Toward Using Morphology in French-English Phrase-based SMT

Marine Carpuat
Center for Computational Learning Systems
Columbia University
475 Riverside Drive, New York, NY 10115
marine@ccls.columbia.edu

Abstract

We describe the system used in our submission to the WMT-2009 French-English translation task. We use the Moses phrase-based Statistical Machine Translation system with two simple modications of the decoding input and word-alignment strategy based on morphology, and analyze their impact on translation quality.

1 Introduction

In this first participation to the French-English translation task at WMT, our goal was to build a standard phrase-based statistical machine translation system and study the impact of French morphological variations at different stages of training and decoding.

Many strategies have been proposed to integrate morphology information in SMT, including factored translation models (Koehn and Hoang, 2007), adding a translation dictionary containing inflected forms to the training data (Schwenk *et al.*, 2008), entirely replacing surface forms by representations built on lemmas and POS tags (Popović and Ney, 2004), morphemes learned in an unsupervised manner (Virpojia *et al.*, 2007), and using Porter stems and even 4-letter prefixes for word alignment (Watanabe *et al.*, 2006). In non-European languages, such as Arabic, heavy effort has been put in identifying appropriate input representations to improve SMT quality (e.g., Sadat and Habash (2006))

As a first step toward using morphology information in our French-English SMT system, this submission focused on studying the impact of different input representations for French based on the POS and lemmatization provided by the Treetagger tool (Schmid, 1994). In the WMT09 French-English data sets, we observe that more than half of the words that are unknown in the translation lexicon actually occur in the training data under different inflected forms. We show that combining a lemma backoff strategy at decoding time and improving alignments by generalizing across verb surface forms improves OOV rates and translation quality.

2 Translation system

2.1 Data sets

We use a subset of the data made available for the official French to English translation task. The evaluation test set consists of French news data from September to October 2008, however the bulk of the training data is not from the same domain. The translation model was trained on the Europarl corpus (europarl-v4) and the small news commentary corpus (news-commentary09). Following Déchelotte *et al.* (2008), we learn a single phrase table and reordering model rather than one for each domain, as it was found to yield better performance in a very similar setting. The language model was trained on the English side of these parallel corpora augmented with non-parallel English news data (news-train08.en). Parameter tuning was performed on the designated development data, which is also in the news domain: news-dev2009a was used as the development set and news-dev2009b as the test set.

Using those data sets, there is therefore a mismatch between the training and evaluation domains, as in the domain adaptation tasks of the previous WMT evaluations. A large automatically extracted parallel corpus was made available, but we were not able to use it due to time constraints. Additional use of this in-domain data would im-

*The author was partially funded by GALE DARPA Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 150–154,
Athens, Greece, 30 March – 31 March 2009.

prove coverage and translation quality.

2.2 Preprocessing

French and English corpora processing followed the same three steps:

First, long sentences are resegmented using simple punctuation-based heuristics.

Second, tokenization, POS tagging and lemmatization are performed with Treetagger (Schmid, 1994) using the standard French and English parameter files[1]. Treetagger is based on Hidden Markov Models where transition probabilities are estimated with decision trees. The POS tag set consists of 33 tags which capture tense information for verbs, but not gender and number.

Third, sentence-initial capitalized words are normalized to their most frequent form as reported by Zollmann *et al.* (2006).

2.3 Core system

We use the Moses phrase-based statistical machine translation system (Koehn *et al.*, 2007) and follow standard training, tuning and decoding strategies.

The translation model consists of a standard Moses phrase-table with lexicalized reordering. Bidirectional word alignments obtained with GIZA++ are intersected using the grow-diag-final heuristic. Translations of phrases of up to 7 words long are collected and scored with translation probabilities and lexical weighting.

The English language model is a 4-gram model with Kneser-Ney smoothing, built with the SRI language modeling toolkit (Stolcke, 2002).

The loglinear model feature weights were learned using minimum error rate training (MERT) (Och, 2003) with BLEU score (Papineni *et al.*, 2002) as the objective function.

Other decoding parameters were selected manually on an earlier version of the system trained and evaluated on the single-domain Europarl data. While the configuration achieved competitive results on the previous, it is not be optimal for this domain adaptation task.

We will first conduct an analysis of this core SMT system, and experiment with two modifications of input representation for decoding and alignment respectively.

[1] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

OOV verbs	**w/ surface form in training corpus**	**w/ lemma+ POS in training corpus**
dev2009a	21 (28%)	48 (63%)
dev2009b	16 (24%)	33 (49%)

Table 1: Unknown verbs statistics

3 Many unknown words are (almost) seen in training

Our baseline system is set up to copy unknown words to the output. This is a helpful strategy to translate unknown names and cognates, but is far from optimal. In this section, we take a closer look at those unknown words.

