One Translation per Discourse

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

Abstract

We revisit the one sense per discourse hypothesis of Gale *et al.* in the context of machine translation. Since a given sense can be lexicalized differently in translation, do we observe one translation per discourse? Analysis of manual translations reveals that the hypothesis still holds when using translations in parallel text as sense annotation, thus confirming that translational differences represent useful sense distinctions. Analysis of Statistical Machine Translation (SMT) output showed that despite ignoring document structure, the one translation per discourse hypothesis is strongly supported in part because of the low variability in SMT lexical choice. More interestingly, cases where the hypothesis does not hold can reveal lexical choice errors. A preliminary study showed that enforcing the one translation per discourse constraint in SMT can potentially improve translation quality, and that SMT systems might benefit from translating sentences within their entire document context.

1 Introduction

The one sense per discourse hypothesis formulated by Gale *et al.* (1992b) has proved to be a simple yet powerful observation and has been successfully used in word sense disambiguation (WSD) and related tasks (e.g., Yarowsky (1995); Agirre and Rigau (1996)). In this paper, we investigate its potential usefulness in the context of machine translation.

A growing body of work suggests that translational differences represent observable sense distinctions that are useful in applications. In monolingual WSD, word alignments in parallel corpora have been successfully used as learning evidence (Resnik and Yarowsky, 1999; Diab and Resnik, 2002; Ng *et al.*, 2003). In Statistical Machine Translation (SMT), recent work shows that WSD helps translation quality when the WSD system directly uses translation candidates as sense inventories (Carpuat and Wu, 2007; Chan *et al.*, 2007; Giménez and Màrquez, 2007).

In this paper, we revisit the one sense per discourse hypothesis using word translations in parallel text as senses. Our first goal is to empirically evaluate whether the one translation per document hypothesis holds on French-English reference corpora, thus verifying whether translations exhibit the same properties as monolingual senses. Our second goal consists in evaluating whether the one translation per discourse hypothesis has the potential to be as useful to statistical machine translation as the one sense per discourse hypothesis to WSD. Current Statistical Machine Translation (SMT) systems translate one sentence at a time, ignoring any document level information. Implementing a one translation per document constraint might help provide consistency in translation for sentences drawn from the same document.

After briefly discussing related work, we will show that the one translation per discourse hypothesis holds on automatic word alignments of manually translated data. Despite ignoring any information beyond the sentential level, automatic SMT out-

*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 NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 19–27,
Boulder, Colorado, June 2009.

put also strongly exhibits the one translation per discourse property. In addition, we will show that having more than one translation per discourse in SMT output often reveals lexical choice errors, and that enforcing the constraint might help improve overall consistency across sentences and translation quality throughout documents.

2 Related Work

In the original one sense per discourse study, Gale *et al.* (1992b) considered a sample of 9 polysemous English words. A total of 5 judges were showed pairs of concordance lines for these words taken from Grolier's Encyclopedia and asked to identify whether they shared the same sense. Results strongly support the one sense per discourse hypothesis: 94% of polysemous words drawn from the same document have the same sense. The experiment was replicated with the same conclusion on the Brown corpus. Yarowsky (1995) successfully used this observation as an approximate annotation technique in an unsupervised WSD model.

A subsequent larger scale study of polysemy based on the WordNet sense inventory in the SEMCOR corpus does not support the hypothesis as strongly (Krovetz, 1998). Only 77% of ambiguous words have a single sense per discourse. Analysis revealed that the one sense per discourse hypothesis is only supported for homonymous senses and not for finer-grained sense distinction.

