A Semantic Evaluation of Machine Translation Lexical Choice

Marine Carpuat
National Research Council Canada
1200 Montreal Rd,
Ottawa, ON K1A 0R6
Marine.Carpuat@nrc.gc.ca

Abstract

While automatic metrics of translation quality are invaluable for machine translation research, deeper understanding of translation errors require more focused evaluations designed to target specific aspects of translation quality. We show that Word Sense Disambiguation (WSD) can be used to evaluate the quality of machine translation lexical choice, by applying a standard phrase-based SMT system on the SemEval2010 Cross-Lingual WSD task. This case study reveals that the SMT system does not perform as well as a WSD system trained on the exact same parallel data, and that local context models based on source phrases and target n-grams are much weaker representations of context than the simple templates used by the WSD system.

1 Introduction

Much research has focused on automatically evaluating the quality of Machine Translation (MT) by comparing automatic translations to human translations on samples of a few thousand sentences. Many metrics (Papineni et al., 2002; Banerjee and Lavie, 2005; Giménez and Márquez, 2007; Lo and Wu, 2011, for instance) have been proposed to estimate the adequacy and fluency of machine translation and evaluated based on their correlatation with human judgements of translation quality (Callison-Burch et al., 2010). While these metrics have proven invaluable in driving progress in MT research, finer-grained evaluations of translation quality are necessary to provide a more focused analysis of translation errors. When developing complex MT systems, comparing BLEU or TER scores is not sufficient to understand what improved or what went wrong. Error analysis can of course be done manually (Vilar et al., 2006), but it is often too slow and expensive to be performed as often as needed during system development.

Several metrics have been recently proposed to evaluate specific aspects of translation quality such as word order (Birch et al., 2010; Chen et al., 2012). While word order is indirectly taken into account by BLEU, TER or METEOR scores, dedicated metrics provide a direct evaluation that lets us understand whether a given system's reordering performance improved during system development. Word order metrics provide a *complementary* tool for targeting evaluation and analysis to a specific aspect of machine translation quality.

There has not been as much work on evaluating the lexical choice performance of MT: does a MT system preserve the meaning of words in translation? This is of course measured indirectly by commonly used global metrics, but a more focused evaluation can help us gain a better understanding of the behavior of MT systems.

In this paper, we show that MT lexical choice can be framed and evaluated as a standard Word Sense Disambiguation (WSD) task. We leverage existing WSD shared tasks in order to evaluate whether word meaning is preserved in translation. Let us emphasize that, just like reordering metrics, our WSD evaluation is meant to *complement* global metrics of translation quality. In previous work, intrinsic evaluations of lexical choice have been performed using either semi-automatically constructed data sets

Proceedings of the 7th Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 1–10,
Atlanta, Georgia, 13 June 2013.

based on MT reference translations (Giménez and Màrquez, 2008; Carpuat and Wu, 2008), or manually constructed word sense disambiguation test beds that do not exactly match MT lexical choice (Carpuat and Wu, 2005). We will show how existing Cross-Lingual Word Sense Disambiguation tasks (Lefever and Hoste, 2010; Lefever and Hoste, 2013) can be directly seen as machine translation lexical choice (Section 2): their sense inventory is based on translations in a second language rather than arbitrary sense representations used in other WSD tasks (Carpuat and Wu, 2005); unlike in MT evaluation settings, human annotators can more easily provide a complete representation of all correct meanings of a word. Second, we show how using this task for evaluating the lexical choice performance of several phrase-based SMT systems (PB-SMT) gives some insights into their strengths and weaknesses (Section 5).

2 Selecting a Word Sense Disambiguation Task to Evaluate MT Lexical Choice

Word Sense Disambiguation consists in determining the correct sense of a word in context. This challenging problem has been studied from a rich variety of persectives in Natural Language Processing (see Agirre and Edmonds (2006) for an overview.) The Senseval and SemEval series of evaluations (Edmonds and Cotton, 2001; Mihalcea and Edmonds, 2004; Agirre et al., 2007) have driven the standardization of methodology for evaluating WSD systems. Many shared tasks were organized over the years, providing evaluation settings that vary along several dimensions, including:

- target vocabulary: in *all word* tasks, systems are expected to tag all content words in running text (Palmer et al., 2001), while in *lexical sample* tasks, the evaluation considers a smaller predefined set of target words (Mihalcea et al., 2004; Lefever and Hoste, 2010).
- language: English is by far the most studied language, but the disambiguation of words in other languages such as Chinese (Jin et al., 2007) has been considered.
- sense inventory: many tasks use WordNet senses (Fellbaum, 1998), but other sense representations have been used, including alternate semantic databases such as HowNet (Dong, 1998), or lexicalizations in one or more languages (Chklovski et al., 2004).

