ANALYSIS OF MULTIWORD EXPRESSION TRANSLATION ERRORS IN STATISTICAL MACHINE TRANSLATION

Natalia Klyueva

Charles University in Prague

kljueva@ufal.mff.cuni.cz

Jeevanthi Liyanapathirana

Copenhagen Business School, Denmark.

jl.ibc@cbs.dk

In this paper, we are going to evaluate a statistical machine translation (SMT) system, Moses, trained for several language pairs to explore how it cope with multiword expression (MWE) translation. We will experiment with Czech-Russian, English-French and English-Sinhala language pairs to make sure that our conclusions are as language-independent as possible. Multiword Expressions present a sequence of words with non-compositional meaning, they differ from language to language and are highly idiosyncratic. Even for the related languages we can not be sure if the structure of MWE is similar or not to say nothing about typologically different languages.

We translated some frequent MWEs using Moses and checked if they were translated properly. We speculate under which conditions MWEs are translated properly and under which context they got mistranslated. We will distinguish several types of the multiword expressions based on their part of speech and function in a sentence: noun multiword expressions, auxiliary multiword expressions, light verbs, idioms.

Noun multiword expressions. Multiword expressions in our test set are mainly named entities(NE) or belong to domain specific terminology(e.g. english - french : military coup - coup d'etat). They generally contain a noun and some other part of speech. Those terms and NEs get translated properly if they were seen in the training data.

Auxiliary multiword expressions present mainly multiword prepositions (e.g. english - french with regard to / en ce qui concerne) and SMT also does not have a problem to handle them properly because their co-occurence in the data is quite frequent and parts of an expression are not separated by other words.

Light verb constructions (LVC) are generally formed by a verb and a noun where a verb does not bare its initial meaning, so that the whole construction takes the semantics of the noun. Some multiword verbs have identical component words in the languages(cz: hrát úlohu,ru: играть роль – to play role), and some not(cz: dát smysl – give sense vs. ru: иметь смысл – have sense). Generally, multiword expressions are translated properly within SMT when an LVC presents an n-gram, but when a verb is separated from a noun, this LVC is often mistranslated.

Idioms are MWEs that can include words of any part of speech and they generally bear a meaning that has very little to do with any component of MWE. Idiomatic constructions often present a challenge to MT systems because they might be equal in the languages (contain the same words), but that is not always the case. As our data belong to the domain of news, we have not found much idioms in the test set. An example : kick the bucket in English would mean □□in Sinhalese, which means to die. However, the machine translation system for English to SInhalese has very little resources, so it translates this expression into "□□□ , which just gives the literal meaning of the expression.

We have found out that SMT cope with MWE as soon as a multiword unit fits into a respective bigram or n-gram, which is present and is relatively frequent in the training data. In order to analyse the translation, we use rules to extract "potential" MWEs from the source text. We then investigate how the possible translation errors can be avoided , for example by training the MT system with the extracted MWEs. We also exploited the cases when the languages involved are under-resourced.

References

KOEH, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., AND HERBST, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions* (pp. 177-180). Association for Computational Linguistics.

LIYANAPATHIRANA, J.U. AND WEERASINGHE A.R., (2011). English to Sinhala Machine Translation: Towards Better information access for Sri Lankans . *Conference on Human Language Technology for Development,* Alexandria, Egypt.

GHONEIM, M. AND DIAB, M. (2013). Multiword expressions in the context of statistical machine translation. In *Proc. of IJCNLP* (pp. 1181-1187)

BOUAMOR, D., SEMMAR, N., AND ZWEIGENBAUM, P. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *LREC* (pp. 674-679).