Statistical Machine Translation with Factored Translation Model: MWEs, Separation of Affixes, and Others

Tsuyoshi Okita **Alexandru Ceausu** **Andy Way**

Dublin City University
Glasnevin, Dublin 9, Ireland

Abstract

This paper discusses Statistical Machine Translation when the target side is morphologically richer language. This paper intends to discuss the issues which are not covered by a factored translation model of Moses especially targetting EN–JP translation: the effect of Multi-Word Expressions, the separation of affixes, and other monolingual morphological issues. We intend to discuss these over a factored translation model.

Introduction

The factored translation model in Moses (Koehn and Hoang 2007; Koehn 2010) intends to handle morphologically rich languages in the target side by integrating additional linguistic markup at the word level, where each type of additional word-level information is called a factor. Typical factors include surface form, lemma, POS-tag and morphological features such as case, number, gender, person, tense, and aspect. This model allows users to decide how to handle factors jointly or not jointly in its translation processes and a generation process in details.

The first note is that in order to capture the noun cases agreement and the verb person conjugation, additional algorithms, such as the case identification algorithm for noun phrases and the person identification algorithm for verbs, may be required (Avramidis and Koehn 2008). The second note is that in order to use the factored translation model correctly, one key is to take care about whether the translation options do not explode, causing the problem in German for example (Graham and van Genabith 2010). The third note is that the annotation of morphological information may typically include ambiguity if we only base on the parsing results, which require some additional disambiguating process (Ceausu 2006).

Some Issues in EN–JP Translation

Although a factored translation model covers wide issues, there seems to be several issues missing considering translation between EN–JP.

The first issue is related to the correct word correspondences between the source and the target, especially related to Multi-Word Expressions (MWEs) which is problematic in word alignment. Since the precision of word alignment is at most around 90% for the easiest language pairs such as FR–EN, the training data for the factored translation model may be often contaminated by various kinds of noise (Okita, Graham, and Way 2010). The language pairs such as EN–JP which often consist of non-literal translation would be problematic. If we can correctly align bilingual Multi-Word Expressions (MWEs) (Okita et al. 2010), this may improve the overall translation.

The second issue is related to the decision whether we (horizontally) separate affixes and word stem or not [1] is already made. For example in EN–JP, the empirical evidences suggest that we separate affix(es) and word stem(s) since it obtains better BLEU score than the case when we do not separate them although the adequacy decreases. This is since when we separate them the meaning of case particle, such as nominative, genitive, dative, accusative, detaches from the word. The combination of word stem(s) with affix(es) in Japanese makes the resulted conjugation in verbs and nouns quite rich.

The third issue is related to (necessary and) sufficient morphological information for particular language pairs. Firstly, sufficient morphological information depends on (monolingual) language: most of the verbs in European language inflect based on person and number, while Japanese verbs inflect based on aspect. Secondly, some missing morphological information depends on (monolingual) language: there is no article and gender for noun phrases in Japanese.

Our Algorithm

Our algorithm is shown in Algorithm 1 which tries to improve BLEU score by examining these three issues. Step 1 relates to the second and the third issues and Step 3 relates to the first issue.

Preliminary Results

Baseline is a plain Moses with 5-gram LM (augmented by factors) by SRILM, and with the MWE-sensitive word align-

[1] The factored translation model vertically separate word / lemma / POS / morphology, but what we mean is to separate 'looks' into 'look' (word stem) and 's' (affix) in the case of JP, if we illustrate in EN.

Algorithm 1 Overall Algorithm

Step 1: Morphological pre-design: we use the knowledge that JP noun phrases are accompanied with case particles and that JP verbs / adjectives / adverbs have conjugation based on six stem forms (imperfective / continuative / terminal / attributive / hypothetical / imperative form) which shows aspect. This step is to decide we separate affixes from word step or not. Default setting for European language is 'no separation' and Japanese is 'separation'.