The Cross-Lingual Word Sense Disambiguation (CLWSD) task introduced at a recent edition of SemEval (Lefever and Hoste, 2010) is an English lexical sample task that uses translations in other European languages as a sense inventory. As a result, it is particularly well suited to evaluating machine translation lexical choice.

2.1 Translations as Word Sense Representations

The CLWSD task is essentially the same task as MT lexical choice: given English target words in context, systems are asked to predict translations in other European languages. The gold standard consists of translations proposed by several bilingual humans, as can be seen in Table 1. MT system predictions can be compared to human annotations directly, without introducing additional sources of ambiguity and mismatches due to representation differences. This contrasts with our previous work on evaluating MT on a WSD task (Carpuat and Wu, 2005), which used text annotated with abstract sense categories from the HowNet knowledge base (Dong, 1998). In HowNet, each word is defined using a concept, constructed as a combination of basic units of meaning, called sememes. Words that share the same concept can be viewed as synonyms. Evaluating MT using a gold standard of HowNet categories requires to map translations from the MT output to the HowNet representation. Some categories are annotated with English translations, but additional effort is required in order to cover all translation candidates produced by the MT system.

2.2 Controlled Learning Conditions

Another advantage of the CLWSD task is that it provides controlled learning conditions (even though it is an unsupervised task with no annotated training data.) The gold labels for CLWSD are learned from parallel corpora. As a result MT lexical choice models can be estimated on the exact same data. Translations for English words in the lexical sample are extracted from a semi-automatic word alignment of

Target word	**ring**
English context	The twelve stars of the European flag are depicted on the outer **ring**.
Gold translations	anillo (3);círculo (2);corona (2);aro (1);
English context	The terrors which Mr Cash expresses about our future in the community have a familiar **ring** about them.
Gold translations	sonar (3);tinte (3);connotación(2);tono (1);
English context	The American containment **ring** around the Soviet bloc had been seriously breached only by the Soviet acquisition of military facilities in Cuba.
Gold translations	cerco (2);círculo (2);cordón (2);barrera (1);blindaje (1);limitación (1);

Table 1: Example of annotated CLWSD instances from the SemEval 2010 test set. For each gold Spanish translation, we are given the number of annotators who proposed it (out of 3 annotators.)

sentences from the Europarl parallel corpus (Koehn, 2005). These translations are then manually clustered into senses. When constructing the gold annotation, human annotators are given occurrences of target words in context. For each occurrence, they select a sense cluster and provide all translations from this cluster that are correct in this specific context. Since three annotators contribute, each test occurrence is therefore tagged with a set of translations in another language, along with a frequency which represents the number of annotators who selected it. A more detailed description of the annotation process can be found in (Lefever and Hoste, 2010).

Again, this contrasts with our previous work on evaluating MT on a HowNet-based Chinese WSD task, where Chinese sentences were manually annotated with HowNet senses which were completely unrelated to the parallel corpus used for training the SMT system. Using CLWSD as an evaluation of MT lexical choice solves this issue and provides controlled learning conditions.

2.3 CLWSD evaluates the semantic adequacy of MT lexical choice

A key challenge in MT evaluation lies in deciding whether the meaning of the translation is correct when it does not exactly match the reference translation. METEOR uses WordNet synonyms and learned paraphrases tables (Denkowski and Lavie, 2010). MEANT uses vector-space based lexical similarity scores (Lo et al., 2012). While these methods lead to higher correlations with human judgements on average, they are not ideal for a fine-grained evaluation of lexical choice: similarity scores are defined independently of context and might give credit to incorrect translations (Carpuat et al., 2012). In contrast, CLWSD solves this difficult problem by providing all correct translation candidates in context according to several human annotators. These multiple translations provide a more complete representation of the correct meaning of each occurrence of a word in context.

