Contextual Modeling for Meeting Translation Using Unsupervised Word Sense Disambiguation

Yang Mei
Department of Electrical Engineering
University of Washington
yangmei@u.washington.edu

Katrin Kirchhoff
Department of Electrical Engineering
University of Washington
katrin@ee.washington.edu

Abstract

In this paper we investigate the challenges of applying statistical machine translation to meeting conversations, with a particular view towards analyzing the importance of modeling contextual factors such as the larger discourse context and topic/domain information on translation performance. We describe the collection of a small corpus of parallel meeting data, the development of a statistical machine translation system in the absence of genre-matched training data, and we present a quantitative analysis of translation errors resulting from the lack of contextual modeling inherent in standard statistical machine translation systems. Finally, we demonstrate how the largest source of translation errors (lack of topic/domain knowledge) can be addressed by applying document-level, *unsupervised* word sense disambiguation, resulting in performance improvements over the baseline system.

1 Introduction

Although statistical machine translation (SMT) has made great progress over the last decade, most SMT research has focused on the translation of structured input data, such as newswire text or parliamentary proceedings. Spoken language translation has mostly concentrated on two-person dialogues, such as travel expressions or patient-provider interactions in the medical domain. Recently, more advanced spoken-language data has been addressed, such as speeches (Stüker et al., 2007), lectures (Waibel and Fügen, 2008), and broadcast conversations (Zheng et al., 2008). Problems for machine translation in these genres include the nature of spontaneous speech input (e.g. disfluencies, incomplete sentences, etc.) and the lack of high-quality training data. Data that match the desired type of spoken-language interaction in topic, domain, and, most importantly, in style, can only be obtained by transcribing and translating conversations, which is a costly and time-consuming process. Finally, many spoken-language interactions, especially those involving more than two speakers, rely heavily on the participants' shared contextual knowledge about the domain and topic of the discourse, relationships between speakers, objects in the real-world environment, past interactions, etc. These are typically not modelled in standard SMT systems.

The problem of speech disfluencies has been addressed by disfluency removal techniques that are applied prior to translation (Rao et al., 2007; Wang et al., 2010). Training data sparsity has been addressed by adding data from out-of-domain resources (e.g. (Matusov et al., 2004; Hildebrandt et al., 2005; Wu et al., 2008)), exploiting comparable rather than parallel corpora (Munteanu and Marcu, 2005), or paraphrasing techniques (Callison-Burch et al., 2006). The lack of contextual modeling, by contrast, has so far not been investigated in depth, although it is a generally recognized problem in machine translation. Early attempts at modeling contextual information in machine translation include (Mima et al., 1998), where information about the role, rank and gender of speakers and listeners was utilized in a transfer-based spoken-language translation system for travel dialogs. In (Kumar et al., 2008)

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1227–1235,
Beijing, August 2010

statistically predicted dialog acts were used in a phrase-based SMT system for three different dialog tasks and were shown to improve performance. Recently, contextual source-language features have been incorporated into translation models to predict translation phrases for traveling domain tasks (Stroppa et al., 2007; Haque et al., 2009). However, we are not aware of any work addressing contextual modeling for statistical translation of spoken meeting-style interactions, not least due to the lack of a relevant corpus.

The first goal of this study is to provide a quantitative analysis of the impact of the lack of contextual modeling on translation performance. To this end we have collected a small corpus of parallel multi-party meeting data. A baseline SMT system was trained for this corpus from freely available data resources, and contextual translation errors were manually analyzed with respect to the type of knowledge sources required to resolve them. Our analysis shows that the largest error category consists of word sense disambiguation errors resulting from a lack of topic/domain modeling. In the second part of this study we therefore present a statistical way of incorporating such knowledge by using a graph-based unsupervised word sense disambiguation algorithm at a *global* (i.e. document) level. Our evaluation on real-world meeting data shows that this technique improves the translation performance slightly but consistently with respect to position-independent word error rate (PER).

