Word Sense Induction for Better Lexical Choice

Neha Prabhugaonkar[1], Jyoti Pawar[1] and Pushpak Bhattacharyya[2]

[1] Department of Computer Science and Technology,
Goa University, Goa
nehapgaonkar.1920@gmail.com, jyotidpawar@gmail.com
[2] Department of Computer Science and Engineering,
IIT Bombay, Powai
pb@cse.iitb.ac.in

Abstract. Most words in natural languages are *polysemous* in nature that is they have multiple possible meanings or *senses*. The sense in which the word is used determines the translation of the word. We show that incorporating a sense-based translation model into statistical machine translation model consistently improves translation quality across all different test sets of five different language-pairs, according to all eight most commonly used evaluation metrics. This paper is an investigation on how to initiate research in word sense disambiguation and statistical machine translation for under-resourced languages by applying *Word Sense Induction*.

1 Introduction

Word Sense Disambiguation or WSD is the ability to identify the meaning of words in context in a computational manner [1]. A wide variety of approaches ranging from supervised to unsupervised algorithms have been proposed. Supervised approaches ([2] and [3]) which rely on sense annotated corpora have proven to be more successful, and they substantially outperform knowledge-based and unsupervised approaches ([4] and [5]). However, creation of sense annotated corpora is always costly and time-consuming, especially for the resource scarce languages.

1.1 Use of WSD models in SMT

WSD is often assumed to be an intermediate task, which should then help higher level applications such as Machine Translation or Information Retrieval. However, WSD is usually performed and evaluated as a standalone task but there have been very few efforts to integrate the learned WSD models into full SMT systems. Some of the reasons are:

– Most of the WSD approaches assign senses with the aid of dictionaries, or other lexical resources such as WordNet; it is difficult to adapt them to new domains or to languages where such resources are scarce.

- A related problem concerns the granularity of the sense distinctions which is fixed, and may not be entirely suitable for different applications [6].
- There is a risk that an important sense will be missed, or an irrelevant sense will influence the results.
- In many cases, lexical resources like WordNet is very precise, defining senses which are similar and hard to distinguish.

1.2 Why WSI for SMT?

Initially, WSD was mainly applied and developed on English texts, because of the broad availability and the prevalence of lexical resources compared to other languages. Due to the lack of availability of large lexical resources i.e. sense inventories (dictionaries, lexical databases, WordNets, etc.) and parallel sense-tagged corpora it is difficult to start working on WSD for under-resourced languages (Tamil, Konkani, Telugu, etc.). To account for under-resourced languages, one can easily adopt techniques aimed at the automatic discovery of word senses from text, a task called Word Sense Induction.

Word Sense Induction (WSI) is a task of automatically inducing the underlying senses of word tokens given the surrounding contexts where the word tokens occur. The biggest difference from word sense disambiguation lies in that WSI does not rely on a predefined sense inventory.

Recent work in Machine Translation ([7] and [8]) and Information Retrieval [9] indicates that induced senses can lead to substantial improvement in performance where methods based on a fixed sense inventory such as HowNet have previously failed ([10] and [11]). Therefore, We adopt the similar approach of Xiong and Zhang [8] by resorting to Word Sense Induction (WSI) that is related to but different from WSD.

The advantages of using WSI are:

- It actually performs word sense disambiguation.
- Aims to divide the occurrence of a word into a number of classes.
- Makes objective evaluation easy if it is domain-specific.

The rest of the paper is structured as follows: Section 2 describes the Related work. In Section 3, we describe the SMT system and its essential components. In Section 4, we provide details about the experiments conducted and results obtained. Finally, Section 5 concludes the paper.

2 Related Work

2.1 Standard WSD for SMT

Carpuat and Wu [10] integrated the translation predictions from a state-of-the-art Chinese WSD system [12] into a Chinese-English word-based SMT system using the ISI ReWrite decoder [13]. They used the WSD model predictions either *to substitute for translation candidates of their translation model* or *to post edit the output of their SMT system.* The authors reported that WSD does not yield significantly better translation quality than the SMT system alone.

2.2 Redefined WSD for SMT

Vickrey et al., [7], redefined the standard WSD problem for SMT as a *word translation task* - predicting possible target translations rather than senses for ambiguous source words. The translation choices for a word w were defined as the set of words or phrases aligned to w, as gathered from a word-aligned parallel corpus. The authors reported that they were able to improve their models accuracy on a simplified word translation task.

Chan et al., [14], successfully integrated a state-of-the-art WSD system into a state-of-the-art Hierarchical phrase-based system, Hiero [15]. They introduced two WSD-related additional features into the log-linear model of SMT. Carpuat and Wu [10] also used the redefined WSD for SMT and further adapted it for multi-word phrasal disambiguation. They both reported that redefined WSD system improves the performance of a state-of-the-art SMT system on actual translation task.