The CLWSD annotation procedure is designed to easily let human annotators provide many correct translation alternatives for a word. Producing many correct annotations for a complete sentence is a much more expensive undertaking: crowdsourcing can help alleviate the cost of obtaining a small number of reference translation (Zbib et al., 2012), but acquiring a complete representation of all possible translations of a source sentence is a much more complex task (Dreyer and Marcu, 2012). Machine translation evaluations typically use between one and four reference translations, which provide a very incomplete representation of the correct semantics of the input sentence in the output language. CLWSD provides a more complete representation through the multiple gold translations available.

2.4 Limitations

The main drawback of using CLWSD to evaluate lexical choice is that CLWSD is a lexical sample task, which only evaluates disambiguation of 20 English nouns. This arbitrary sample of words does not let us target words or phrases that might be specifically interesting for MT.

In addition, the data available through the shared task does not let us evaluate complete translations of the CLWSD test sentences, since full references translations are not available. Instead of using

a WSD dataset for MT purposes, we could take the converse approach andautomatically construct a WSD test set based on MT evaluation corpora (Vickrey et al., 2005; Giménez and Màrquez, 2008; Carpuat and Wu, 2008; Carpuat et al., 2012). However, this approach suffers from noisy automatic alignments between source and reference, as well as from a limited representation of the correct meaning of words in context due to the limited number of reference translations.

Other SemEval tasks such as the Cross-Lingual Lexical Substitution Task (Mihalcea et al., 2010) would also provide an appropriate test bed. We focused on the CLWSD task, since it uses senses drawn from the Europarl parallel corpus, and therefore offers more constrained settings for comparison between systems. The lexical substitution task targets verbs and adjectives in addition to nouns, and would therefore be an interesting test case to consider in future work.

2.5 Official and MT-centric Evaluation Metrics

In order to make comparison with other systems possible, we follow the standard evaluation framework defined for the task and score the output of all our systems using four different metrics, computed using the scoring tool made available by the organizers.

The difference between system predictions and gold standard annotations are quantified using *precision* and *recall* scores[1] , defined as follows. Given a set T of test items and a set H of annotators, H_i is the set of translation proposed by all annotators h for instance $i \in T$. Each translation type res in H_i has an associated frequency $freq_{res}$, which represents the number of human annotators which selected res as one of their top 3 translations. Given a set of system answers A of items $i \in T$ such that the system provides at least one answer, $a_i : i \in A$ is is the set of answers from the system for instance i. For each i, the scorer computes the intersection of the system answers a_i and the gold standard H_i.

Systems propose as many answers as deemed necessary, but the scores are divided by the number of guesses in order not to favor systems that output many answers per instance.

$$\text{Precision} = \frac{1}{|A|} \sum_{a_i : i \in A} \frac{\sum_{res \in a_i} freq_{res}}{|a_i||H_i|}$$

$$\text{Recall} = \frac{1}{|T|} \sum_{a_i : i \in T} \frac{\sum_{res \in a_i} freq_{res}}{|a_i||H_i|}$$

We also report *Mode Precision* and *Mode Recall* scores: instead of comparing system answers to the full set of gold standard translations H_i for an instance $i \in T$, the Mode Precision and Recall scores only use a single gold translation, which is the translation chosen most frequently by the human annotators.

In addition, we compute the *1-gram precision* component of the BLEU score (Papineni et al., 2002), denoted as BLEU1 in the result tables[2]. In contrast with the official CLWSD evaluation scores described above, BLEU1 gives equal weight to all translation candidates, which can be seen as multiple references.

3 PBSMT system

We use a typical phrase-based SMT system trained for English-to-Spanish translation. Its application to the CLWSD task affects the selection of training data and its preprocessing, but the SMT model design and learning strategies are exactly the same as for conventional translation tasks.

3.1 Model

We use the NRC's PORTAGE phrase-based SMT system, which implements a standard phrasal beam-search decoder with cube pruning. Translation hypotheses are scored according to the following features:

- 4 phrase-table scores: phrasal translation probabilites with Kneser-Ney smoothing and Zens-Ney lexical smoothing in both translation directions (Chen et al., 2011)
- 6 hierarchical lexicalized reordering scores, which represent the orientation of the current phrase with respect to the previous block that could have been translated as a single phrase (Galley and Manning, 2008)

[1] In this paper, we focus on evaluating translation systems whose task is to produce a single complete translation for a given sentence. As a result, we only focus on the 1-best MT output and do not report the relaxed out-of-five evaluation setting also considered in the official SemEval task.