In machine translation, discourse level information has only been indirectly used by adaptation of translation or language models to specific genre or topics (e.g., Foster and Kuhn (2007); Koehn and Schroeder (2007)). While phrase-based SMT models incorporate the one sense per collocation hypothesis by attempting to translate phrases rather than single words (Koehn *et al.*, 2007), the one sense per discourse hypothesis has not been explicitly used in SMT modeling. Even the recent generation of SMT models that explicitly use WSD modeling to perform lexical choice rely on sentence context rather than wider document context and translate sentences in isolation (Carpuat and Wu, 2007; Chan *et al.*, 2007; Giménez and Màrquez, 2007; Stroppa *et al.*, 2007; Specia *et al.*, 2008). Other context-sensitive SMT approaches (Gimpel and Smith, 2008) and global lexical choice models (Bangalore *et al.*, 2007) also translate sentences independently.

3 One translation per discourse in reference translations

In this section we investigate whether the one sense per discourse hypothesis holds in translation. Does one sense per discourse mean one translation per discourse?

On the one hand, one translation per discourse might be too strict a constraint to allow for variations in lexicalization of a given sense. While a WSD task produces a set of predefined sense labels, a single sense might be correctly translated in many different ways in a full sentence translation.

On the other hand, if the author of the source language text is assumed to consistently use one sense per word per document, translators might also prefer consistent translations of the same source language word throughout a document. In addition, translated text tends to exhibit more regularities than original text, as shown by machine learning approches to discriminate between "translationese" and original texts (Baroni and Bernardini, 2006) although patterns of syntactic regularity seemed more informative than lexical choice for those experiments.

3.1 Manual translation data

We will test the one translation per discourse hypothesis on a corpus of French and English translations, using standard freely available MT data sets and software.

We use a corpus of 90 French-English news articles made available for the WMT evaluations[1]. All development data that contained article boundaries were used. The data is split into two sets of about 27k words each as described in Table 1. The articles cover topics ranging from international and local politics to sports and music. They are drawn from a wide variety of newspapers and magazines originally published in various European languages. As a result, even though only a single English reference translation is available, it was produced by several different interpreters. It would have been interesting to perform this analysis with multiple references, but this is unfortunately not possible with

[1] http://www.statmt.org/wmt09/translation-task.html

Test set	Language	Sentences	Tokens	Types	Singletons
no. 1	French	1070	27440	5958	3727
	English (ref)	1070	24544	5566	3342
	English (SMT)	1070	24758	5075	2932
no. 2	French	1080	27924	6150	3839
	English (ref)	1080	24825	5686	3414
	English (SMT)	1080	25128	5240	3080

Table 1: Data statistics for the bilingual corpus, including the French side, the manually translated English side (ref) and the automatic English translations (SMT)

the French-English data currently available.

Since golden word-alignments are not available, we automatically word align the corpus using standard SMT training techniques. Using IBM-4 alignment models learned on the large WMT training corpus (see Section 4.1 for more details), we align GIZA++(Och and Ney, 2003) to obtain the IBM-4 alignments in both translation directions, expand their intersection with additional links using the grow-diag-final-and heuristic (Koehn *et al.*, 2007). This creates a total of 51660 alignment links, and about 89% of French tokens are aligned to at least one English token. Note that all links involving stopwords are not considered for the rest of the study.

3.2 One translation per discourse holds

For every French lemma that occurs more than once in a document, we compute the number of English translations. In order to allow for morphological and syntactic variations, we compute those statistics using English lemmas obtained by running Treetagger (Schmid, 1994) with the standard French and English parameter settings[2]. A higher level of generalization is introduced by conducting the same analysis using stems, which are simply defined as 4-letter prefixes.

We have a total of 2316 different lemma types and 6603 lemma-document pairs. The scale of this empirical evaluation is much larger than in Gale *et al.* (1992a) where only 9 target words were considered and in Krovetz (1998) which used the entire SEMCOR corpus vocabulary.

The resulting distribution of number of English translations per French word-document pair is given in the first half of Table 2. Remarkably, more than 98% of the French lemmas are aligned to no more than 2 English translations and 80% of French lemmas have a single translation per document. While these numbers are not as high as the 94% agreement reported by Gale *et al.* (1992b) in their empirical study, they still strongly support the one translation per discourse hypothesis.