2 Data

2.1 Parallel Conversational Data

For our investigations we used a subset of the AMI corpus (McCowan, 2005), which is a collection of multi-party meetings consisting of approximately 100 hours of multimodal data (audio and video recordings, slide images, data captured from digital whiteboards, etc.) with a variety of existing annotations (audio transcriptions, topic segmentations, summaries, etc.). Meetings were recorded in English and fall into two broad types: scenario meetings, where participants were asked to act out roles in a pre-defined scenario, and non-scenario meetings where participants were not restricted by role assignments. In the first case, the scenario was a project meeting about the development of a new TV remote control; participant roles were project manager, industrial designer, marketing expert, etc. The non-scenario meetings are about the move of an academic lab to a new location on campus. The number of participants is four. For our study we selected 10 meetings (5 scenario meetings and 5 non-scenario meetings) and had their audio transcriptions translated into German (our chosen target language) by two native speakers each. Translators were able to simultaneously read the audio transcription of the meeting, view the video, and listen to the audio, when creating the translation. The translation guidelines were designed to obtain translations that match the source text as closely as possible in terms of style – for example, translators were asked to maintain the same level of colloquial as opposed to formal language, and to generally ensure that the translation was pragmatically adequate. Obvious errors in the source text (e.g. errors made by non-native English speakers among the meeting participants) were not rendered by equivalent errors in the German translation but were corrected prior to translation. The final translations were reviewed for accuracy and the data were filtered semi-automatically by eliminating incomplete sentences, false starts, fillers, repetitions, etc. Although these would certainly pose problems in a real-world application of spoken language translation, the goal of this study is not to analyze the impact of speech-specific phenomena on translation performance (which, as discussed in Section 1, has been addressed before) but to assess the impact of contextual information such as discourse and knowledge of the real-world surroundings. Finally, single-word utterances such as *yeah, oh, no, sure*, etc. were downsampled since they are trivial to translate and were very frequent in the corpus; their inclusion would therefore bias the development and tuning of the MT system towards these short utterances at the expense of longer, more informative utterances.

Table 1 shows the word counts of the translated meetings after the preprocessing steps described above. As an indicator of inter-translator

ID	type	# utter.	# word	S-BLEU
ES2008a	S	224	2327	21.5
IB4001	NS	419	3879	24.5
IB4002	NS	447	3246	30.5
IB4003	NS	476	5118	24.1
IB4004	NS	593	5696	26.9
IB4005	NS	381	4719	30.4
IS1008a	S	191	2058	25.8
IS1008b	S	353	3661	24.1
IS1008c	S	308	3351	19.6
TS3005a	S	245	2339	28.1

Table 1: Sizes and symmetric BLEU scores for translated meetings from the AMI corpus (S = scenario meeting, NS = non-scenario meeting).

agreement we computed the symmetric BLEU (S-BLEU) scores on the reference translations (i.e. using one translation as the reference and the other as the hypothesis, then switching them and averaging the results). As we can see, scores are fairly low overall, indicating large variation in the translations. This is due to (a) the nature of conversational speech, and (b) the linguistic properties of the target language. Conversational data contain a fair amount of colloquialisms, referential expressions, etc. that can be translated in a variety of ways. Additionally, German as the target language permits many variations in word order that convey slight differences in emphasis, which is turn is dependent on the translators' interpretation of the source sentence. German also has rich inflectional morphology that varies along with the choice of words and word order (e.g. verbal morphology depends on which subject is chosen).

2.2 SMT System Training Data

Since transcription and translation of multi-party spoken conversations is extremely time-consuming and costly, it is unlikely that parallel conversational data will ever be produced on a sufficiently large scale for a variety of different meeting types, topics, and target languages. In order to mimic this situation we trained an initial English-German SMT system on freely available out-of-domain data resources. We considered the following parallel corpora: news text (de-news[1], 1.5M words), EU parliamentary proceedings (Europarl (Koehn, 2005), 24M words) and EU legal documents (JRC Acquis[2], 35M words), as well as two generic English-German machine-readable dictionaries[3],[4] (672k and 140k entries, respectively).

