Evaluating the Word Sense Disambiguation Performance of Statistical Machine Translation

Marine CARPUAT **Dekai WU**[1]
marine@cs.ust.hk dekai@cs.ust.hk

Human Language Technology Center
HKUST
Department of Computer Science
University of Science and Technology
Clear Water Bay, Hong Kong

Abstract

We present the first known empirical test of an increasingly common speculative claim, by evaluating a representative Chinese-to-English SMT model directly on word sense disambiguation performance, using standard WSD evaluation methodology and datasets from the Senseval-3 Chinese lexical sample task. Much effort has been put in designing and evaluating dedicated word sense disambiguation (WSD) models, in particular with the Senseval series of workshops. At the same time, the recent improvements in the BLEU scores of statistical machine translation (SMT) suggests that SMT models are good at predicting the right translation of the words in source language sentences. Surprisingly however, the WSD accuracy of SMT models has never been evaluated and compared with that of the dedicated WSD models. We present controlled experiments showing the WSD accuracy of current typical SMT models to be significantly lower than that of all the dedicated WSD models considered. This tends to support the view that despite recent speculative claims to the contrary, current SMT models do have limitations in comparison with dedicated WSD models, and that SMT should benefit from the better predictions made by the WSD models.

[1]The authors would like to thank the Hong Kong Research Grants Council (RGC) for supporting this research in part through grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09.

1 Introduction

Word sense disambiguation or WSD, the task of identifying the correct sense of a word in context, is a central problem for all natural language processing applications, and in particular machine translation: different senses of a word translate differently in other languages, and resolving sense ambiguity is needed to identify the right translation of a word.

Much work has been done in building dedicated WSD models. The recent Senseval series of workshop promoted controlled comparison of very different WSD models with common accuracy metrics and common data sets. These efforts yielded steady improvements in WSD accuracy, but for WSD evaluated as a standalone task. Senseval focuses on the evaluation of standalone, generic WSD models, even though many application-specific systems—machine translation, information retrieval, and so on—all perform WSD either explicitly or implicitly.

Since the Senseval models have been built and optimized specifically to address the WSD problems, they typically use richer disambiguating information than SMT systems. This, however, raises the question of whether the sophisticated WSD models are in fact needed in practice.

In many machine translation architectures, in particular most current statistical machine translation (SMT) models, the WSD problem is typically not explicitly addressed. However, recent progress in machine translation and the continuous improvement on evaluation metrics such as BLEU (Papineni *et al.*, 2002) suggest that SMT systems are already very good at choosing correct word translations. BLEU score with low order n-grams can be seen as an evaluation of the translation adequacy, which suggests that as SMT systems achieve higher BLEU score, their ability to disambiguate word translations improves.

In other work, we have been conducting comparative studies testing whether state-of-the-art WSD mod-

els can improve SMT translation quality (Carpuat and Wu, 2005). Using a state-of-the-art Chinese word sense disambiguation model to choose translation candidates for a typical IBM statistical MT system, we found that word sense disambiguation does *not* yield significantly better translation quality than the statistical machine translation system alone. The surprising difficulty of this challenge might suggest that SMT models are sufficiently strong at word level disambiguation on their own, and has recently encouraged speculation that SMT performs WSD as well as the dedicated WSD models.

The studies described in this paper are aimed at directly testing this increasingly common speculation. The comparison of SMT and WSD strengths is not obvious; there are strong arguments in support of both the WSD and the SMT models. A controlled empirical comparison is therefore needed to better assess the strengths and weaknesses of each type of model on the WSD task.

We therefore propose to evaluate statistical machine translation models on a WSD task, in terms of standard WSD accuracy metrics. This addresses the inverse, complementary question to the other study mentioned above (of whether WSD models can help SMT systems in terms of machine translation quality metrics). Senseval provides a good framework for this evaluation, and allows a direct comparison of the performance of the SMT model with state-of-the-art WSD models on a common dataset. We built a Chinese-to-English SMT system using freely available toolkits, and show that it does not perform as well as the WSD models specifically built for this task.

2 Statistical machine translation vs. word sense disambiguation

We begin by examining the respective strengths and weaknesses of full SMT models versus dedicated WSD models, which might be expected to be relevant to the WSD task.