Generalizing from lemmas to stems yields a 4.3 point increase in the percentage of French lemmas with a single translation per document. Note that using stems might yield to false positives since different words can share the same prefix, however, since we only compare words that align to the same French word in a given document, the amount of noise introduced should be small. Manual inspection shows that this increase is often due to variations in the POS of the translation, more specifically variations between noun and verb forms which share the same 4-letter prefix as can be seen in the following examples:

verb vs. noun `conclude vs. conclusion, investigate vs. investigation, apply vs. application, inject vs. injection, establish vs. establishment, criticize vs. critism, recruit vs. recruitment, regulate vs. regulation`

3.3 Exceptions: one sense but more than one translation per discourse

We investigate what happens in the 15 to 20% of cases where a French word is not consistenly translated throughout a document. Do these translation differences reflect sense ambiguity in French, or are they close variations in English lexical choice? For

[2]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

	reference		SMT	
	lemmas	stems	**lemmas**	stems
1	**80.82%**	85.14%	**83.03%**	86.38%
2	17.88%	13.91%	15.43%	12.47%
3	01.12%	00.95%	01.25%	00.85%
4	00.18%	00.00%	00.17%	00.22%

Table 2: Distribution of number of English translation per document using the word-aligned reference translations and the automatic SMT translations

a given French word, how semantically similar are the various English translations?

We measure semantic similarity using the shortest path length in WordNet (Fellbaum, 1998) as implemented in the WordNet Similarity package (Pedersen *et al.*, 2004). The path length is defined as the number of WordNet nodes or synsets in a path between two words: words that belong to the same synset therefore have a shortest path length of 1, while words that are related via a common synonym, hypernym or hyponym have a shortest path length of 2. Note that this similarity metric is only defined for two WordNet vocabulary words of the same POS.

For 57% of the French lemmas with multiple translations, those translations can be linked by a WordNet path of no more than 4 nodes. In 19% of the cases, the translations belong to the same synset, another 19% are separated by a path of length 2 only. Given that sense distinctions in WordNet are very fine-grained, these numbers show that the translations have very similar meanings. In other words, while the one sense per translation hypothesis does not hold for those 57%, the one sense per discourse hypothesis still holds.

Examples of those synonymous translations are given below:

synonyms with SPL = 1 `adjust and adapt, earn and gain, movie and film, education and training, holiday and day`

synonyms with SPL = 2 `travel and circulate, scientist and researcher, investigation and inquiry, leave and abandon, witness and eyewitness`

synonyms with SPL = 3 `ratio and proportion, quiet and peace, plane and aircraft`

3.4 Exceptions: more than one sense per discourse

Among the words with a high WordNet path length or no existing path, we find translations that are not synonyms or semantically similar words, but related words sometimes with different POS. They fall within two categories.

The first category is that of fine-grained sense distinctions for which the one sense per discourse hypothesis has been showed to break for monolingual WordNet sense distinctions Krovetz (1998). However, for those closely related words, it would be possible to write correct English translations that use the same English form throughout a document.

Nationality translation `Tibet vs. Tibetan, French vs. France, Paris vs. Parisian, Europe vs. European, French vs. Frenchman`

Agent/entity `policeman vs. police, alderman vs. city`

The second category of not identical but related translations is explained by a limitation of our experiment set-up: we are looking at single-word translations while the translation of a longer multiword phrase should be considered as a whole. In the following example, the French word `émission` is aligned to both `emission` and `greenhouse` in the same document, because French does not repeat the long phrase `émission de gaz à effet de serre` throughout the document, while the more concise English translation `greenhouse gas emissions` is used throughout:

Fr `après la période de réduction des émissions [...] la Hongrie a pris l'engagement de réduire les émissions de gaz à effet de serre de 6 pour cent [...]`

En `[...] to cut greenhouse gas emissions after 2012 [...] Hungary agreed to cut its greenhouse gas emissions by 6 percent [...]`

Finally, there are a few rare instances where the different translations for a French word reflect a

sense distinction in French and could not be correctly translated in English with the same English word. These are cases where both the one sense per discourse hypothesis and the one translation per discourse break, and where it is not possible to paraphrase the English sentences to fullfill either constraints. In these instances, the French word is used in two different senses related by metonymy, but the metonymy relation does not translate into English and two non-synonym English words are used as a result. For instance, the French word `bureau` translates to both `office` and `desk` in the same document, while `retraite` translates both to `retirement` and `pension`.