3 Translation Systems

We trained a standard statistical phrase-based English-German translation system from the resources described above using Moses (Hoang and Koehn, 2008). Individual language models were trained for each data source and were then linearly interpolated with weights optimized on the development set. Similarly, individual phrase tables were trained and were then combined into a single table. Binary indicator features were added for each phrase pair, indicating which data source it was extracted from. Duplicated phrase pairs were merged into a single entry by averaging their scores (geometric mean) over all duplicated entries. The weights for binary indicator features were optimized along with all other standard features on the development set. Our previous experience showed that this method worked better than the two built-in features in Moses for handling multiple translation tables. We found that the JRC corpus obtained very small weights; it was therefore omitted from further system development. Table 2 reports results from six different systems: the first (System 1) is a system that only uses the parallel corpora but not the external dictionaries listed in Section 2.2. System 2 additionally uses the external dictionaries. All systems use two meetings (IB4002 and IS1008b) as a development set for tuning model parameters and five meetings for testing (IB4003-5,IS1008c,TS3005a). For comparison we also trained a version of the system where a small in-domain data set (meetings ES2008a, IB4001, and IS1008a) was added to the training data (System 3). Finally, we also compared our performance against Google Translate, which is a state-of-the-art statistical MT system with unconstrained ac-

[1] www.iccs.inf.ed.ac.uk/~pkoehn/publications/de-news
[2] http://wt.jrc.it/lt/Acquis/
[3] http://www.dict.cc
[4] http://www-user.tu-chemnitz.de/~fri/ding

	System description	Dev set				Eval set			
		OOV (%)		Trans. Scores		OOV (%)		Trans. Scores	
		EN	DE	BLEU	PER	EN	DE	BLEU	PER
System 1	OOD parallel data only	4.1	17.0	23.8	49.0	6.5	20.5	21.1	49.5
System 2	System 1 + dictionaries	1.5	15.9	24.6	47.3	2.8	16.3	21.7	48.4
System 3	System 1 + ID parallel data	3.5	13.4	24.7	47.2	5.8	19.7	21.9	48.3
System 4	System 2 + ID parallel data	1.2	12.9	25.4	46.1	2.5	15.9	22.0	48.2
System 5	System 4 + web data	1.2	12.8	26.0	45.9	2.5	15.8	22.1	48.1
System 6	Google Translate	–	–	25.1	49.1	–	–	23.7	50.8

Table 2: System performance using out-of-domain (OOD) parallel data only vs. combination with a small amount of in-domain (ID) data and generic dictionaries. For each of the development (DEV) and evaluation (Eval) set, the table displays the percentages of unknown word types (OOV) for English (EN) and German (DE), as well as the translation scores of BLEU (%) and PER.

cess to the web as training data (System 6). As expected, translation performance is fairly poor compared to the performance generally obtained on more structured genres. The use of external dictionaries helps primarily in reducing PER scores while BLEU scores are only improved noticeably by adding in-domain data. System 6 shows a more even performance across dev and eval sets than our trained system, which may reflect some degree of overtuning of our systems to the relatively small development set (about 7K words). However, the PER scores of System 6 are significantly worse compared to our in-house systems.

In order to assess the impact of adding web data specifically collected to match our meeting corpus we queried a web portal[5] that searches a range of English-German bilingual web resources and returns parallel text in response to queries in either English or German. As queries we used English phrases from our development and evaluation sets that (a) did not already have phrasal translations in our phrase tables, (b) had a minimum length of four words, and (c) occurred at least twice in the test data. In those cases where the search engine returned results with an exact match on the English side, we word-aligned the resulting parallel text (about 600k words) by training the word alignment together with the news text corpus. We then extracted new phrase pairs (about 3k) from the aligned data. The phrasal scores assigned to the new phrase pairs were set to 1; the lexical scores were computed from a word lexicon trained over both the baseline data resources and the parallel web data. However, results (Row 5 in Table 2) show that performance hardly improved, indicating the difficulty in finding matching data sources for conversational speech.

Table 2 also shows the impact of different data resources on the percentages of unknown word types (OOV) for both the source and target languages. The use of external dictionaries gave the largest reduction of OOV rates (System 1 vs. System 2 and System 3 vs. System 4), followed by the use of in-domain data (System 1 vs. System 3 and System 2 vs. System 4). Since they were retrieved by multi-word query phrases, adding the web data did not lead to significant reduction on the OOV rates (System 4 vs. System 5).