Although the redefined WSD has proved helpful for SMT, recently, Xiong and Zhang [8] re-investigated the question of whether pure senses are useful for SMT by using WSI. They proposed a sense-based translation model to integrate word senses into SMT which enables the decoder to select appropriate translations for the source words according to the inferred senses for these words using Maximum Entropy classifiers. The authors reported that the proposed model substantially outperforms not only baseline but also the previously redefined WSD.

3 The SMT system

To build a representative baseline SMT system, we restricted ourselves to making use of freely available tools. Since our focus is not on a specific SMT architecture, we used the `cdec`[3] [16] toolkit trained in a standard fashion for our experiments. The detailed architecture of the SMT system is shown in Figure 1.

3.1 Data Preprocessing

We preprocess the source side of our bilingual training data as well as development and test set by removing stop words and rare words. From the preprocessed training data, we extract all possible pseudo documents for each source word type. The collection of these extracted pseudo documents is used as a corpus to train a HDP-based WSI model for the source word type. In this way, we can train as many HDP-based WSI models as the number of word types kept after preprocessing.

3.2 Sense Annotation

To obtain word senses for any source words, we build a sense tagger that relies on the nonparametric Bayesian model based word sense induction ([17], [18]) similar to Xiong and Zhang [8]. We used HDP-based WSI[4] [19] to predict sense

[3] `http://www.cdec-decoder.org/`

[4] `http://www.cs.cmu.edu/\~chongw/resource.html`

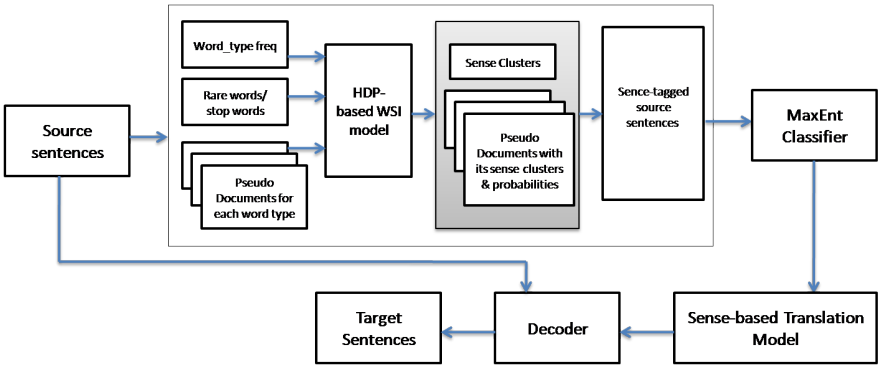

Fig. 1. Architecture of SMT system

clusters and to annotate source words in our training/development/test sets with these sense clusters. We individually build a HDP-based WSI model per word type and train these models on the training data. The sense for a word token is defined as the most probable sense according to the per-document sense distribution estimated for the corresponding pseudo document that represents the surrounding context of the word token.

3.3 Alignment Model

The alignment model was trained with `fast-align` alignment tool which is a variant of the aligner proposed by Dyer et al., [20]. The alignment algorithm is trained in either direction and are symmetrized using grow-diag-final heuristics.

3.4 Language Model

The Hindi language model is a five gram model trained on the **Hindi** side of the parallel corpora using a publicly available software, the KenLM[5] [21] toolkit. We used additional monolingual corpora[6] [22] of ≈45 million lines and included more Hindi monolingual corpora[7] for language model training.

3.5 Sense-based Translation Model

The sense-based translation model estimates the probability that a source word c is translated into a target phrase e given contextual information, i.e. word senses that are obtained using the HDP-based WSI. We adopt the same approach of Xiong and Zhang [8] to build the sense-based Translation Model.

[5] `https://kheafield.com/code/kenlm/`
[6] `https://lindat.mff.cuni.cz/repository/xmlui/browse?value=hin\&type=language`
[7] `http://www.cfilt.iitb.ac.in/wsd/annotated\_corpus/`

Table 1. WSI-based SMT improves BLEU, GTM-3, NIST, WER, PER, TER, METEOR and ROUGE-L across all language-pair datasets.