[2] even though it does not include the length penalty used in the BLEU score.

- a word penalty, which scores the length of the output sentence
- a word-displacement distortion penalty
- a Kneser-Ney smoothed 5-gram Spanish language model

Weights for these features are learned using a batch version of the MIRA algorithm(Cherry and Foster, 2012). Phrase pairs are extracted from IBM4 alignments obtained with GIZA++(Och and Ney, 2003). We learn phrase translation candidates for phrases of length 1 to 7.

Converting the PBSMT output for CLWSD requires a final straightforward mapping step. We use the phrasal alignment between SMT input and output to isolate the translation candidates for the CLWSD target word. When it maps to a multiword phrase in the target language, we use the word within the phrase that has the highest translation IBM1 translation probability given the CLWSD target word of interest. Note that there is no need to perform any manual mapping between SMT output and sense inventories as in (Carpuat and Wu, 2005).

3.2 Data

The core training corpus is the exact same set of sentences from Europarl that were used to learn the sense inventory, in order to ensure that PBSMT knows the same translations as the human annotators who built the gold standard. There are about 900k sentence pairs, since only 1-to1 alignments that exist in all the languages considered in CLWSD were used (Lefever and Hoste, 2010).

We exploit additional corpora from the WMT2012 translation task, using the full Europarl corpus to train language models, and for one experiment the news-commentary parallel corpus (see Section 9.)

These parallel corpora are used to learn the translation, reordering and language models. The log-linear feature weights are learned on a development set of 3000 sentences sampled from the WMT2012 development test sets. They are selected based on their distance to the CLWSD trial and test sentences (Moore and Lewis, 2010).

We tokenize and lemmatize all English and Spanish text using the FreeLing tools (Padró and Stanilovsky, 2012). We use lemma representations to perform translation, since the CLWSD targets and translations are lemmatized.

4 WSD system

4.1 Model

We also train a dedicated WSD system for this task in order to perform a controlled comparison with the SMT system. Many WSD systems have been evaluated on the SemEval test bed used here, however, they differ in terms of resources used, training data and preprocessing pipelines. In order to control for these parameters, we build a WSD system trained on the exact same training corpus, preprocessing and word alignment as the SMT system described above.

We cast WSD as a generic ranking problem with linear models. Given a word in context x, translation candidates t are ranked according to the following model: $f(x,t) = \sum_i \lambda_i \phi_i(x,t)$, where $\phi_i(x,t)$ represent binary features that fire when specific clues are observed in a context x.

Context clues are based on standard feature templates in many supervised WSD approaches (Florian et al., 2002; van Gompel, 2010; Lefever et al., 2011):

- words in a window of 2 words around the disambiguation target.
- part-of-speech tags in a window of 2 words around the disambiguation target
- bag-of-word context: all nouns, verbs and adjectives in the context x

At training time, each example (x,t) is assigned a cost based on the translation observed in parallel corpora: $f(x,t) = 0$ if $t = t_{aligned}$, $f(x,t) = 1$ otherwise . Feature weights λ_i can be learned in many ways. We optimize logistic loss using stochastic gradient descent[3].

4.2 Data

The training instances for the supervised WSD system are built automatically by (1) extracting all occurrences of English target words in context, and (2) annotating them with their aligned Spanish lemma.

[3] we use the optimizer from http://hunch.net/~ vw v7.1.2

System	Prec.	Rec.	Mode Prec.	Mode Rec.	BLEU1
WSD	25.96	25.58	55.02	54.13	76.06
PBSMT	23.72	23.69	45.49	45.37	62.72
MFStest	21.35	21.35	44.50	44.50	65.50
MFStrain	19.14	19.14	42.00	42.00	59.70

Table 2: Main CLWSD results: PBSMT yields competitive results, but WSD outperforms PBSMT

We obtain a total of 33139 training instances for all targets (an average of 1656 per target, with a minimum of 30 and a maximum of 5414). Note that this process does not require any manual annotation.