We found a single instance where two homonym senses of the French word `coffre` are in the same sentence. This sentence seems to be a headline, which suggests that the author or translator deliberately used the ambiguous repetition to attract the attention of the reader.

Fr `un coffre dans le coffre`

En `a trunk in the boot`

4 One translation per discourse in SMT

We now turn to empirically testing the one translation per discourse hypothesis on automatically translated text.

While there is an implicit assumption that a well-written document produced by a human writer will not introduce unncessary ambiguities, most SMT systems translate one sentence at a time, without any model of discourse or document. This might suggest that the one translation per discourse hypothesis will not be as strongly supported as by manual translations.

However, this effect might be compensated by the tendency of automatically translated text to exhibit little variety in lexical choice as MT systems tend to produce very literal word for word translations. As can be seen in Table 1 the reference translations use a larger vocabulary than the automatic translations for the same text.

4.1 Automatically translated data

We build a standard SMT system and automatically translate the data set described in Section 3.1. We strictly follow the instructions for building a phrase-based SMT system that is close to the state-of-the-art in the WMT evaluations[3], using the large training sets of about 460M words from Europarl and news.

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 GIZA++ word alignments are intersected using the grow-diag-final-and heuristic. Translations of phrases of up to 7 words long are collected and scored with translation probilities 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 word alignment between French input sentences and English SMT output is easily obtained as a by-product of decoding. We have a total of 56003 alignment links, and 96% of French tokens are linked to a least one English translation.

4.2 One translation per discourse holds

We perform the same analysis as for the manual translations. The distribution of the number of translations for a given French word that occurs repeatedly in a document still strongly supports the one translation per document hypothesis (Table 2). In fact, SMT lexical choice seems to be more regular than in manual translations.

4.3 Exceptions: where SMT and reference disagree

Again, it is interesting to look at instances where the hypothesis is not verified. We will not focus on the exceptions that fall in the categories previously observed in Section 3. Instead, we take a closer look at cases where the reference consistently uses the same English translation, while SMT selects different translation candidates.

There are cases where the SMT system arbitrarily chooses different synonymous translation candidates for the same word in different sentences. This is not incorrect but will affect translation quality as measured by automatic metrics which compare

[3] http://www.statmt.org/wmt09/baseline.html

Test set	Decoding Input	METEOR	BLEU	NIST
	Moses	49.05	20.45	6.135
	+postprocess (transprob)	48.73	19.93	6.064
	+postprocess (bestmatch)	50.01	20.64	6.220
no. 1	+decode (transprob)	49.04	20.44	6.128
	+decode (bestmatch)	49.36	20.70	6.179
	Moses	49.60	21.10	6.211
	+postprocess (transprob)	49.20	20.43	6.128
	+postprocess (bestmatch)	50.56	21.19	6.291
no. 2	+decode (transprob)	49.58	21.02	6.201
	+decode (bestmatch)	50.60	21.21	6.243

Table 3: Enforcing one translation per discourse can help METEOR, BLEU and NIST scores when using the supervised sense disambiguation technique (bestmatch). Relying on the unsupervised context-independent SMT translation probabilities (transprob) does not help.

matches between SMT output and manually translated references. For instance, in a single document, the French `agents pathogènes` translates to both (1) `pathogens` and (2) `disease-causing agents` while the reference consistently translates to `pathogens`. Similarly, the French phrase `parmi les détenus` is inconsistently translated to `among detainees` and `among those arrested` in the same document.