Finally, we also explored a hierarchical phrase-based system as an alternative baseline system. The system was trained using the Joshua toolkit (Li et al., 2009) with the same word alignments and language models as were used in the standard phrase-based baseline system (System 4). After extracting the phrasal (rule) tables for each data source, they were combined into a single phrasal (rule) table using the same combination approach as for the basic phrase-based system. However, the translation results (BLEU/PER of 24.0/46.6 (dev) and 20.8/47.6 (eval), respectively) did not show any improvement over the basic phrase-based system.

[5]http://www.linguee.com

4 Analysis of Baseline Translations: Effect of Contextual Information

The output from System 5 was analyzed manually in order to assess the importance of modeling contextual information. Our goal was not to determine how translation of meeting style data can be improved in general – better translations could certainly be generated by better syntactic modeling, addressing morphological variation in German, and generally improving phrasal coverage, in particular for sentences involving colloquial expressions. However, these are fairly general problems of SMT that have been studied previously. Instead, our goal was to determine the relative importance of modeling different contextual factors, such as discourse-level information or knowledge of the real-world environment, which have not been studied extensively.

We considered three types of contextual information: discourse coherence information (in particular anaphoric relations), knowledge of the topic or domain, and real-world/multimodal information. Anaphoric relations affect the translation of referring expressions in cases where the source and target languages make different grammatical distinctions. For example, German makes more morphological distinctions in noun phrases than English. In order to correctly translate an expression like "the red one" the grammatical features of the target language expression for the referent need to be known. This is only possible if a sufficiently large context is taken into account during translation and if the reference is resolved correctly. Knowledge of the topic or domain is relevant for correctly translating content words and is closely related to the problem of word sense disambiguation. In our current setup, topic/domain knowledge could be particularly helpful because in-domain training data is lacking and many word translations are obtained from generic dictionaries that do not assign probabilities to competing translations. Finally, knowledge of the real-world environment, such as objects in the room, other speakers present, etc. determines translation choices. If a speaker utters the expression "that one" while pointing to an object, the correct translation might depend on the grammatical features of the linguistic expression for that object; e.g. in German, the translation could be "die da", "der da" or "das da". Since the participants in our meeting corpus use slides and supporting documents we expect to see some effect of such exophoric references to external objects.

Error type	% (dev)	% (eval)
Word sense	64.5	68.2
Exophora (addressee)	24.3	23.4
Anaphora	10.2	7.8
Exophora (other)	1.0	0.6

Table 3: Relative frequency of different error types involving contextual knowledge. The total number of errors is 715, for 315 sentences.

In order to quantify the influence of contextual information we manually analyzed the 1-best output of System 5, identified those translation errors that require knowledge of the topic/domain, larger discourse, or external environment for their resolution, classified them into different categories, and computed their relative frequencies. We then corrected these errors in the translation output to match at least one of the human references, in order to assess the maximum possible improvement in standard performance scores that could be obtained from contextual modeling. The results are shown in Tables 3 and 4. We observe that out of all errors that can be related to the lack of contextual knowledge, word sense confusions are by far the most frequent. A smaller percentage of errors is caused by anaphoric expressions. Contrary to our expectations, we did not find a strong impact of exophoric references; however, there is one crucial exception where real-world knowledge does play an important role. This is the correct translation of the addressee *you*. In English, this form is used for the second person singular, second person plural, and the generic interpretation (as in "one", or "people"). German has three distinct forms for these cases and, additionally, formal and informal versions of the second-person pronouns. The required formal/informal pronouns can only be determined by prior knowledge of the relationships among the meeting participants. However, the singular-plural-generic distinction can potentially be resolved by multimodal informa-

	Original		Corrected	
	BLEU (%)	PER	BLEU (%)	PER
dev	26.0	45.9	27.5	44.0
eval	22.1	48.1	23.3	46.0

Table 4: Scores obtained by correcting errors due to lack of contextual knowledge.

tion such as gaze, head turns, body movements, or hand gestures of the current speaker. Since these errors affect mostly single words as opposed to larger phrases, the impact of the corrections on BLEU/PER scores is not large. However, for practical applications (e.g. information extraction or human browsing of meeting translations) the correct translation of content words and referring expressions would be very important. In the remainder of the paper we therefore describe initial experiments designed to address the most important source of contextual errors, viz. word sense confusions.