5 WSD systems can outperform PBSMT

Table 2 summarizes the main results. PBSMT outperforms the most frequent sense baseline by a wide margin, and interestingly also yields better results than many of the dedicated WSD systems that participated in the SemEval task. However, PBSMT performance does not match that of the most frequent sense oracle (which uses sense frequencies observed in the test set rather than training set). The WSD system trained on the same word-aligned parallel corpus as the PBSMT system achieves the best performance. It also obtains better results than all but the top system in the official results (Lefever and Hoste, 2010).

The results in Table 2 are quite different from those reported by Carpuat and Wu (2005) on a Chinese WSD task. The Chinese-English PBSMT system performed much worse than any of the dedicated WSD systems on that task. While our WSD system outperforms PBSMT on the CLWSD task too, the difference is not as large, and the PBSMT system is competitive when compared to the full set of systems that were evaluated on this task. This confirms that the CLWSD task represents a more fair benchmark for comparing PBSMT with WSD systems.

6 Impact of PBSMT Context Models

What is the impact of PBSMT context models on lexical choice accuracy? Table 3 provides an overview of experiments where we vary the context size available to the PBSMT system. The main PBSMT system in the top row uses the default settings presented in Section 3.

System	Prec.	Rec.	Mode Prec.	Mode Rec.	BLEU1
PBSMT	23.72	23.69	45.49	45.37	62.72
max source phrase length l					
$l = 1$	24.44	24.36	44.50	44.38	65.43
$l = 3$	24.27	24.22	46.52	46.41	64.33
n-gram LM order					
$n = 3$	23.60	23.55	44.58	44.47	61.62
$n = 7$	23.58	23.53	46.06	45.94	62.22
$n = 2$	23.40	23.35	44.75	44.63	63.02
$n = 1$	22.92	22.87	43.00	42.89	58.62
+bilingual LM					
4-gram	23.89	23.84	45.49	45.37	62.62

Table 3: Impact of source and target context models on PBSMT performance

In the first set of experiments, we evaluate the impact of the source side context on CLWSD performance. Phrasal translations represent the core of PBSMT systems: they capture collocational context in the source language, and they are therefore are less ambiguous than single words (Koehn and Knight, 2003; Koehn et al., 2003). The default PBSMT learns translations for sources phrases of length ranging from 1 to 7 words.

Limiting the PBSMT system to translate shorter phrases (Rows $l = 1$ and $l = 3$ in Table 3) surprisingly improves CLWSD performance, even though it degrades BLEU score on WMT test sets. The source context captured by longer phrases therefore does not provide the right disambiguating information in this context.

In the second set of experiments, we evaluate the impact of the context size in the target language, by varying the size of the n-gram language model used. The default PBSMT system used a 5-gram language model. Reducing the n-gram order to 3, 2, 1 and increasing it to 7 both degrade performance. Shorter n-grams do not provide enough disambiguating context, while longer n-grams are more sparse and perhaps do not generalize well outside of the training corpus.

Finally, we report a last experiment which uses a bilingual language model to enrich the context representation in PBSMT (Niehues et al., 2011). This language model is estimated on word pairs formed

System	Prec.	Rec.	Mode Prec.	Mode Rec.	BLEU1
+ hier	23.72	23.69	45.49	45.37	62.72
+ lex	23.69	23.64	46.66	46.54	62.22
dist	23.42	23.37	45.43	45.30	62.22

Table 4: Impact of reordering models: lexicalized reodering does not hurt lexical choice only when hierarchical models are used

by target words augmented with their aligned source words. We use a 4-gram model, trained using Good-Turing discounting. This only results in small improvements (< 0.1) over the standard PBSMT system, and remains far below the performance of the dedicated WSD system.

These results show that source phrases are weak representations of context for the purpose of lexical choice. Target $n-gram$ context is more useful than source phrasal context, which can surprisingly harm lexical choice accuracy.

7 Impact of PBSMT Reordering Models

While the phrase-table is the core of PBSMT system, the reordering model used in our system is heavily lexicalized. In this section, we evaluate its impact on CLWSD performance. The standard PBSMT system uses a hierarchical lexicalized reordering model (Galley and Manning, 2008) in addition to the distance-based distortion limit. Unlike lexicalized reordering(Koehn et al., 2007), which models the orientation of a phrase with respect to the previous phrase, hierarchical reordering models define the orientation of a phrase with respect to the previous block that could have been translated as a single phrase.

