Lightly-Supervised Word Sense Translation Error Detection for an Interactive Conversational Spoken Language Translation System

Dennis N. Mehay, Sankaranarayanan Ananthakrishnan and Sanjika Hewavitharana
Speech, Language and Multimedia Processing Unit
Raytheon BBN Technologies
Cambridge, MA, 02138, USA
{dmehay,sanantha,shewavit}@bbn.com

Abstract

Lexical ambiguity can lead to concept transfer failure in conversational spoken language translation (CSLT) systems. This paper presents a novel, classification-based approach to accurately detecting word sense translation errors (WSTEs) of ambiguous source words. The approach requires minimal human annotation effort, and can be easily scaled to new language pairs and domains, with only a word-aligned parallel corpus and a small set of manual translation judgments. We show that this approach is highly precise in detecting WSTEs, even in highly skewed data, making it practical for use in an interactive CSLT system.

1 Introduction

Lexical ambiguity arises when a single word form can refer to different concepts. Selecting a contextually incorrect translation of such a word — here referred to as a *word sense translation error* (WSTE) — can lead to a critical failure in a conversational spoken language translation (CSLT) system, where accuracy of concept transfer is paramount. Interactive CSLT systems are especially prone to mis-translating less frequent word senses, when they use phrase-based statistical machine translation (SMT), due to its limited use of source context (source phrases) when constructing translation hypotheses. Figure 1 illustrates a typical WSTE in a phrase-based English-to-Iraqi Arabic CSLT system, where the English word *board*

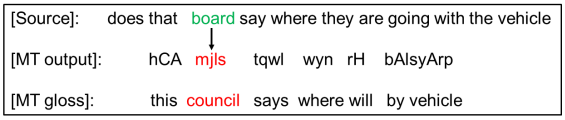

Figure 1: Example WSTE in English-to-Iraqi SMT.

is mis-translated as *mjls* ("*council*"), completely distorting the intended message.

Interactive CSLT systems can mitigate this problem by automatically detecting WSTEs in SMT hypotheses, and engaging the operator in a clarification dialogue (e.g. requesting an unambiguous rephrasing). We propose a novel, two-level classification approach to accurately detect WSTEs. In the first level, a bank of word-specific classifiers predicts, given a rich set of contextual and syntactic features, a distribution over possible target *translations* for each ambiguous source word in our inventory. A single, second-level classifier then compares the predicted target words to those chosen by the decoder and determines the likelihood that an error was made.

A significant novelty of our approach is that the first-level classifiers are fully unsupervised with respect to manual annotation and can easily be expanded to accommodate new ambiguous words and additional parallel data. The other innovative aspect of our solution is the use of a small set of manual translation judgments to train the second-level classifier. This classifier uses high-level features derived from the output of the first-level classifiers to produce a binary WSTE prediction, and can be re-used unchanged even when the first level of classifiers is expanded.

Our goal departs from the large body of work devoted to lightly-supervised word sense disambiguation (WSD) using monolingual and bilingual corpora (Yarowsky, 1995; Schutze, 1998; Diab and Resnik, 2002; Ng et al., 2003; Li and Li, 2002; Purandare and Pedersen, 2004), which seeks to la-

Disclaimer: This paper is based upon work supported by the DARPA BOLT program. The views expressed here are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Distribution Statement A (Approved for Public Release, Distribution Unlimited)

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 54–58,
Gothenburg, Sweden, April 26-30 2014.

bel and group unlabeled sense instances. Instead, our approach detects *mis-translations* of a known set of ambiguous words.

The proposed method also deviates from existing work on global lexical selection models (Mauser et al., 2009) and on integration of WSD features within SMT systems with the goal of improving offline translation performance (Chan et al., 2007). Rather, we *detect* translation errors due to ambiguous source words with the goal of providing feedback to and soliciting clarification from the system operator in real time. Our approach is partly inspired by Carpuat and Wu's (2007b; 2007a) unsupervised sense disambiguation models for offline SMT. More recently, Carpuat et al. (2013) identify unseen target senses in new domains, but their approach requires the full test corpus upfront, which is unavailable in spontaneous CSLT. Our approach can, in principle, identify novel senses when unfamiliar source contexts are encountered, but this is not our current focus.