2.1 Features unique to SMT

Unlike lexical sample WSD models, SMT models simultaneously translate complete sentences rather than isolated target words. The lexical choices are made in a way that heavily prefers *phrasal cohesion* in the output target sentence, as scored by the language model. That is, the predictions benefit from the sentential context of the *target* language. This has the general effect of improving translation fluency.

Another major difference with most lexical sample WSD models is that SMT models are always unsupervised. SMT models learn from large sets of bisentences but the correct word alignment between the two sentences is unknown. SMT models cannot therefore directly exploit sense-annotated data, or at least not as easily as classification-based WSD models do.

2.2 Features unique to WSD

Dedicated WSD is typically cast as a classification task with a predefined sense inventory. Sense distinctions and granularity are often manually predefined, which means that they can be adapted to the task at hand, but also that the translation candidates are limited to an existing set.

To improve accuracy, dedicated WSD models typically employ features that are not limited to the local context, and that include more linguistic information than the surface form of words. For example, a typical dedicated WSD model might employ features as described by Yarowsky and Florian (2002) in their "feature-enhanced naive Bayes model", with position-sensitive, syntactic, and local collocational features. The feature set made available to the WSD model to predict lexical choices is therefore much richer than that used by a statistical MT model.

Also, dedicated WSD models can be supervised, which yields significantly higher accuracies than unsupervised. For the experiments described in this study we employed supervised training, exploiting the annotated corpus that was produced for the Senseval-3 evaluation.

Again, this brief comparison shows that both models have important and very different strengths, which motivates our controlled empirical comparison of their WSD performance.

3 The SMT system

To build a representative baseline SMT system, we restricted ourselves to making use of freely available tools.

3.1 Alignment model

The alignment model was trained with GIZA++ (Och and Ney, 2003), which implements the most typical IBM and HMM alignment models. Translation quality could be improved using more advanced hybrid phrasal or tree models, but this would interfere with the questions being investigated here. The alignment model used is IBM-4, as required by our decoder. The training scheme is IBM-1, HMM, IBM-3 and IBM-4, as specified in (Och and Ney, 2003).

The training corpus consists of about 1 million sentences from the United Nations Chinese-English parallel corpus from LDC. This corpus was automatically sentence-aligned, so the training data does not require as much manual annotation as for the WSD model.

3.2 Language model

The English language model is a trigram model trained on the Gigaword newswire data and on the English side of the UN and Xinhua parallel corpora. The language model is also trained using a publicly available software, the CMU-Cambridge Statistical Language Modeling Toolkit (Clarkson and Rosenfeld, 1997).

3.3 Decoding

The ISI ReWrite decoder (Germann, 2003), which implements an efficient greedy decoding algorithm, is used to translate the Chinese sentences, using the alignment model and language model previously described.

Notice that very little contextual information is available to the IBM SMT models. Lexical choice during decoding essentially depends on the translation probabilities learned for the target word, and on the English language model scores.

4 The WSD system

The WSD system used here is based on the model that achieved the best performance on the Senseval-3 Chinese lexical sample task, outperforming other systems by a large margin (Carpuat *et al.*, 2004).

The model consists of an ensemble of four highly accurate classifiers combined by majority vote: a naive Bayes classifier, a maximum entropy model (Jaynes, 1978), a boosting model (Freund and Schapire, 1997), and a Kernel PCA-based model (Wu *et al.*, 2004), which has the advantage of having a signficantly different bias. All these classifiers have the ability to handle large numbers of sparse features, many of which may be irrelevant. Moreover, the maximum entropy and boosting models are known to be well suited to handling features that are highly interdependent.

The feature set used consists of position-sensitive, syntactic, and local collocational features, as described by Yarowsky and Florian (2002).

5 Experimental method

5.1 Senseval-3 Chinese lexical sample task

The Senseval-3 Chinese lexical sample task includes 20 target word types. For each word type, several senses are defined using the HowNet knowledge base. There are an average of 3.95 senses per target word type, ranging from 2 to 8. Only about 37 training instances per target word are available.

The dedicated WSD models described in Section 4 are trained to predict HowNet senses for a set of new occurrences of the target word in context.

