Bootstrapping Phrase-based Statistical Machine Translation via WSD Integration

Hien Vu Huy[†,♯] , **Phuong-Thai Nguyen**[†,♯] , **Tung-Lam Nguyen**[†,♯] and **M.L Nguyen**[†,‡]
[†] University of Engineering and Technology, VNU Hanoi
{hienvuhuy, thainp, lamnt_52}@vnu.edu.vn
[‡] Japan Advanced Institute of Science and Technology (JAIST)
nguyenml@jaist.ac.jp

Abstract

Beside the word order problem, word choice is another major obstacle for machine translation. Though phrase-based statistical machine translation (SMT) has an advantage of word choice based on local context, exploiting larger context is an interesting research topic. Recently, there have been a number of studies on integrating word sense disambiguation (WSD) into phrase-based SMT. The WSD score has been used as a feature of translation. In this paper, we will show that by bootstrapping WSD models using unlabeled data, we can bootstrap an SMT system. Our experiments on English-Vietnamese translation showed that BLEU scores have been improved significantly.

1 Introduction

Conventional phrase-based systems use local context information from phrase table and language model. Though phrase based SMT achieves a jump in translation quality in comparison with word based SMT, there are still cases in which local context cannot capture correctly the meanings of source words. WSD can use features from much larger contexts and those features can overlap each other. The idea of integrating WSD into SMT rises naturally from this perspective. Previously, Varea et al. (2001) directly used context sensitive lexical models, applying these models for re-ranking n-best for their word-based maximum entropy model (MEM) SMT and achieving slight improvements in translation quality.

Chan et al. (2007) made use of WSD for hierarchical phrase-based translation for Chinese-English by utilizing two new WSD features for SMT and proposing an algorithm for scoring synchronous rules. Phrases which do not exceed a length of two were computed WSD models. Their experiments showed that WSD can improve SMT significantly.

Simultaneously with Chan et al. (2007), Carpuat and Wu (2007) used a similar approach to the problem. The main difference was that they focused on conventional phrase-based SMT in Koehn et al. (2003) and used only one WSD feature for SMT. The limit of phrase length was the same as the value used by their SMT system. Their experiments led to the same conclusion: WSD can improve SMT.

However, approaches based on statistic frequently against deficiencies of parallel and specific domain corpora. Only a few popular languages are derived continuous financial support and interest of researchers. Therefore, it becomes an immense obstacle to apply these approaches for the remaining languages.

Recently, there are several approaches to address this impediment. Ambati et al. (2011) applied multi-strategy methods in active learning for machine translation by combining several techniques in sentence selection process. They attained significantly results while parallel training data was scarce.

In this paper, we present our study on this topic. First, by integrating WSD as a model of SMT system as shown in the Figure 1, we present how we use WSD for SMT. Then we demonstrate a method to bootstrap WSD models by using unlabelled data. Finally, we show our experimental results. We analyse various settings of WSD-SMT integration. Our results give a thorough view into the problem.

2 WSD for SMT

2.1 WSD Task

In order to use WSD for SMT, the precondition is that training data must be large enough.

International Joint Conference on Natural Language Processing, pages 1042–1046,
Nagoya, Japan, 14-18 October 2013.

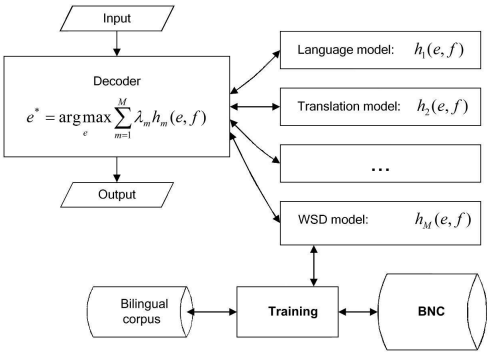

Figure 1: Integrating WSD into phrase-based SMT system

Manually-created data sets such as SENSEVAL and SemCor, which are often used in WSD studies, are too small for applications like machine translation. We overcome this difficulty by using an approach based on Carpuat and Wu (2007) and Chan et al. (2007) to extract training data from bilingual data. Word alignment information serves as a map between source words and target words. Target words are seen as senses. Since word alignment usually performs incorrectly, the resulting WSD training data is noisy. When carrying out this research, we consider WSD for word and phrase levels.