Synonym translations `detainees vs. arrested, apartment vs. flat, good vs. beautiful, unit vs. cell`

However, the majority of differences in translation reflect lexical choice errors. For instance, the French adjective `biologique` is incorrectly disambiguated as `organic` in the phrase `fille biologique` which should be translated as `biological daughter`.

SMT lexical choice errors `conseiller: advisor vs. councillor, arrondissement: district vs. rounding-off, bal: ball vs. court, biologique: biological vs. organic, assurance: insurance vs. assurance, franchise: frankness vs. deductible`

While some of those translation distinctions can be explained by differences in topics, all of those French words occur in a large number of documents and cannot be disambiguated by topic alone. This suggests that local sentential context is not sufficient to correctly disambiguate translation candidates.

5 Detecting SMT errors

Based on the observations from the previous section, we further evaluate whether breaking the one translation per discourse hypothesis is indicative of a translation error. For this purpose, we attempt to correct the translations provided by the Moses SMT system by enforcing the one translation per discourse constraint and evaluate the impact on translation quality.

5.1 Enforcing one translation per discourse

In order to get a sense of the potential impact of the one translation per discourse constraint in SMT, we attempt to enforce it using two simple postprocessing techniques.

First, we select a set of French words which are not consistently translated to a single English words in a given document. We apply a document frequency-based filter to select content words for each document. This yields a set of 595 French target word types occurring in a total of 89 documents.

Second, we propose a single English translation for all the occurrences of the French target in a document. We used two different strategies: (1) the fully unsupervised strategy consists in selecting the translation with highest probability among those produced by the baseline SMT system, and

Moses	Young people under 25 years face various drawbacks when a contract with an *assurance* at an accessible price , as can be the low experience in the conduct and seniority of driving licences .
+postprocess	young people under 25 years against various drawbacks when a contract with an **insurance** at an accessible price , as can be the small experience in the conduct and seniority of driving licences .
Moses	drivers the most far-sighted can opt for insurance any risk with *frankness* , so that they get *blankets* insurance to any risk but at a price more accessible .
+postprocess	drivers the most far-sighted can opt for insurance any risk with *exemption* , so that they get *blankets* insurance to any risk but at a price more accessible .
Moses	" These *ill* are isolated , nurses puts gloves rubber and masks of protection and we have antibiotics adapted to treat them , " said Tibor Nyulasi .
+postprocess	" These **patient** are isolated , personnel puts gloves rubber and masks of protection and we have antibiotics appropriate to treat them , " say Tibor Nyulasi .
Moses	according to the Ministry of Defence , they also **served** to make known to the public the real **aims** of the presence of the army abroad .
+postprocess	according to the Ministry of Defence , they also *use* to make known to the public the real **purpose** of the presence of the army abroad .
Moses	the public authorities also **prepare** Christmas .
+postprocess	the public authorities also *puritan* Christmas .

Table 4: Examples of translation improvement (bold) and degradation (italics) by enforcing the one translation per discourse constraint through postprocessing

(2) the supervised strategy picks, among the baseline SMT translations, the one that matches the reference. Note that the supervised strategy does not predict perfect translations, but an approximation of the golden translations: in addition to noise in word alignments due to phrasal translations, the translations selected are lemmas that might not be in the correctly inflected form for use in the full sentence translation.

Third, we integrate the selected translation candidates by (1) postprocessing the baseline SMT output - the translations of the French target word are simply replaced by the recommended translation, and (2) encouraging the SMT system to choose the recommended translations by annotating SMT input using the xml input markup scheme - again, this approach is not optimal as it introduces additional translation candidates without probability scores and forces single word translation to compete with phrasal translation even if they are consistent.

5.2 Impact on translation quality

As reported in Table 3, small increases in METEOR (Banerjee and Lavie, 2005), BLEU (Papineni *et al.*, 2002) and NIST scores (Doddington, 2002) suggest that SMT output matches the references better after postprocessing or decoding with the suggested lemma translations. Examples of both improved and degraded lexical choice are given in Table 4.