We use the SMT sytem described in Section 3 to translate the Chinese sentences of the Senseval evaluation test set, and extract the translation chosen for each of the target word occurrences. In order to evaluate the predictions of the SMT model just like any WSD model, we need to map the English translations to HowNet senses. This mapping is done using HowNet, which provides English glosses for each of the senses of every Chinese word.

Note that Senseval-3 also defined a translation or multilingual lexical sample task (Chklovski *et al.*, 2004), which is just like the English lexical sample task, except that the WSD systems are expected to predict Hindi translations instead of WordNet senses. This translation task might seem to be a more natural evaluation framework for SMT than the monolingual Chinese lexical sample task. However, in practice, there is very little data available to train an English-to-Hindi SMT model, which would significantly hinder its performance and bias the study in favor of the dedicated WSD models.

5.2 Allowing the SMT model to exploit the Senseval data

Comparing the Senseval WSD models with a regular SMT model is not entirely fair, since, unlike the SMT model, the dedicated WSD models are trained and evaluated on similar data. We address this problem by adapting our SMT model to the lexical sample task domain in two ways.

First, we augment the training set of the SMT model with the Senseval training data. Since the training set consists of sense-annotated Chinese sentences, and not of Chinese-English bisentences, we artificially create sentence pairs for each training instance, where the Chinese sentence consists of the Chinese target word, and the English sense is the English gloss given by HowNet for that particular target word and HowNet sense.

Second, we restrict the translation candidates considered by the decoder for the target words to the set of all the English glosses given by HowNet for all the senses of the target word considered. With this modification, the degree of ambiguity faced by the SMT model is closer to that of the WSD model.

Table 1 shows that each of these modifications help the accuracy, overall yielding a 28.9% relative improvement over the regular SMT system.

6 Results

Table 2 summarizes the results of the SMT and WSD models on the Senseval-3 Chinese lexical sample task.

Note that the accuracy of the most frequent sense baseline is extremely low, which shows that the evaluation set contains instances that are particularly difficult to disambiguate. All our SMT and WSD models significantly outperform this baseline.

Table 1: Evaluation of the different variations of the SMT model on the Senseval-3 Chinese Lexical Sample task.

System	**Accuracy**
SMT	25.6
SMT with augmented training set	26.9
SMT with augmented training set and restricted translation candidates	*33.0*

Table 2: Accuracy of all our SMT and WSD models on the Senseval-3 Chinese Lexical Sample task.

System	**Accuracy**
Most Frequent Sense Baseline	7.7
Best SMT system	*33.0*
UMD unsupervised system	44.5
WSD naive Bayes	60.4
WSD KPCA	63.6
WSD Boosting	64.1
WSD Maximum Entropy	64.4
WSD Ensemble (current best Senseval-3 model)	*66.2*

6.1 The dedicated supervised WSD models all significantly outperform SMT

Table 2 clearly shows that even the best of the SMT model considered performs significantly worse than any of the dedicated WSD models considered. The accuracy of the best Senseval-3 system is double the accuracy of the best SMT model.

Since the best Senseval-3 system is a classifier ensemble that benefits from the predictions of four individual WSD models which have very different biases, we also compare the performance of the SMT model with that of the individual WSD models. All the individual component WSD models, including the simplest naive Bayes model, also significantly outperform the SMT model.

6.2 The SMT model prefers phrasal cohesion in the output sentences to WSD accuracy

Inspection of the output reveals that the main cause of errors is that the SMT model tends to prefer phrasal cohesion to word translation adequacy: lexical choice is essentially performed by the English language model, therefore translations are primarily chosen to preserve phrasal cohesion in the output sentence, and only local context is used. We will give three different examples to illustrate this effect.

The Chinese word “活动”(huodong) has the senses “move/exercise” vs. “act”. A Chinese sentence is incorrectly translated as “the party leadership which develop the constitution and laws and in constitutional and legal framework exercise”. Here, “exercise” is not the right translation, “act” should be used instead. However, the language model prefers the use of the phrase “legal framework exercise”, where the word “exercise” is used in a different sense than the one meant in the “move/exercise” category. Note that choosing the wrong translation for this word not only affects the adequacy, but also the grammaticality and fluency of the translated sentence.

