MWE-sensitive Word Alignment in Factored Translation Model

Tsuyoshi Okita, Andy Way
CNGL / School of Computing, Dublin City University

Factored Translation Model

The factored translation model in Moses (Koehn et al. 2007; Avramidis and Koehn 2008; Koehn 2010), which consists of translation processes followed by a generation process, intends to handle morphologically rich languages by integrating additional linguistic markup at the word level, where each type of additional word-level information is called a factor with the independent assumptions shown in (1):

$$\begin{aligned} p(s_e, m_e | s_f, m_f) &= p(s_e | s_f, m_f) p(m_e | s_e, s_f, m_f) \\ &\simeq p(s_e | s_f) p(m_e | m_f)) \end{aligned} \quad (1)$$

When the target side is morphologically rich language, there are two ways to design this: one way is to have sufficiently rich morphological information in the target side to obtain the mappings in a generation process (Koehn et al. 2007), and the other is to have sufficient information in the source side to discriminate different inflected forms in the target side to obtain the mappings in the translation processes (Avramidis and Koehn 2008). These two are mutual exclusive in the sense that if the translation process supplies the information about the inflected forms in the target side, there is no need for the generation process to generate inflected forms. Similarly, if the generation process supplies them in the target side, there is no need for the translation process to consider the morphological transformation. Both methods uses morphological features including case, number, gender, person, tense, and aspect. The latter additionally uses the case identification algorithm for noun phrases and the person identification algorithm for verbs. Decoding steps do early pruning of expansions and has limitation of the number of translation options per input phrase to a maximum number.

Strong Assumptions

We intend to examine the following assumptions, which are often made without much examination before we apply to the factored translation model.

The first assumption is on the source and the target correct word correspondences. A direct effect of this is on the translation processes. Since the precision of word alignment is around 90% (Moore 2005) for the easiest language pairs such as FR–EN. Inevitablly, the training data for the factored translation model is often contaminated by various kinds of noise. The language pairs such as EN–JP which often consist of non-literal translation would be problematic.

The second assumption is that the decision is already made whether we (horizontally) separate a word and morpheme(s) or not. [1] For example in EN–JP, the empirical evidences suggest that we separate word(s) and morpheme(s) since it obtains better BLEU score than the case when we do not separate them although the adequacy decreases. The reason of decrese in adequacy may be due to the detachment of the case information, such as the nominative, genitive, dative, and accusative cases, from the word. The combination of word(s) with morpheme(s) in Japanese may make the resulted conjugation in verbs and nouns moderately rich.

The third assumption is that we know (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.

Similarly, there is another assumption on the generation process in the target side which generates surface forms given the lemma and linguistic factors. This process has an assumption that the target side is correctly parsed.[2] We do not discuss this item.

Our Algorithm

Our algorithm tries to improve BLEU score by examining these three assumptions. [3] Step 1 relates to the third assumption, Step 3 the second assumption, and Step 4 the first

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

[2] In practice, although English parser has accuracy around 93-4% with coverage around 90%, the target language with morphologically rich language may often decrease these figures.

[3] This is intended to be a general method, but we demonstrate this using the preliminary example between EN–JP here. Although the morphologically 'rich' target side is JP (JP is often not recognized as 'rich'.), we plan to extend the same strategy to Turkish and Arabic later.

assumption.

Algorithm 1 Overall Algorithm

Step 1: Morphological predesign: 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.
Step 2: Do segmentation of JP sentence into morphemes by a morphological analyzer.
(**Step 3**: Combine verb and morphemes with attaching case information. By this construction, we aim at not losing the information by morphemes).
Step 4: Do word alignment by a multi-word expression-sensitive word aligner (MWE-sensitive word aligner)(Okita et al. 2010) instead of GIZA++. First, we supply prior knowledge about bilingual terminology and nominal / (verbal) compounds. We use the same bilingual terminology extraction algorithm described in (Okita et al. 2010). Then, we run a MWE-sensitive word aligner with these prior knowledge.

Preliminary Results

Baseline is a plain Moses with 5-gram LM (augmented by factors) by SRILM, and with the MWE-sensitive word alignment followed by phrase extraction. We used NTCIR-8 corpus (Fujii et al. 2010) for EN-JP (200k randomly extracted sentence pairs as training corpus). We proceeded the items mentioned in Section 3.

The experiment was on the target side generating process. The baseline by the plain factored model was 21.67 BLEU point absolute. With step 3, our algorithm decreases the score 18.35. Without step 3, our algorithm obtains 22.23 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 (prefixes)	66729
5 forms	111	5%	ph (OOVs)	66461
6 forms	33	1%	ph (conjunctions)	65159
			ph (attributives)	59633
			Adverbial phrases	33781

Table 1: Statistics of observed verb forms (left) and number of phrase types(right) in JP side. In right figures, the inside of parenthesis means that the top of the phrase starts with symbols, and so forth.

Conclusion

The factored translation model is intended to handle morphologically rich language in the target side. Our motivation is to augment the word correspondences by the MWE-sensitive word aligner, examining several preconditions for the factored translation model. It is observed that the MWE-sensitive word aligner slightly increases the BLEU score 0.56 absolute and 2.5% relative. (Work in progress).

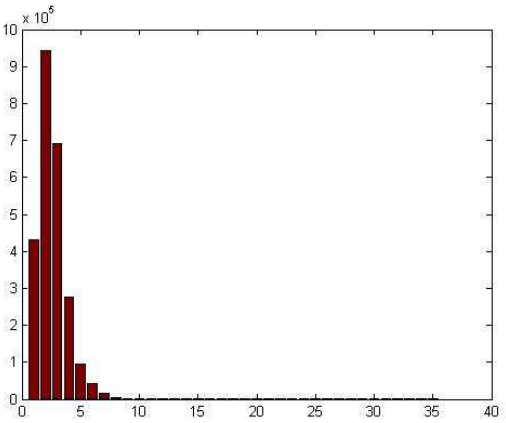

Figure 1: Statistics of number of nouns in NP.

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)*.

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*.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; and Herbst, E. 2007. Moses: Open source toolkit for Statistical Machine Translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* 177–180.

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

Moore, R. C. 2005. A discriminative framework for bilingual word alignment. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computationa Linguistics and the Empirical Methods in Natural Language Processing (HLT/EMNLP)*.

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.