In Table 4, we show that lexicalized reordering model benefit CLWSD performance, and that the hierarchical model performs slightly better than the non-hierarchical overall.

8 Impact of phrase translation selection

In this section, we consider the impact of various methods for selecting phrase translations on the lexical choice performance of PBSMT.

First, we investigate the impact of limiting the number of translation candidates considered for each source phrase in the phrase-table. The main PBSMT system uses $t = 50$ translation candidates per source phrase. Limiting that number to 20 and increasing it to 100 both have a very small impact on CLWSD.

System	Prec.	Rec.	Mode Prec.	Mode Rec.	BLEU1
PBSMT	23.72	23.69	45.49	45.37	62.72
Number t of translations per phrase					
$t = 20$	23.68	23.63	45.66	45.54	62.32
$t = 100$	23.65	23.60	45.65	45.53	62.52
Other phrase-table pruning methods					
stat sig	23.71	23.66	45.19	45.07	62.62

Table 5: Impact of translation candidate selection on PBSMT performance

Second, we prune the phrase-table using a statistical significance test to measure (Johnson et al., 2007). This pruning strategy aims to drastically decrease the size of the phrase-table without degrading translation performance by removing noisy phrase pairs.

9 Impact of training corpus

Since increasing the amount of training data is a reliable way to improve translation performance, we evaluate the impact of training the PBSMT system on more than the Europarl data used for controlled comparison with WSD. We increase the parallel training corpus with the WMT-12 News Commentary parallel data [4]. This yields an additional training set of roughly160k sentence pairs. We build linear mixture models to combine translation, reordering and language models learned on Europarl and News Commentary corpora (Foster and Kuhn, 2007). As can be seen in Table 6, this approach improves all CLWSD scores except for 1-gram precision. The decrease in 1-gram precision indicates that the addition of the news corpus introduces new translation candidates that differ from those used in the gold inventory. Interestingly, the additional data is not sufficient to match the performance of the WSD system learned on Europarl only (see Table 2). While additional data should be used when available, richer context features are valuable to make the most of existing data.

[4] http://www.statmt.org/wmt12/translation-task.html

System	Prec.	Rec.	Mode Prec.	Mode Rec.	BLEU1
Europarl	23.72	23.69	45.49	45.37	62.72
+ News	24.34	24.28	47.49	47.37	61.22

Table 6: Impact of training corpus on PBSMT performance: adding news parallel sentences helps Precision and Recall, but does not match WSD on the Europarl only.

10 Conclusion

We use a SemEval Cross-Lingual WSD task to evaluate the lexical choice performance of a typical phrase-based SMT system. Unlike conventional WSD task that rely on abstract sense inventories rather than translations, cross-lingual WSD provides a fair setting for comparing SMT with dedicated WSD systems. Unlike conventional evaluations of machine translation quality, the cross-lingual WSD task lets us isolate a specific aspect of translation quality and show how it is affected by different components of the phrase-based SMT system.

Unlike in previous evaluations on conventional WSD tasks (Carpuat and Wu, 2005), phrase-based SMT performance is on par with many dedicated WSD systems. However, the phrase-based SMT system does not perform as well as a WSD system trained on the exact same parallel data. Analysis shows that while many SMT components can potentially have an impact on SMT lexical choice, CLWSD accuracy is most affected by the length of source phrases and order of target n-gram language models. Using shorter source phrases actually improves lexical choice accuracy. The official results for the CLWSD task at SemEval 2013 evaluation provide further insights (Lefever and Hoste, 2013): our PBSMT system can achieve top precision as measured using the top prediction as in this paper, but does not perform as well as other submitted systems when taking into account the top 5 predictions (Carpuat, 2013). This suggests that local context models based on source phrases and target n-grams are much weaker representations of context than the simple templates used by WSD systems, and that even strong PBSMT systems can benefit from context models developed for WSD.

New learning algorithms (Chiang et al., 2009; Cherry and Foster, 2012, for instance) finally make it possible for PBSMT to reliably learn from many more features than the typical system used here. Evaluations such as the CLWSD task will provide useful tools for analyzing the impact of these features on lexical choice and inform feature design in increasingly large and complex systems.