In one of the target sentences, the SMT model has to choose between two translations for the Chinese word “材料” (cailiao): “data” or “material”. The two closest left neighbors can be translated as “provide proof”, and the SMT incorrectly picks the “material” sense, because the phrase “provide proof of material...” is more frequent than “provide proof of data”. In contrast, the WSD model has access to a wider context to correctly pick the “data” translation.

Similarly, in a test sentence where the Chinese word “分子” (fengzi) is used in the sense “element/component” vs. “member”, the SMT system incorrectly chooses the “member” translation because the neighboring word translates to “active”, and the language model prefers the phrase “active member” to “active element” or “active component”.

6.3 WSD models are consistently better than SMT models for all target word types

Computing accuracies per target word type shows that the previous observations hold for each target word. Table 3 compares the accuracies of the best SMT vs. the best WSD system per target word type and shows that the WSD system always yields significantly higher scores for the set of target words considered.

Also this breakdown reveals, interestingly, that the

Table 3: Accuracy of the best SMT and best WSD models for each target word type in the Senseval-3 Chinese Lexical Sample task.

Target word	SMT accuracy	WSD accuracy
冲击	38.5	53.8
分子	62.5	81.2
包	13.8	66.6
地方	29.4	64.7
坐	16.7	58.3
少	15.8	89.5
把握	40.0	66.7
日子	28.6	61.9
材料	40.0	60.0
没有	66.7	60.0
活动	43.7	62.5
研究	53.3	80.0
穿	28.6	57.1
突出	40.0	73.3
老	26.9	57.7
走	20.8	62.5
起来	25.0	55.0
路	17.8	85.7
运动	44.4	55.5
钱	25.0	70.0

most difficult words for the SMT model consist of a single character. Eight out of the 20 target words considered consist of a unique character, and appear as such in the test set, while these characters were typically segmented within longer words in the parallel training corpus. However, this is of course not the only reason for the difference in accuracies, as the WSD system also signficantly outperforms the SMT model on target words that consist of more than 1 character.

6.4 A dedicated unsupervised WSD model also outperforms SMT

One might speculate that the difference in performance obtained with SMT vs. WSD models can be explained by the fact that we are essentially comparing unsupervised models with fully supervised models.

To address this we can again take advantage of the Senseval framework, and compare the performance of our SMT system with other published results on the same dataset. The system described in (Cabezas *et al.*, 2004) is of particular interest as it uses an unsupervised approach. An unsupervised Chinese-English bilingual WSD model is learned from automatically word-aligned parallel corpora. In order to use this bilingual model, the Chinese lexical sample task is artificially converted into a bilingual task, by automatically translating the Chinese test sentences into English, using an alignment-template based SMT system.

This unsupervised, but dedicated, WSD model yields an accuracy of 44.5%, thus outperforming all the SMT model variations. It yields a 35% relative improvement over the best SMT model, which remains relatively little compared to the best supervised dedicated WSD system, which doubles the accuracy score of the SMT model.

7 Related work

To the best of our knowledge, this is the first evaluation of SMT models on standard WSD performance metrics and datasets. One might argue that traditional MT evaluation metrics such as word error rate (WER) also evaluate the WSD performance of MT models. WER is defined as the percentage of words to be inserted, deleted or replaced in the translation in order to obtain the sentence of reference. However, WER does not isolate WSD performance since it also encompasses many other types of errors. Also, since the choice of a translation for a particular word affects the translation of other words in the sentence, the effect of WSD performance on WER is unclear. In contrast, the Senseval accuracy metric counts each incorrect translation choice only once.

Apart from the voted WSD system described in section 4, and the unsupervised system (Cabezas *et al.*, 2004) mentioned in section 6.4, systems built and optimized for the Senseval-3 Chinese lexical sample task, include Niu *et al.* (2004). Many of the Senseval-type

WSD system are not language specific and the presentation of the results in the English lexical sample task (Midhalcea *et al.*, 2004), English-Hindi multilingual task (Chklovski *et al.*, 2004), or any of the lexical sample tasks defined in other languages, give a good overview of the variety of approaches to WSD.