Step 2: Run a parser and / or morphological analyzer to obtain the necessary information for a given training corpus. Run a tiered tagger (Ceausu 2006) to disambiguate the annotation.

Step 3: Run a training procedure of factored translation model where a word aligner is replaced by a multi-word expression-sensitive word aligner (MWE-sensitive word aligner)(Okita et al. 2010) instead of GIZA++, with the bilingual terminology (verbal / nominal compounds) extracted from parallel corpus (Okita et al. 2010).

ment followed by phrase extraction. We used NTCIR-8 corpus (Fujii et al. 2010) for EN-JP (50k randomly extracted sentence pairs as training corpus). We proceeded the items mentioned in Section 3. We used Cabocha (Kudo and Matsumoto 2003) for morphological analysis for JP.

We use both sides with the factors of surface, lemma, POS-tag, and morphology. The baseline by the plain factored model was 21.67 BLEU point absolute. With affixes separation in step 1, our algorithm decreases the score 18.35 BLEU point absolute. Without affixes separation in step 1, our algorithm obtains 22.25 BLEU point absolute.

observed	#	%	type	#
1 form	911	40%	NP	1831012
2 forms	445	20%	VP	259432
3 forms	506	22%	ph (symbols)	68298
4 forms	270	12%	ph (prefix)	66729
5 forms	111	5%	ph (OOV)	66461
all forms	33	1%	ph (conjunction)	65159
			ph (attributives)	59633
			ph Adverbial phrases	33781

Table 1: Statistics of observed verb forms (left) and number of phrase types(right) in JP side.

Conclusion

The factored translation model intends to handle morphologically richer language in the target side. We extend the original target to handle Multi-Word Expressions, affixes separation, and other monolingual morphological information for EN–JP. Preliminary results for EN–JP show that the combination of MWEs and the separation of affixes improved the results, and the separation of MWEs and the combination of affixes did not improve the results.

There are various further studies. Firstly, although our preliminary results show the strategy to combine affixes with word stems negative in Japanese, our intuition is opposite. We would like to find a way how to obtain the improved results if we do not separate affixes and word stems in Japanese. This might be related to the free word order phenomenon in Japanese. Secondly, we would like to extend the scale of parallel corpus and language pairs.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`http://www.cngl.ie`) at Dublin City University. We would also like to thank the Irish Centre for High-End Computing and Machine Translation Marathons.

References

Avramidis, E., and Koehn, P. 2008. Enriching morphologically poor languages for statistical machine translation. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2008)* 763–770.

Ceausu, A. 2006. Maximum entropy tiered tagging. *In Proceedings of the 11th ESSLLI Student Session* 173–179.

Fujii, A.; Utiyama, M.; Yamamoto, M.; Utsuro, T.; Ehara, T.; Echizen-ya, H.; and Shimohata, S. 2010. Overview of the patent translation task at the NTCIR-8 workshop. *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access* 293–302.

Graham, Y., and van Genabith, J. 2010. Factored templates for factored machine translation models. *In Proceedings of the International Workshop on Spoken Language Translation 2010* 275–282.

Koehn, P., and Hoang, H. 2007. Factored translation models. *In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2007)* 868–876.

Koehn, P. 2010. Statistical machine translation. *Cambridge University Press. Cambridge. UK.*

Kudo, T., and Matsumoto, Y. 2003. Fast methods for kernel-based text analysis. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)* 24–31.

Okita, T.; Guerra, A. M.; Graham, Y.; and Way, A. 2010. Multi-word expression-sensitive word alignment. *In Proceedings of the Fourth International Workshop On Cross Ling ual Information Access (CLIA2010, collocated with COLING2010)* 26–34.

Okita, T.; Graham, Y.; and Way, A. 2010. Gap between theory and practice: Noise sensitive word alignment in machine translation. *In Journal of Machine Learning Research Workshop and Conference Proceedings Volume 11: Workshop on Applications of Pattern Analysis (WAPA2010)* 119–126.