References

E. Agirre and P.G. Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Text, Speech, and Language Technology Series. Springer Science+Business Media B.V.

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of SemEval-2007: 4th International Workshop on Semantic Evaluation*, Prague, July.

Satanjeev Banerjee and Alon Lavie. 2005. 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.

Alexandra Birch, Mile Osborne, and Phil Blunsom. 2010. Metrics for MT evaluation: Evaluating reordering. *Machine Translation*, 24:15–26.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 17–53.

Marine Carpuat and Dekai Wu. 2005. Evaluating the Word Sense Disambiguation Performance of Statistical Machine Translation. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP)*, pages 122–127, Jeju Island, Republic of Korea.

Marine Carpuat and Dekai Wu. 2008. Evaluation of Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation. In *Proceedings of the sixth conference on Language Resouces and Evaluation (LREC 2008)*, Marrakech, May.

Marine Carpuat, Hal Daumé III, Alexander Fraser, Chris Quirk, Fabienne Braune, Ann Clifton, Ann Irvine, Jagadeesh Jagarlamudi, John Morgan, Majid Razmara, Aleš Tamchyna, Katharine Henry, and Rachel

Rudinger. 2012. Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins Summer Workshop Final Report*.

Marine Carpuat. 2013. Nrc: A machine translation approach to cross-lingual word sense disambiguation (SemEval-2013 Task 10). In *Proceedings of SemEval*.

Boxing Chen, Roland Kuhn, George Foster, and Howard Johnson. 2011. Unpacking and transforming feature functions: New ways to smooth phrase tables. In *Proceedings of Machine Translation Summit*.

Boxing Chen, Roland Kuhn, and Samuel Larkin. 2012. Port: a precision-order-recall mt evaluation metric for tuning. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 930–939.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June.

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *NAACL-HLT 2009: Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 218–226, Boulder, Colorado.

Timothy Chklovski, Rada Mihalcea, Ted Pedersen, and Amruta Purandare. 2004. 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.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR paraphrase tables: improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 339–342.

Zhendong Dong. 1998. Knowledge description: what, how and who? In *Proceedings of International Symposium on Electronic Dictionary*, Tokyo, Japan.

Markus Dreyer and Daniel Marcu. 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada, June.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: Overview. In *Proceedings of Senseval-2: Second International Workshop on Evaluating Word Snese Disambiguation Systems*, pages 1–5, Toulouse, France.

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

Radu Florian, Silviu Cucerzan, Charles Schafer, and David Yarowsky. 2002. Combining classifiers for word sense disambiguation. *Natural Language Engineering*, 8(4):327–241.

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

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856.

Jesús Giménez and Lluís Márquez. 2007. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Prague

Jesús Giménez and Lluís Màrquez. 2008. Discriminative Phrase Selection for Statistical Machine Translation. *Learning Machine Translation. NIPS Workshop Series*.

Peng Jin, Yunfang Wu, and Shiwen Yu. 2007. Semeval-2007 task 05: Multilingual chinese-english lexical sample. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 19–23, Prague, Czech Republic, June.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975.

Philipp Koehn and Kevin Knight. 2003. Feature-Rich Statistical Translation of Noun Phrases. In *Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of HLT/NAACL-2003*, Edmonton, Canada, May.

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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, Phuket, Thailand, September.

Els Lefever and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden, July.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, USA, May.

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–322, Portland, Oregon, USA, June.

Chi-kiu Lo and Dekai Wu. 2011. Meant: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 220–229.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252.

Rada Mihalcea and Phiip Edmonds, editors. 2004. *Proceedings of Senseval-3: Third international Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain.

Rada Mihalcea, Timothy Chklovski, and Adam Killgariff. 2004. 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.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden, July.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA.

Jan Niehues, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider context by using bilingual language models in machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 198–206, Stroudsburg, PA, USA.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hao Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of Senseval-2, Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24, Toulouse, France, July. SIGLEX, Association for Computational Linguistic.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July.

Maarten van Gompel. 2010. Uvt-wsd1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 238–241, Uppsala, Sweden, July.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-Sense Disambiguation for Machine Translation. In *Joint Human Language Technology conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702, Genoa, Italy, May.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 49–59.