2.2 WSD Training Data Generation

A procedure for WSD-training-data extraction:

Input: a bilingual corpus, a POS-tagged version of the source text and word alignment information.

Output: WSD training sets for source phrases.

- Step 1: Collect phrase pair instances associated with position in the bilingual corpus. Group phrase pairs according to source phrase.
- Step 2: For each group, generate a training set for its corresponding source phrase.

Phrase pairs (s,t) which are consistent with the word alignment will be generated. The criteria of consistence with word alignment in Koehn et al. (2003) are as follows: First, there exists links from words of s to words of t. Second, for every word outside s, there is no link to any word of t. Third, for every word outside t, there is no link to any word of s.

When extracting WSD training data from a bilingual corpus, the number of training sets resulting from the extractive procedure is often much larger than vocabulary size of the source text. Additionally, raw data extracted from a bilingual corpus is a miscellany of semantic, lexical, morphological, an syntactic ingredients. It is very different from conventional WSD data style. This data can be refined in several ways such as lemmatization.

2.3 WSD Features

In our work, we use six kinds of knowledge and represent them as subsets of features, as follows:

- *bag-of-words*, $F_1(l, r) = \{w_{-l}, \ldots, w_{+r}\}$: We investigate three sets of this knowledge including $F_1^a = F_1(-5, +5)$, $F_1^b = F_1(-10, +10)$, $F_1^c = F_1(-100, +100)$, corresponding to small, medium and large size respectively.
- *collocation of words*, $F_2 = \{w_{-l} \ldots w_{+r}\}$: As a result of the work in Le and Shimazu (2004) we choose such collocations that their lengths (including the target words) are less than or equal to 4, it means $(l + r + 1) \leq 4$.
- *ordered words*, $F_3 = \{w_i | i = -l, \ldots, +r\}$: We choose $l = r = 3$
- *collocation of POSs*, $F_4 = \{p_{-l} \ldots p_{+r}\}$: Like collocation of words, we choose their lengths including the target words are less than or equal to 4.
- *ordered POSs*: $F_5 = \{p_i | i = -l, \ldots, +r\}$: We choose $l = r = 3$

In cases that we are working with a training set of a source phrase, features will be extracted from surrounding context of that phrase.

2.4 Integration

After having been trained, WSD models can be used as a feature for SMT as shown in the Figure 1. Since we use a log linear translation model, the use of a new feature is easy. Feature's weight is tuned using minimum error rate training (MERT) in Och (2003). In decoding phase, when translation options are generated, their WSD score is computed and then can be used in searching process. Among other features, this new feature is sensitive to large contexts.

Given a source phrase, the simplest way is to train its own WSD model and then apply that model in new contexts. The number of WSD models is equal to the number of source phrases in the SMT phrase table. An alternative is to score a phrase using shorter phrases. That means only WSD models for phrases whose length is smaller than a threshold to be trained. This setting could

reduce computational time. Suppose that we are considering a phrase pair (s, t) in which s is a source phrase, t is a target phrase. If this phrase pair can be split into a sequence (s_i, t_i) of n sub phrase pairs which are consistent with the word alignment of (s, t), then the probability of t given s and its context can be computed using (1) here

$$P_{wsd}(t|s) \approx \prod_{i=1}^{n} P_{wsd}(t_i|s_i) \quad (1)$$

$P_{wsd}(t_i|s_i)$ calculates the probability of t_i conditioning on s_i and its surrounding context. If there are more than one possible split, we use a greedy method. This method gives preferences to sub phrases according to their length and score.

3 Using Unlabelled Data

3.1 Basic Algorithm

Suppose that we have two data sets, one labelled (eg., the data extracted from a bilingual corpus) and the other unlabelled. First, a classifier is trained using the labelled data set, then it can be used to classify the unlabelled data set. Among newly labelled examples, the ones with high score will be chosen to enlarge the training data. These steps are repeated until a stopping condition is matched. Stopping condition can be a maximum number of iterations, or a minimum increase in classification accuracy, etc.

Input: L = a labelled data set.
U = an unlabelled data set.

Output: L_{new}, a new labelled data set.

1. Train a classifier C using L.
2. For each u ∈ U:
   a. use C to classify u.
   b. find the label assigned with highest score.
   c. if the score is above a threshold, choose u.
3. L_{new} = L $\bigcup$ { u ∈ U : u has been labelled}. and U_{new} = { u ∈ U : u unlabelled}.
4. If the stopping condition is not matched, repeat from step 1, else stop.