5 Resolving Word Sense Disambiguation Errors

The problem of word sense disambiguation (WSD) in MT has received a fair amount of attention before. Initial experiments designed at integrating a WSD component into an MT system (Carpuat and Wu, 2005) did not meet with success; however, WSD was subsequently demonstrated to be successful in data-matched conditions (Carpuat and Wu, 2007; Chan et al., 2007). The approach pursued by these latter approaches is to train a supervised word sense classifier on different phrase translation options provided by the phrase table of an initial baseline system (i.e. the task is to separate different phrase senses rather than word senses). The input features to the classifier consist of word features obtained from the immediate context of the phrase in questions, i.e. from the same sentence or from the two or three preceding sentences. The classifier is usually trained only for those phrases that are sufficiently frequent in the training data.

By contrast, our problem is quite different. First, many of the translation errors caused by choosing the wrong word sense relate to words obtained from an external dictionary that do not occur in the parallel training data; there is also little in-domain training data available in general. For these reasons, training a supervised WSD module is not an option without collecting additional data. Second, the relevant information for resolving a word sense distinction is often not located in the immediately surrounding context but it is either at a more distant location in the discourse, or it is part of the participants' background knowledge. For example, in many meetings the opening remarks refer to slides and an overhead projector. It is likely that subsequent mentioning of *slide* later on during the conversation also refer to overhead slides (rather than e.g. *slide* in the sense of "playground equipment"), though the contextual features that could be used to identify this word sense are not located in the immediately preceding sentences. Thus, in contrast to supervised, local phrase sense disambiguation employed in previous work, we propose to utilize unsupervised, *global* word sense disambiguation, in order to obtain better modeling of the topic and domain knowledge that is implicitly present in meeting conversations.

5.1 Unsupervised Word Sense Disambiguation

Unsupervised WSD algorithms have been proposed previously (e.g. (Navigli and Lapata, 2007; Cheng et al., 2009)). The general idea is to exploit measures of word similarity or relatedness to jointly tag all words in a text with their correct sense. We adopted the graph-based WSD method proposed in (Sinha and Mihalcea, 2007), which represents all word senses in a text as nodes in an undirected graph $G = (V, E)$. Pairs of nodes are linked by edges weighted by scores indicating the similarity or relatedness of the words associated with the nodes. Given such a graph, the likelihood of each node is derived by the PageRank algorithm (Brin and Page, 1998), which measures the relative importance of each node to the entire graph by considering the amount of "votes" the node receives from its neighboring nodes. The PageRank algorithm was originally designed for directed graphs, but can be easily extended to an undirected graph. Let $PR(v_i)$ denote the PageRank score of v_i. The PageRank algorithm itera-

tively updates this score as follows:

$$PR(v_i) = (1-d) + d \sum_{(v_i,v_j)\in E} PR(v_j) \frac{w_{ij}}{\sum_k w_{kj}}$$

where w_{ij} is the similarity weight of the undirected edge (v_i, v_j) and d is a damping factor, which is typically set to 0.85 (Brin and Page, 1998). The outcome of the PageRank algorithm is numerical weighting of each node in the graph. The sense with the highest score for each word identifies its most likely word sense. For our purposes, we modified the procedure as follows. Given a document (meeting transcription), we first identify all content words in the source document. The graph is then built over all target-language translation candidates, i.e. each node represents a word translation. Edges are then established between all pairs of nodes for which a word similarity measure can be obtained.

5.2 Word Similarity Measures

We follow (Zesch et al., 2008a) in computing the semantic similarity of German words by exploiting the Wikipedia and Wiktionary databases. We use the publicly available toolkits JWPL and JWKTL (Zesch et al., 2008b) to retrieve relevant articles in Wikipedia and entries in Wiktionary for each German word – these include the first paragraphs of Wikipedia articles entitled by the German word, the content of Wiktionary entries of the word itself as well as of closely related words (hypernyms, hyponyms, synonyms, etc.). We then concatenate all retrieved material for each word to construct a pseudo-gloss. We then lowercase and lemmatize the pseudo-glosses (using the lemmatizer available in the TextGrid package [6]), exclude function words by applying a simple stop-word list, and compute a word similarity measure for a given pair of words by counting the number of common words in their glosses.