Most previous work on multilingual WSD has focused on the different problem of exploiting bilingual resources (e.g., parallel or comparable corpora, or even full MT systems) to help WSD. For instance, Ng *et al.* (2003) showed that it is possible to use word aligned parallel corpora to train accurate supervised WSD models. Other work includes Li and Li (2002) who propose a bilingual bootstrapping method to learn a translation disambiguation WSD model, and Diab (2004) who exploited large amounts of automatically generated noisy parallel data to learn WSD models in an unsupervised bootstrapping scheme. In all this work, the goal is to achieve accurate WSD with minimum amounts of annotated data. Again, this differs from our objective which is to directly evaluate an SMT model as a WSD model.

8 Conclusion

We presented empirical results casting doubt on the increasingly common assumption that SMT models are very good at WSD, even though they do not explicitly address WSD as an independent task.

Using the Senseval-3 Chinese lexical sample task as a testbed, we directly compared the performance of a typical Chinese-to-English SMT model, built from off-the-shelf toolkits, with that of state-of-the-art Senseval models and found that the SMT model does not achieve the same high accuracies as any the dedicated WSD models considered. Even after attempting to compensate for the difference between training and evaluation data in favor of the SMT model, the accuracy of the SMT model is still significantly lower than that of the dedicated WSD systems.

Error analysis confirms the weaknesses of the SMT models for the WSD task. Unlike dedicated WSD models, SMT models only rely on the local context to make translation choices, and tend to prefer phrasal cohesion in the target language and fluency, rather than adequacy of the translation of each source word.

These results cast doubt on the speculative claim that SMT systems do not need sophisticated WSD models, and suggest on the contrary that the predictions of the dedicated models should be useful. Puzzlingly, in converse experiments, using a state-of-the-art WSD model to choose translation candidates for a typical IBM SMT system, we find that WSD does *not* yield significantly better translation quality than the SMT system alone (Carpuat and Wu, 2005). Taken together, these results suggest that another SMT formulation might be needed. In particular, more grammatically structured statistical MT models that are better equipped to handle long distance dependencies, such as the ITG based "grammatical channel" translation model (Wu and Wong, 1998), might make better use of the WSD predictions.

References

Clara Cabezas, Indrajit Bhattacharya, and Philip Resnik. The university of maryland senseval-3 system descriptions. In *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 83–87, Barcelona, Spain, July 2004. SIGLEX, Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 387–394, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Marine Carpuat, Weifeng Su, and Dekai Wu. Augmenting ensemble classifi cation for word sense disambiguation with a kernel pca model. In *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems*, Barcelona, July 2004. SIGLEX, Association for Computational Linguistics.

Timothy Chklovski, Rada Midhalcea, Ted Pedersen, and Amruta Purandare. The senseval-3 multilingual english-hindi lexical sample task. In *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 5–8, Barcelona, Spain, July 2004. SIGLEX, Association for Computational Linguistics.

Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings of Eurospeech '97*, pages 2707–2710, Rhodes, Greece, 1997.

Mona Diab. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.

Yoram Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Journal of Computer and System Sciences, 55(1)*, pages 119–139, 1997.

Ulrich Germann. Greeedy decoding for statistical machine translation in almost linear time. In *Proceedings of HLT-NAACL-2003. Edmonton, AB, Canada*, 2003.

E.T. Jaynes. *Where do we Stand on Maximum Entropy?* MIT Press, Cambridge MA, 1978.

Cong Li and Hang Li. Word translation disambiguation using bilingual bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 343–351, 2002.

Rada Midhalcea, Timothy Chklovski, and Adam Killgariff. The senseval-3 english lexical sample task. In *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 25–28, Barcelona, Spain, July 2004. SIGLEX, Association for Computational Linguistics.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of ACL-03, Sapporo, Japan*, pages 455–462, 2003.

Zheng-Yu Niu, Dong-Hong Ji, and Chew-Lim Tan. Optimizing feature set for chinese word sense disambiguation. In *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems*, Barcelona, July 2004. SIGLEX, Association for Computational Linguistics.

Franz Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 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.

Dekai Wu and Hongsing Wong. Machine translation with a stochastic grammatical channel. In *Proceedings of COLING-ACL'98*, Montreal,Canada, August 1998.

Dekai Wu, Weifeng Su, and Marine Carpuat. A kernel pca method for superior word sense disambiguation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, July 2004.

David Yarowsky and Radu Florian. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310, 2002.