3.2 A New Algorithm with Sense Distribution Control

A problem with the basic algorithm is that after extension, the resulting labelled data set can be highly imbalanced in sense distribution with dominating senses, due to which the classification accuracy decreases. To handle the problem, the change of sense distribution during extending process should be controlled. We propose to use the relative entropy or Kullback-Leibler distance in Cover and Thomas (2006) to measure the change in sense distribution and control the amount of new examples. After extending using the previous algorithm, we will remove examples one by one until the KL distance is smaller than a threshold. The threshold need not to be a fixed number.

Algorithm: Input: a labelled data set $L_{initial}$ and its expanded set L_{new}.

Output: a labelled data set $L_{extending}$ whose sense distribution is controlled

1. p = $(p_1, p_2, \ldots, p_n)$ and q = $(q_1, q_2, \ldots, q_n)$ are sense distributions over $L_{initial}$ and L_{new}.
2. Compute the Kullback Leibler distance between p and q: Δ = KL(p,q) = $\sum_{i}^{n} p_i log(\frac{p_i}{q_i})$
3. Repeat
   a. for each u ∈ L_{new}:
      - compute t = $(t_1, t_2, \ldots\ t_n)$, the sense distribution over T = $L_{new} \setminus \{u\}$, then compute KL(t,p)
      - find u_m minimizing KL(t, p), then u_m is the element that when removing it, the KL distance decreases a maximum amount.
   b. Remove u_m from L_{new}
   c. Compute KL(p,q)
4. The iteration stop when KL(p,q) $< \frac{\Delta}{2}$
5. $L_{extending}$ = L_{new}.
6. Return $L_{extending}$.

4 Evaluation

4.1 Corpora and Tools

The corpus in our experiments is English-Vietnamese bilingual corpus from several different fields which includes approximately 135,000 sentence pairs. It is divided into three parts: training, developing and testing in Table 1. We used the developing set in the evaluation of MERT of SMT system in all experiments. In addition to the testing set extracted from the bilingual corpus, we used an additional corpus consisting of ambiguous words that are labelled by evaluators to test the external domain. The rate of Out-of-Vocabulary in testing sets is roughly 2%.

In our experiments, the British National Corpus (BNC) in Clear (1993) has been used for our expansion. We used a word-segmentation program in Nguyen et al. (2003), Moses in Koehn et al. (2007), GIZA++ in Och and Ney (2000), SRILM in Stolcke (2002), a rule-based morpho-

logical analyser in Pham et al. (2003) and Natural Language Toolkit in (Bird et al., 2009) for segmenting Vietnamese sentences, learning phrase translations, creating word alignment, learning language models, analysing morphology and exploiting BNC respectively.

	Number of sentences	**Average length of sentences**	**Number of words**
Training corpus			
English	131,118	15.9	2,096,073
Vietnamese	131,118	17.0	2,236,847
Developing corpus			
English	218	15.4	3,367
Vietnamese	218	16.5	3,609
Testing corpus			
English	2,000	17.8	35,797
Vietnamese	2,000	19.4	38,814
External-domain testing corpus			
English	123	18.7	2,308

Table 1: Statistics for training, testing and developing corpora

4.2 Experiments and results

	Without WSD	WSD integration	WSD integration with BNC
BLEU	34.93	35.43	36.47
NIST	7.4491	7.4937	7.7971

Table 2: BLEU scores of SMT based on phrase-based with WSD and BNC-extended WSD

As indicated from the Table 2, that SMT system utilizes WSD with expanded information of BNC corpus leads to the high translation quality with growths by 1.04 and 1.54 in BLUE score and 0.3034 and 0.3488 in NIST score in comparison with non-extended WSD integrated SMT system and baseline SMT system. Let consider the example:

Input: *hard water is water that has high mineral content (in contrast with soft water).*

SMT: **chăm_chỉ**/(hard) nước/(water) là/(is) nước/(water) cao/(hight) nội_dung/(content) khoáng_sản/(mineral) trái/(in contrast) với/(with) nước/(water) mềm/(soft) .

SMT + WSD: **khó**/(hard) nước/(water) là/(is) nước/(water) có/(has) hàm_lượng/(content) khoáng_sản/(mineral) cao/(hight) mềm/(soft) (ngược_lại)/(in contrast) với/(with) nước/(water).

