifying MWEs, the translation pairs are extracted from the corpus. In the first approach the aligner probability is kept, while in the second it is set to 1 in both translation directions. The factored model is used, which enables using lemmas beside surface forms of words.

A study in [19] shows that different MWE types require different integration methods in the SMT framework. Beside the two approaches proposed in [15], they also take zone integration approach in which they define reordering zones for all MWEs found in the test data. In that way the decoder is forced to respect the boundaries while constructing the hypothesis. However, although it is not allowed to translate out of zone phrases unless it fully finished translating the words in the zone, it is allowed to divide the zone into any combination of phrases and translate them individually and in any order. That is why the approach does not help to increase automatic scores.

The impact of idioms on SMT is evaluated in [20]. Unlike in this study, the authors focus on idiomatic expressions formed from the combination of a verb and a noun as its direct object (e.g. lose head). They show that even in that limited scenario idiomatic expressions pose a challenge to PBSMT systems, as witnessed by a drop in the BLEU score (BLEU score is proposed and presented in [21]).

A substitution based technique for improving SMT on idiomatic MWEs is evaluated in [22]. The method first performs substitution on the original idiom with its literal meaning before translation, and then replaces literal meanings with idioms following translation. Although a statistically significant improvement is reported, the authors conclude that there is still a lot of work to be done to solve the problems posed by idioms to SMT.
IV. EXPERIMENTAL SETUP

We compare the way that two online translation systems, i.e. GT\(^1\) and Asistent\(^2\), manipulate texts, more precisely idioms, in English-Croatian and Croatian-English language directions. The texts used in this analysis are constructed by hand and care was taken to ensure that the texts are abundant with idioms but make up coherent stories. More details on the test sets are given in Table I.

In the analysis of generated translations, we ignore punctuation errors, inappropriate upper case and lower case letters and discordance in gender, number and case. The focus is primarily on the translations of idioms\(^3\). For the purpose of analysis, we divide machine translations of idioms into the following categories: Equivalent, Appropriate meaning, Literal translation, Anomaly in form, Untranslated or partially translated, Wrong idiom, and Other.

V. RESULTS

The results are presented in Table II. For the purpose of comparison of translation directions, the distribution of error categories in percentages is given in Figure I.

The results consist mostly of literal translations of idioms, i.e. word by word translations ("stretch a dollar is translated into Croatian as "produžim dolar"). Some of the idioms are not translated into their equivalents in the target language, but their translation conveys appropriate meaning, which is in accordance with the context ("cold hearted is translated into Croatian as "hladnog srca", instead of "mrtav hladan"). For some of the idioms in this category it is not possible to use an equivalent idiom because it does not exist in the target language ("Jack of all trades is translated into Croatian as "dobar u svemu").

To a smaller extent, we identify idioms that are translated using an equivalent idiom in the target language ("armed to the teeth is translated into Croatian as “naoružan do zuba”), as well as idioms with some sort of anomaly in their form. The latter are idioms that are recognizable and translated appropriately, but have some sort of anomaly (an extra word, inappropriate preposition, wrong word). For example, idiom crying like the rain is translated as “plakala kao kiša s”, where the preposition s is unnecessary, but the rest of the idiom is translated appropriately.

The translations contain several idioms that are translated using an inappropriate idiom ("is down for the count is translated into Croatian as “manji od makova zrna”). Additionally, some idioms are not even translated, i.e. they remain in their original form ("pomrsio konce is translated into English as pomprio strings), because of training data sparsity, which is mainly caused by rich morphology. We group these idioms into separate categories.

VI. CONCLUSION

Although one would expect Croatian-English translations to be better, as English is well resourced, the opposite proves to be true. We attribute this to the fact that English texts are more prevalent on the Internet than texts in Croatian, and therefore, translations from English into Croatian are much more accessible. Moreover, in our analysis we ignore discordance in gender, number and

---

\(^1\) https://translate.google.hr/

\(^2\) http://server1.nlp.insight-centre.org/asistent/

\(^3\) Translations performed on January 19, 2017
case, which could also greatly affect results when translating from morphologically poor to morphologically rich language. Furthermore, due to a great number of existing idioms in both the source and target language, it was impossible to include them all in our corpus. This could have affected our final analysis.

The majority of generated translations are literal translations of idioms and other elements of the texts. This leads to the conclusion that the systems involved do not have special treatment for MWEs. The last three categories, namely Untranslated or partially translated, Wrong idiom and Other are so poorly represented that they can be dismissed from further investigations.

In some instances, GT achieves better results, while in others Asistent offers much better translations. Both give similar translations although we observe slight superiority of Asistent especially in English-Croatian translations.

The main conclusion that can be deducted from this study is that PBSMT systems do not capture MWEs to a sufficient degree and special attention needs to be paid first to their detection, and afterwards to finding appropriate integration method. Since literal translations predominate, back translations would yield similar if not even worse results.

ACKNOWLEDGMENT

This work has been fully supported by the University of Rijeka under the project number 16.13.2.2.01.

REFERENCES