About 25% of the dev and test set sentences contain at least one unknown token. After eliminating number expressions, which can be handled with translation rules, the majority of unknown words are content words, nouns, verbs and adjectives. As reported in Table 1, we find that many of the verbs that are not in the phrase-table vocabulary were actually seen in the training data in the exact same form: they are therefore out of vocabulary due to alignment errors. In addition, for more than half of the unknown verb occurrences, another inflexion form for the same lemma and POS tag are observed in the training corpus.

Using only the surface form of words therefore leads us to ignore potentially useful information available in our training corpus. Additional training data would naturally improve coverage, but will not cover all possible morphological variations of all verbs, especially for tenses and persons that are not used frequently in news coverage. It is therefore necessary to generalize beyond word surface forms.

4 Using morphological information in decoding

A simple strategy for handling unknown words at decoding time consists in replacing their occurrences in the test set with their lemma, when it is part of the translation lexicon vocabulary. Unlike with factored models (Koehn and Hoang, 2007) or additional translation lexicons (Schwenk *et al.*, 2008), we do not generate the surface form back from the lemma translation, which means that tense, gender and number information are

news-dev2009a	**representation**	**OOV %**	**METEOR**	**BLEU**	**NIST**
baseline	surface form only	*2.24*	*49.05*	*20.45*	*6.135*
decoding	lemma backoff	2.13	49.12	20.44	6.143
word alignment	lemma+POS for all	2.24	48.87	20.36	6.145
	lemma+POS for adj	2.25	48.94	20.46	6.131
	lemma+POS for verbs	2.21	49.05	20.47	6.137
decoding + alignment	backoff + all	2.10	48.97	20.36	6.147
	backoff + adj	2.12	49.05	20.48	6.140
	backoff + verbs	**2.08**	**49.15**	**20.50**	**6.148**
news-dev2009b	**representation**	**OOV %**	**METEOR**	**BLEU**	**NIST**
baseline	surface form only	*2.52*	*49.60*	**21.10**	*6.211*
decoding	lemma backoff	2.43	49.66	21.02	6.210
word alignment	lemma+POS for all	2.53	49.56	21.03	6.199
	lemma+POS for adj	2.52	49.74	21.00	6.213
	lemma+POS for verbs	2.47	49.73	**21.10**	**6.217**
decoding+alignment	backoff + all	2.44	49.59	20.92	6.194
	backoff + adj	2.43	49.80	21.03	**6.217**
	backoff + verbs	**2.39**	**49.80**	21.03	**6.217**

Table 2: Evaluation of the decoding backoff strategy, the modified word alignment strategy and their combination

Input	Même s'il démissionnait, la situation ne changerait pas.
Baseline	even if it *démissionnait*, the situation will not change.
Lemma backoff	even if it **resign**, the situation will not change.
Reference	even if he resigned, the situation would remain the same.
Input	Tant que tu gagnes, on te laisse en paix
Baseline	As you *gagnes*, it leaves you in peace
Lemma backoff	As you **win**, it leaves you in peace
Reference	As Long as You Gain, We Let You
Input	Le groupe a réagi comme il faut, il a sorti un nouveau et meilleur disque.
Baseline	The group has reacted properly, it has *emerged* a new and better records.
Lemma+POS for verbs	The group has reacted properly, it has **produced** a new and better records.
Reference	The group responded with a new and even better CD.
Input	Un trader qui ne prend pas de vacances est un trader qui ne veut pas laisser son book à un autre", conclut Kerviel.
Baseline	A senior trader which does not take holiday is a senior trader which does not *allow* his book to another, " concludes Kerviel.
Lemma+POS for verbs	A senior trader which does not take holiday is a senior trader who do not wish to **leave** his book to another, " concludes Kerviel.
Reference	A broker who does not take vacations is a broker who does not want anybody to look into his records," Kerviel concluded.

Table 3: Examples of improved translations by morphological analysis

Input	54 pour cent ne font pas du tout confiance au premier ministre et 27 pour cent au président du Fidesz.
Baseline	54% are not all confidence to Prime Minister and the President of Fidesz 1.27%.
Backoff + verbs	54% do not all confidence to Prime Minister and 27% to the President of Fidesz.
Reference	Fifty-four percent said they did not trust the PM, while 27 percent said they mistrusted the Fidesz chairman.
Input	Le prsident Václav Klaus s'est nouveau prononc sur la problématique du rchauffement plantaire.
Baseline	President Václav Klaus has once again voted on the problem of global warming.
Backoff+verbs	President Václav Klaus has again pronounced on the problem of global warming.
Reference	President Václav Klaus has again commented on the problem of global warming.
Input	Mais les supérieurs étaient au courant de tout, ou plutôt, ils s'en doutaient.
Baseline	But superiors were aware of everything, or rather, they knew.
Backoff+verbs	But superiors were aware of everything, or rather, they doubted.
Reference	But his superiors are said to have known, or rather suspected the whole thing.