Lang-pair	Expts	BLEU	GTM-3	NIST	WER	PER	TER	METEOR	ROUGE
Eng-Hin	SMT	0.2619	0.3253	5.8787	0.649	0.5155	0.6346	0.2581	0.0727
	WSI+SMT	0.2747	0.3394	6.1792	0.62	0.4873	0.6021	0.2665	0.0771
Ben-Hin	SMT	0.3674	0.4063	7.4327	0.4424	0.3918	0.4537	0.315	0.1008
	WSI+SMT	0.3761	0.4151	7.5075	0.4347	0.3856	0.4479	0.3184	0.1004
Mar-Hin	SMT	0.4096	0.4231	7.8353	0.4211	0.3866	0.4265	0.3335	0.1172
	WSI+SMT	0.4156	0.4319	7.5009	0.4581	0.4262	0.4475	0.3526	0.118
Tam-Hin	SMT	0.2057	0.2386	5.1119	0.671	0.5334	0.6544	0.223	0.0967
	WSI+SMT	0.2157	0.2529	4.8329	0.7222	0.5843	0.7067	0.2386	0.096
Tel-Hin	SMT	0.2822	0.3415	5.8713	0.606	0.556	0.592	0.288	0.1215
	WSI+SMT	0.2976	0.3556	6.4549	0.5088	0.4666	0.5208	0.273	0.1218

4 Experimental Details

4.1 Datasets and Resources Used

We used five different language pairs in our experiments - representing a wide range of diversities, such as language family (Indo-Aryan: Hindi, Bengali and Marathi, Dravidian: Tamil and Telugu and West Germanic: English), languages with high structural divergence and morphological manifestations (English is structurally classified as a Subject-Verb-Object (SVO) language with poor morphology whereas Hindi is a morphologically rich, Subject-Object-Verb (SOV) language), etc. The target language for all the languages is **Hindi**.

The datasets belonged to the tourism and health domains (25,000+25,000 sentences) from the ILCI corpora. We normalized the corpus to solve issues related to incorrect characters, redundant Unicode representation of some Indic characters, etc. The English corpus was tokenized using the Stanford tokenizer[8] [23] and for Indian languages, we used NLP Indic Library[9] [24].

For every language pair, the corpus was split up as follows: training set of 48000 sentences, development test set of 1000 sentences and test set of 1000 sentences. The training, development test and test splits are completely parallel across the five language-pairs involved.

[8] `http://nlp.stanford.edu/software/`

[9] `https://github.com/anoopkunchukuttan/indic\_nlp\_library`

Table 2. Examples of translations drawn from the English-Hindi test set.

Example 1	Input	in Delhi many types of food of India and abroad are served
	Sense-based SMT output / Reference	दिल्ली में भारत और विदेशों के अनेक प्रकार के भोजन परोसे जाते है ।
Example 2	Input	this medicine is mainly used for ulcer , asthma and bronchitis.
	Sense-based SMT output / Reference	इस औषधि का विशेष रूप से प्रयोग अल्सर , दमा और ब्रोंकाटिस के लिये किया जाता है ।
Example 3	Input	the journey of namdapha is easy and also inexpensive .
	Baseline	नमदफा का **सफर** भी आसान और सस्ता है ।
	Sense-based SMT output / Reference	नमदफा की **यात्रा** आसान है और सस्ती भी ।
Example 4	Input	along with sunrise the stir of the devotees start at the ramghat .
	Baseline	सूर्योदय के साथ ही रामघाट पर श्रद्धालुओ की हलचल **शुरू** हो जाता है ।
	Sense-based SMT output / Reference	सूर्योदय के साथ ही रामघाट पे श्रद्धालुओ की हलचल **आरंभ** हो जाती है
Example 5	Input	shampoo a little while after the massage .
	Baseline	मसाज के थोडी देर बाद शेम्पू ।
	Sense-based SMT output / Reference	मसाज के थोडी देर बाद शेम्पू **कर ले** ।

Table 3. Number of translations which exactly matched with the reference sentences.

Language-pair	**Baseline SMT**	**WSI-based SMT**	**Overlap between Baseline & WSI-based SMT**
English-Hindi	39	44	11
Bengali-Hindi	63	66	23
Marathi-Hindi	57	59	38
Tamil-Hindi	16	19	9
Telugu-Hindi	29	29	19

4.2 Results and Analysis

As mentioned, our experiments were on Indian language (Bengali, Marathi, Tamil, Telugu) to Hindi translation and English to Hindi translation. To measure the impact of using sense-based Translation Model on translation quality, we used the most commonly used automatic evaluation metrics to evaluate the translations obtained. Apart from the widely used BLEU [25] and NIST [26], we also evaluate translation quality with METEOR [27] without using WordNet synonyms to match translation candidates and references, General Text Matcher (GTM-3), Word Error Rate (WER), Position-independent word Error Rate (PER), Translation Edit Rate (TER) [28] and ROUGE. These metrics have proved to relate well with both adequacy and fluency. The results are shown in Table 1.