2 Baseline SMT System

In this paper, we focus on WSTE detection in the context of phrase-based English-to-Iraqi Arabic SMT, an integral component of our interactive, two-way CSLT system that mediates conversation between monolingual speakers of English and Iraqi Arabic. The parallel training corpus of approximately 773K sentence pairs (7.3M English words) was derived from the DARPA TransTac English-Iraqi two-way spoken dialogue collection and spans a variety of domains including force protection, medical diagnosis and aid, etc. Phrase pairs were extracted from bidirectional IBM Model 4 word alignment after applying a merging heuristic similar to that of Koehn et al. (2003). A 4-gram target LM was trained on Iraqi Arabic transcriptions. Our phrase-based decoder, similar to Moses (Koehn et al., 2007), performs beam search stack decoding based on a standard log-linear model, whose parameters were tuned with MERT (Och, 2003) on a held-out development set (3,534 sentence pairs, 45K words). The BLEU and METEOR scores of this system on a separate test set (3,138 sentence pairs, 38K words) were 16.1 and 42.5, respectively.

3 WSTE Detection

The core of the WSTE detector is a novel, two-level classification pipeline. Our approach avoids the need for expensive, sense-labeled training data based on the observation that knowing the *sense* of an ambiguous source word is distinct from knowing whether a *sense translation error* has occurred. Instead, the target (Iraqi Arabic) words typically associated with a given sense of an ambiguous source (English) word serve as implicit sense labels, as the following describes.

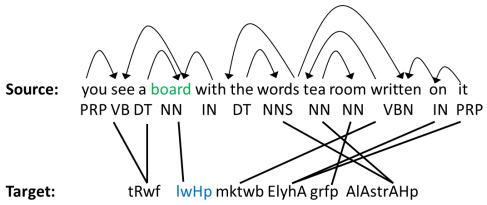

Figure 2: An English–Iraqi training pair.

3.1 A First Level of Unsupervised Classifiers

The main intuition behind our approach is that strong disagreement between the expanded context of an ambiguous source word and the corresponding SMT hypothesis indicates an increased likelihood that a WSTE has occurred. To identify such disagreement, we train a bank of maximum-entropy classifiers (Berger et al., 1996), one for each ambiguous word. The classifiers are trained on the same word-aligned parallel data used for training the baseline SMT system, as follows.

For each instance of an ambiguous source word in the training set, and for each target word it is aligned to, we emit a training instance associating that target word and the wider source context of the ambiguous word. Figure 2 illustrates a typical training instance for the ambiguous English word *board*, which emits a tuple of contextual features and the aligned Iraqi Arabic word *lwHp* ("*placard*") as a target label. We use the following contextual features similar to those of Carpuat and Wu (2005), which are in turn based on the classic WSD features of Yarowsky (1995).

Neighboring Words/Lemmas/POSs. The tokens, t, to the left and right of the current ambiguous token, as well as all trigrams of tokens that span the current token. Separate features for word, lemma and parts of speech tokens, t.

Lemma/POS Dependencies. The lemma-lemma and POS-POS labeled and unlabeled directed syntactic dependencies of the current ambiguous token.

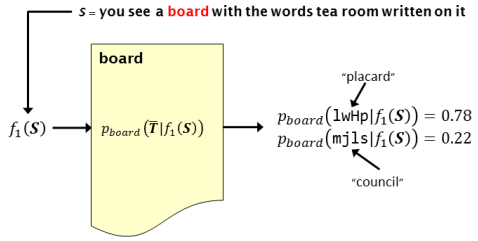

Figure 3: An unsupervised first-level classifier.