Since we are modifying translations for a limited set of single-words only, only 10% to 30% of the test set sentences are translated differently. We manually inspected a random sample of 100 of those sentence pairs for two different systems: postprocess (bestmatch) and decode (bestmatch). For each sentence pair, we determined whether the "one sense per discourse" processing improved, degraded or made no difference in translation quality compared to the baseline Moses output. Among the sentence pairs where a real change in translation quality was observed, the postprocessing heuristic yielded improvements in 62.5% (decode) and 64.5% (postprocess) of sentences considered. For 41% (decode) and 57% (postprocess) of the sentences in the sam-

ple, changes only consisted of synonym substitution, morphological variations or local reorderings which did not impact translation quality.

Taken together, these results suggest that the "one sense per discourse" constraint should be useful to SMT and that it would be worthwile to integrate it directly into SMT modeling.

6 Conclusion

We investigated the one sense per discourse hypothesis (Gale *et al.*, 1992b) in the context of machine translation. Analysis of manual translations showed that the hypothesis still holds when using translations in parallel text as sense annotation, thus confirming that translational differences represent useful sense distinctions. Analysis of SMT output showed that despite ignoring document structure, the one translation per discourse hypothesis is strongly supported in part because of the low variability in SMT lexical choice. More interestingly, cases where the hypothesis does not hold can reveal lexical choice errors in an unsupervised fashion. A preliminary study showed that enforcing the one translation per discourse constraint in SMT can potentially improve translation quality, and that SMT systems might benefit from translating sentences within their entire document context.

In future work, we will (1) evaluate whether one translation per discourse holds for other language pairs such as Arabic-English and Chinese-English, which are not as closely related as French-English and for which multiple reference corpora are available, and (2) directly implement the one translation per discourse constraint within SMT.

References

Eneko Agirre and German Rigau. Word sense disambiguation using conceptual density. In *Proceedings of COLING'96*, pages 16–22, Copenhagen, Denmark, 1996.

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.

Srinivas Bangalore, Patrick Haffner, and Stephan Kanthak. Statistical machine translation through global lexical selection and sentence reconstruction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 152–159, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Marco Baroni and Silvia Bernardini. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, 2006.

Marine Carpuat and Dekai Wu. 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 2007)*, pages 61–72, Prague, June 2007.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, Prague, June 2007.

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

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology conference (HLT-2002)*, San Diego, CA, 2002.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

George Foster and Roland Kuhn. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June 2007.

William A. Gale, Kenneth W. Church, and David Yarowsky. A method for disambiguating word

senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1992.

William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, Harriman, NY, February 1992.

Jesús Giménez and Lluís Màrquez. Context-aware discriminative phrase selection for statistical machine translation. In *Workshop on Statistical Machine Translation*, Prague, June 2007.

Kevin Gimpel and Noah Smith. Rich source-side context for statistical machine translation. In *Workshop on Statistical Machine Translation*, Columbus, Ohio, June 2008.

Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Workshop on Statistical Machine Translation*, Prague, June 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.

Robert Krovetz. More than one sense per discourse. In *NEC Princeton NJ Labs., Research Memorandum*, 1998.

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.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52, 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.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet::Similarity - Measuring the relatedness of concepts. In *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04)*, pages 38–41, Boston, MA, May 2004.

Philip Resnik and David Yarowsky. Distinguising systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133, 1999.

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.

Lucia Specia, Baskaran Sankaran, and Maria das Graças Volpe Nunes. n-best reranking for the efficient integration of word sense disambiguation and statistical machine translation. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing'08)*, Haifa, Israel, February 2008.

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

Nicolas Stroppa, Antal van den Bosch, and Andy Way. Exploiting source similarity for smt using context-informed features. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skovde, Sweden, September 2007.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA, 1995.