Table 4: Examples of translations that are not improved morphological analysis

lost. However, imperfect lemma translations can be more useful to understand the meaning of the input sentence than copying the unknown word to the output.

We report the impact of this strategy on automatic evaluation scores in the decoding section of Table 2. Since only a small subset of the test sentences are affected by the change, the score variation is small, but the OOV rate decreases and translation quality is not degraded. In addition to the BLEU and NIST n-gram precision metrics which only count exact matches between system output and reference, we report METEOR scores which take into account matches after lemmatization using both the Porter stemmer and the WordNet lemmas (Banerjee and Lavie, 2005). The improvement in METEOR scores results from more matches with the references, yielding both improved precision and recall.

Manual inspection of the output sentences shows that the translations are better to the human eye and potentially more useful to subsequent text understanding applications (Table 3).

5 Using morphological information in word-alignment

In this experiment, we would like to use morphological analysis to alleviate the alignment errors because of which some words from the parallel corpus are not in the phrase-table. We adopt a two-step approach: (1) before word alignment, replace surface forms by lemma and POS tags. In our experiments, this replacement is performed for 3 categories of words: verbs only, adjectives only and all words. (2) the phrase-table and reordering models are learned as usual using word surface forms, but with the alignment links from step 1.

In constrast with Watanabe *et al.* (2006), we attempt to generalize for specific word categories only, rather than use lemmas across all surface forms, as we found in earlier experiments that this approach did not help translation quality in our particular setting.

Unlike other approaches which use morphological analysis to change the representation of the input (e.g., Popović and Ney (2004), Sadat and Habash (2006), Virpojia *et al.* (2007)), our system still uses word surface forms as input during decoding. This is a constraint imposed by the relatively coarse analysis given by the default Treetagger lemmas and POS tags. Since they do not capture information that is crucial in translation such as number and gender, we need to keep surface forms as the input for translation.

The impact of this strategy on automatic evaluation metrics is reported in the word alignment section of Table 2. Note that all experiments were performed using the parameters learned by MERT on news-dev2009a using the baseline configuration. Again the impact in numbers is small, but does not degrade translation quality. The METEOR score is slightly improved on the real test set. As expected given our POS tag set, it seems better to restrict the modifications of the input for word alignment to verbs or adjectives.

This simple modification of the training procedure improves the coverage of the phrase-table, but the OOV rate remains higher than with the lemma backoff strategy. For the news-dev2009b test set, 1186 additional phrases are available in the phrase-table after replacing verb surface forms by their lemma and POS combination. About half of the test sentences are changed. As reflected by the scores, most of the changes are small and do not yield significantly different sentences. However, some translations are improved as can be seen in Table 3.

The impact of both strategies combined is reported in the decoding + alignment section of Table 2. Tables 3 and 4 show positive and negative examples of translations using the best combination.

6 Conclusion

We have described the system used for our submission, which is based on Moses with two simple modifications of the decoding input and word-alignment strategy in order to improve coverage without using additional training data. While the improvements on automatic metrics are small, manual inspection suggests that better morphological analysis for the French side has potential to improve translation quality. In future work, we plan to improve the core model by including the new large in-domain parallel corpus in training, and to further experiment with French input representations at different stages of training and decoding using more expressive POS tags such as the MULTITAG tag set (Allauzen and Bonneau-Maynard, 2008).

References

Alexandre Allauzen and Hélène Bonneau-Maynard. Training and evaluation of pos taggers on the french multitag corpus. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgement. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan, June 2005.

Daniel Déchelotte, Gilles Adda, Alexandre Allauzen, Hélène Bonneau-Maynard, Olivier Galibert, Jean-Luc Gauvain, Philippe Langlais, and François Yvon. LIMSIs statistical translation systems for WMT08. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 107–110, Columbus, Ohio, 2008.

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, 2007.

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. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June 2007.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 2003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

Maja Popović and Hermann Ney. Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, 2004.

Fatiha Sadat and Nizar Habash. Combination of arabic preprocessing schemes for statistical machine translation. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1–8, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

Helmut Schmid. Probabilistic part–of–speech tagging using decision trees. In *Proceedings of the Conference on New Methods in Language Processin*, pages 44–49, Manchester, UK, 1994.

Holger Schwenk, Jean-Baptiste Fouet, and Jean Senellart. First steps towards a general purpose French/English statistical machine translation system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 119–122, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Andreas Stolcke. SRILM—an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*, Denver, Colorado, September 2002.

Sami Virpojia, Jaako J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Machine Translation Summit XI*, pages 491–498, Copenhagen, Denmark, September 2007.

Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. Ntt system description for the wmt2006 shared task. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 122–125, New York City, June 2006. Association for Computational Linguistics.

Andreas Zollmann, Ashish Venugopal, Stephan Vogel, and Alex Waibel. The CMU-UKA Syntax Augmented Machine Translation System for IWSLT-06. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 138–144, Kyoto, Japan, 2006.