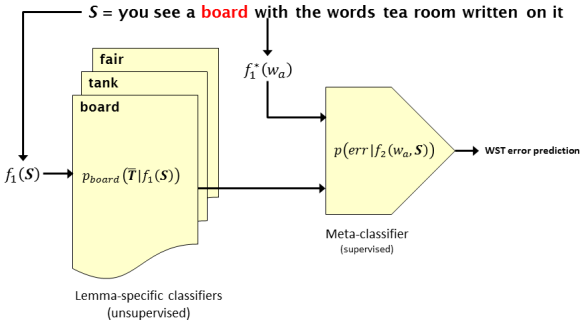

Figure 4: The two-level WSTE architecture.

Bag-of-words/lemmas. Distance decayed bag-of-words-style features for each word and lemma in a seven-word window around the current token.

Figure 3 schematically illustrates how this classifier operates on a sample test sentence. The example assumes that the ambiguous English word *board* is only ever associated with the Iraqi Arabic words *lwHp (“placard”)* and *mjls (“council”)* in the training word alignment. We emphasize that even though the first-level maximum entropy classifiers are intrinsically supervised, their training data is derived via unsupervised word alignment.

3.2 A Second-Level Meta-Classifier

The first-level classifiers do not directly predict the presence of a WSTE, but induce a distribution over possible target words that could be generated by the ambiguous source word in that context. In order to make a binary decision, this distribution must be contrasted with the corresponding target phrase hypothesized by the SMT decoder. One straightforward approach, which we use as a baseline, is to threshold the posterior probability of the word in the SMT target phrase which is ranked highest in the classifier-predicted distribution. However, this approach is not ideal because each classifier has a different target label set and is trained on a different number of instances.

To address this issue, we introduce a second meta-classifier, which is trained on a small number of hand-annotated translation judgments of SMT hypotheses of source sentences containing ambiguous words. The bilingual annotator was simply asked to label the phrasal translation of source phrases containing ambiguous words as *correct* or *incorrect*. We obtained translation judgments for 511 instances from the baseline SMT development and test sets, encompassing 147 pre-defined ambiguous words obtained heuristically from WordNet, public domain homograph lists, etc.

The second-level classifier is trained on a small set of meta-features drawn from the predictions of the first-level classifiers and from simple statistics of the training corpus. For an ambiguous word w_a in source sentence S, with contextual features $f_1(S)$, and aligned to target words $t \in T$ (the set of words in the target phrase) in the SMT hypothesis, we extract the following features:

1. The first-level classifier’s maximum likelihood of any decoded target word: $\max_{t \in T} p_{w_a}(t|f_1(S))$
2. The entropy of the predicted distribution: $\sum_t p_{w_a}(t|f_1(S)) \cdot ln(p_{w_a}(t|f_1(S)))$
3. The number of training instances for w_a
4. The inverse of the number of distinct target labels for w_a.
5. The product of meta-features (1) and (4)

A high value for feature 1 indicates that the first-level model and the SMT decoder agree. By contrast, a high value for feature 2 indicates uncertainty in the classifier’s prediction, due either to a novel source context, or inadequate training data. Feature 3 indicates whether the second scenario of meta-feature 2 might be at play, and feature 4 can be thought of as a simple, uniform prior for each classifier. Finally, feature 5 attenuates feature 1 by this simple, uniform prior. We feed these features to a random forest (Breiman, 2001), which is a committee of decision trees, trained using randomly selected features and data points, using the implementation in Weka (Hall et al., 2009). The target labels for training the second-level classifier are obtained from the binary translation judgments on the small annotated corpus. Figure 4 illustrates the interaction of the two levels of classification.

3.3 Scalability and Portability

Scalability was an important consideration in designing the proposed WSTE approach. For instance, we may wish to augment the inventory with new ambiguous words if the vocabulary grows due to addition of new parallel data or due to a change in the domain. The primary advantage of the two-level approach is that new ambiguous words can be accommodated by augmenting the unsupervised first-level classifier set with additional word-specific classifiers, which can be done by simply extending the pre-defined list of ambiguous words. Further, the current classification stack requires only ≈1.5GB of RAM and performs per-word WSTE inference in only a few milliseconds on a commodity, quad-core laptop, which is critical for real-time, interactive CSLT.

The minimal annotation requirements also allow a high level of *portability* to new language pairs. Moreover, as our results indicate (below), a good quality WSTE detector can be bootstrapped for a new language pair *without any annotation effort* by simply leveraging the first-level classifiers.