Using sense-based Translation Model in SMT yields better translation quality on all language-pair test sets, as measured by all eight commonly used automatic evaluation metrics.

Table 2 show examples of translations drawn from the English-Hindi test set. Analysis says that WSI-based translation model helps decoder to give better rankings and lexical choices then the baseline translation probabilities (see Example 3 and 4). Examples 1-5 are the translations which exactly matched with reference sentences. We came across many such examples where the lexical item proposed by the WSI-based translation model was better than the baseline system which resulted in increase in performance of the MT system.

5 Conclusion and Future Work

We have shown that sense-based Translation Model improves the translation performance of an Indian language SMT system and its improvement is statistically significant in terms of all eight evaluation metrics. Word senses induced automatically by the HDP-based WSI are very useful for Machine Translation for under-resourced languages. The sense-based Translation Model in SMT is effective at choosing the correct and appropriate lexical choice for an ambiguous word.

Our future work will be to build a sense-based Hindi language model by inducing sense clusters for words in the target language. We would also like to explore whether integrating learned WSD Model in SMT for same Indian language-pairs improves translation quality or not and perform a comparative study.

References

1. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. **41** (2009) 10:1–10:69
2. Ng, H.T., Lee, H.B.: Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In: Proceedings of the 34th Annual Meeting

on Association for Computational Linguistics. ACL '96, Stroudsburg, PA, USA, Association for Computational Linguistics (1996) 40–47
3. Lee, Y., Ng, H., Chia, T.: Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources. In: Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. (2004) 137–140
4. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation. SIGDOC '86, New York, NY, USA, ACM (1986) 24–26
5. Mihalcea, R.: Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 411–418
6. Brody, S., Lapata, M.: Bayesian word sense induction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. EACL '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 103–111
7. Vickrey, D., Biewald, L., Teyssier, M., Koller, D.: Word-sense disambiguation for machine translation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), Vancouver, Canada (2005)
8. Xiong, D., Zhang, M.: A sense-based translation model for statistical machine translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, Association for Computational Linguistics (2014) 1459–1469
9. Veronis, J.: Hyperlex: lexical cartography for information retrieval. Computer Speech and Language **18** (2004) 223–252
10. Carpuat, M., Wu, D.: Word sense disambiguation vs. statistical machine translation. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 387–394
11. Voorhees, E.M.: Using wordnet to disambiguate word senses for text retrieval. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '93, New York, NY, USA, ACM (1993) 171–180
12. Wu, D., Su, W., Carpuat, M.: A kernel pca method for superior word sense disambiguation. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. ACL '04, Stroudsburg, PA, USA, Association for Computational Linguistics (2004)
13. Germann, U.: Greedy decoding for statistical machine translation in almost linear time. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. NAACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 1–8
14. Chan, Y.S., Ng, H.T., Chiang, D.: Word sense disambiguation improves statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, Association for Computational Linguistics (2007) 33–40

15. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 263–270
16. Dyer, C., Weese, J., Setiawan, H., Lopez, A., Ture, F., Eidelman, V., Ganitkevitch, J., Blunsom, P., Resnik, P.: Cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In: Proceedings of the ACL 2010 System Demonstrations. ACLDemos '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 7–12
17. Yao, X., Durme, B.V.: Nonparametric bayesian word sense induction. In: Graph-based Methods for Natural Language Processing, The Association for Computer Linguistics (2011) 10–14
18. Lau, J.H., Cook, P., McCarthy, D., Newman, D., Baldwin, T.: Word sense induction for novel sense detection. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. EACL '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 591–601
19. Wang, C., Blei, D.M.: A split-merge mcmc algorithm for the hierarchical dirichlet process. CoRR **abs/1201.1657** (2012)
20. Dyer, C., Chahuneau, V., Smith, A.N.: A simple, fast, and effective reparameterization of ibm model 2. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics (2013) 644–648
21. Heafield, K.: Kenlm: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. WMT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 187–197
22. Bojar, O., Diatka, V., Rychly, P., Stranak, P., Suchomel, V., Tamchyna, A., Zeman, D.: Hindencorp - hindi-english and hindi-only corpus for machine translation. In Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014)
23. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1. ACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 423–430
24. Kunchukuttan, A., Mishra, A., Chatterjee, R., Shah, R.M., Bhattacharyya, P.: Shata-anuvadak: Tackling multiway translation of indian languages. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014. (2014) 1781–1787
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 311–318
26. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research. HLT '02, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2002) 138–145
27. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or

Summarization, Ann Arbor, Michigan, Association for Computational Linguistics (2005) 65–72
28. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: In Proceedings of Association for Machine Translation in the Americas. (2006) 223–231