SMT + WSD + BNC: nước/(water) rất **cứng**/(hard) là/(is) nước/(water) cao/(hight) hàm_lượng/(content) khoáng_sản/(mineral) trái/(in contrast) với/(with) mềm/(soft) ra nước/(water).

Clearly, ambiguous words in above example were translated precisely in the target language when utilizing WSD and BNC. In the first example, the word *hard* in *hard water* is translated to *cứng* (a type of water) which is more accurate than *chăm chỉ* (a personality) and *khó* (a difficulty).

4.3 The impact of context on WSD and WSD on SMT system

In many cases, the evaluation result of WSD is incorrect, resulting in the effect on the translation outcome of SMT. Below are two main reasons for this phenomenon: First, after the BNC expansion, the context could not embrace all possible cases due to limitation of contexts of BNC. Second, in several situations, information contexts of surrounding sentences should be used to determine labels of ambiguous words, whereas the system only uses the information in one sentence.

Besides, in the integration of WSD system into SMT system, WSD system occupies only a certain weight thus translation results are depend majorly on other models such as language model, translation model even though WSD gave precise results.

5 Conclusions

In this paper, we indicated a considerable effect of WSD which is bootstrapped on SMT system. The analyses and results on experiments point out that the approach of enhancing quality of WSD model contributes to the improvement of translation quality. The explanation for the increase of BLEU point is the impact of sparse data on the training set in WSD model. The expansion of training data from BNC whereby not only increases the degree of accurateness of WSD system but also improves the quality of translation. In the future, we would like to continue to experiment with the expansion of the training set on other sources to enhance the quality of translation.

Acknowledgments

This paper has been supported by VNU project "Exploiting Very Large Monolingual Corpora for Statistical Machine Translation" (code QG.12.49).

References

Vamshi Ambati, Stephan Vogen and Jaime Carbonell. 2011. *Multi-Strategy Approaches to Active Learning for Statistical Machine Translation*. Proc of the 13th Machine Translation Summit.

Marine Carpuat and Dekai Wu. 2007. *Improving Statistical Machine Translation Using Word Sense Disambiguation*. Proceedings of EMNLP-CoNLL.

Y. S. Chan, H. T. Ng, and D. Chiang. 2007. *Word Sense Disambiguation Improves Statistical Machine Translation*. Proceedings of ACL.

Jeremy H. Clear 1993. *The British National Corpus* MIT Press, Cambridge, MA, USA, pages 163–187.

Philipp Koehn, Franz Josef Och, Daniel Marcu. 2003. *Statistical Phrase-based Translation*. In Proceedings of HLT-NAACL.

Philipp Koehn et al. June, 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* In ACL, demonstration session, Prague, Czech Republic.

C.A. Le and A. Shimazu. 2004. *High Word Sense Disambiguation Using Naive Bayesian Classifier with Rich Features*. The 18th Pacific Asian Conference on Linguistic Information and Computation (PACLIC18), pages 105–113.

Nguyen, T. P., Nguyen V. V. and Le A. C. 2003. *Vietnamese Word Segmentation Using Hidden Markov Model*. In Proceedings of International Workshop for Computer, Information, and Communication Technologies in Korea and Vietnam.

Franz Josef Och and Hermann Ney. 2000. *Improved Statistical Alignment Models*. In Proceedings of ACL.

Pham, N. H., Nguyen L. M., Le A. C., Nguyen P. T., and Nguyen V. V. 2003. *LVT: An English-Vietnamese Machine Translation System*. In Proceedings of FAIR.

Stolcke, A. September, 2002. *SRILM - An Extensible Language Modeling Toolkit*. In Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado.

Varea, I. G., F. J. Och, H. Ney, and F. Casacuberta. 2001. *Refined Lexicon Models for Statistical Machine Translation using a Maximum Entropy Approach*. Proceedings of ACL, pages 204–211.

Bird, Steven, Ewan Klein and Edward Loper. 2006. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media, 2009

Thomas M. Cover and Joy A. Thomas: *Elements of Information Theory*. New Jersey, John Wiley & Son.

Och F.J. 2003 *Minimum Error Rate Training in Statistical Machine Translation*. Proceedings of the 41st International Conference on Computational Linguistics, pages 160–167.