4 Experimental Results

The 511 WSTE-annotated instances used for training the second-level classifier doubled as an evaluation set using the leave-one-out cross-validation method. Of these, 115 were labeled as errors by the bilingual judge, while the remaining 396 were translated correctly by the baseline SMT system. The error prediction score from the second-level classifier was thresholded to obtain the receiver operating characteristic (ROC) curve shown in the top (black) curve of Figure 5. We obtain a 43% error detection rate with only 10% false alarms and 71% detection with 20% false alarms, in spite of the highly skewed label distribution. In absolute terms, true positives outnumber false alarms at both the 10% (49 to 39) and 20% (81 to 79) false alarm rates. This is important for deployment, as we do not want to disrupt the flow of conversation with more false alarms than true positives.

For comparison, the bottom (red) ROC curve shows the performance of a baseline WSTE predictor comprised of just meta-feature (1), obtainable directly from the first-level classifiers. This performs slightly worse than the two-level model at 10% false alarms (40% detection, 46 true positives, 39 false alarms), and considerably worse at 20% false alarms (57% detection, 66 true positives, 78 false alarms). Nevertheless, this result indicates the possibility of bootstrapping a good quality baseline WSTE detector in a new language or domain without any annotation effort.

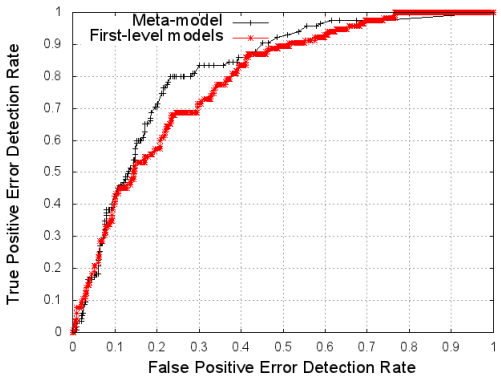

Figure 5: WST error detection ROC curve.

5 Conclusion

We proposed a novel, lightly-supervised, two-level classification architecture that identifies possible mis-translations of pre-defined ambiguous source words. The WSTE detector pre-empts communication failure in an interactive CSLT system by serving as a trigger for initiating feedback and clarification. The first level of our detector comprises of a bank of word-specific classifiers trained on automatic word alignment over the SMT parallel training corpus. Their predicted distributions over target words feed into the second-level meta-classifier, which is trained on a small set of manual translation judgments. On a 511-instance test set, the two-level approach exhibits WSTE detection rates of 43% and 71% at 10% and 20% false alarm rates, respectively, in spite of a nearly 1:4 skew against actual WSTE instances.

Because adding new ambiguous words to the inventory only requires augmenting the set of first-level unsupervised classifiers, our WSTE detection approach is *scalable* to new domains and training data. It is also easily *portable* to new language pairs due to the minimal annotation effort required for training the second-level classifier. Finally, we show that it is possible to bootstrap a good quality WSTE detector in a new language pair *without any annotation effort* using only unsupervised classifiers and a parallel corpus.

References

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

Leo Breiman. 2001. Random Forests. Technical report, Statistics Department, University of California, Berkeley, Berkeley, CA, USA, January.

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 387–394.

Marine Carpuat and Dekai Wu. 2007a. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skovde, Sweden, September.

Marine Carpuat and Dekai Wu. 2007b. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic, June.

Marine Carpuat, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013. Sensespotting: Never let your parallel data tie you to an old domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1435–1445, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic, June.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 255–262, July.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Arne Mauser, Saša Hasan, and Hermann Ney. 2009. Extending statistical machine translation with discriminative and trigger-based lexicon models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 210–218, Singapore, August. Association for Computational Linguistics.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of 41st Annual Meeting on Association for Computational Linguistics*, pages 455–462, July.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Amruta Purandare and Ted Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 41–48.

Hinrich Schutze. 1998. Automatic word sense discrimination. *Journal of Computational Linguistics*, 24:97–123.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.