We need to point out that one drawback in this approach is the low coverage of German content words in the Wikipedia and Wiktionary databases. Although the English edition contains millions of entries, the German edition of Wikipedia and Wiktionary is much smaller – the coverage of all content words in our task ranges between 53% and 56%, depending on the meeting, which leads to graphs with roughly 3K to 5K nodes and 8M to 13M edges. Words that are not covered mostly include rare words, technical terms, and compound words.

5.3 Experiments and Results

For each meeting, the derived PageRank scores were converted into a positive valued feature, referred to as the WSD feature, by normalization and exponentiation:

$$f_{WSD}(w_g|w_e) = \exp\left\{\frac{PR(w_g)}{\sum_{w_g\in H(w_e)} PR(w_g)}\right\}$$

where $PR(w_g)$ is the PageRank score for the German word w_g and $H(w_e)$ is the set of all translation candidates for the English word w_e. Since they are not modeled in the graph-based method, multi-words phrases and words that are not found in the Wikipedia or Wiktionary databases will receive the default value 1 for their WSD feature. The WSD feature was then integrated into the phrase table to perform translation. The new system was optimized as before.

It should be emphasized that the standard measures of BLEU and PER give an inadequate impression of translation quality, in particular because of the large variation among the reference translations, as discussed in Section 4. In many cases, better word sense disambiguation does not result in better BLEU scores (since higher gram matches are not affected) or even PER scores because although a feasible translation has been found it does not match any words in the reference translations. The best way of evaluating the effect of WSD is to obtain human judgments – however, since translation hypotheses change with every change to the system, our original error annotation described in Section 4 cannot be re-used, and time and resource constraints prevented us from using manual evaluations at every step during system development.

In order to loosen the restrictions imposed by having only two reference translations, we utilized a German thesaurus[7] to automatically extend the content words in the references with synonyms. This can be seen as an automated way of

[6] http://www.textgrid.de/en/beta.html

[7] http://www.openthesaurus.de

	No WSD			With WSD		
	BLEU (%)	PER	XPER	BLEU (%)	PER	XPER
dev	25.4	46.1	43.4	25.4	45.6	42.9
eval	22.0	48.2	44.6	22.0	47.9	44.0
IB4003	21.4	48.3	44.4	21.4	47.5	43.8
IB4004	22.4	48.5	44.4	23.1	48.4	43.9
IB4005	25.4	45.9	42.4	25.3	45.6	42.2
IS1008c	15.9	52.9	50.0	14.9	52.3	48.6
TS3005a	23.1	45.2	41.9	23.2	45.3	41.7

Table 5: Performance of systems with and without WSD for dev and eval sets as well as individual meetings in the eval set.

approximating the larger space of feasible translations that could be obtained by producing additional human references. Note that the thesaurus provided synonyms for only roughly 50% of all content words in the dev and eval set. For each of them, on average three synonyms are found in the thesaurus. We use these extended references to recompute the PER score as an indicator of correct word selection. All results (BLEU, PER and extended PER (or XPER)) are shown in Table 5. As expected, BLEU is not affected but WSD improves the PER and XPER slightly but consistently. Note that this is despite the fact that only roughly half of all content words received disambiguation scores.

Finally, we provide a concrete example of translation improvements, with improved words highlighted:
Source:
on the balcony
there's that **terrace**
there's no place inside the building
Translation, no WSD:
auf dem balkon
es ist das **absatz**
es gibt keinen platz innerhalb des gebäudes
Translation, with WSD:
auf dem balkon
es ist das **terrasse**
es gibt keinen platz gebäudeintern
References:
auf dem balkon / auf dem balkon
da gibt es die **terrasse** */ da ist die* **terrasse**
es gibt keinen platz im gebäude / es gibt keinen platz innen im gebäude

6 Summary and Conclusions

We have presented a study on statistical translation of meeting data that makes the following contributions: to our knowledge it presents the first quantitative analysis of contextual factors in the statistical translation of multi-party spoken meetings. This analysis showed that the largest impact could be obtained in the area of word sense disambiguation using topic and domain knowledge, followed by multimodal information to resolve addressees of *you*. Contrary to our expectations, further knowledge of the real-world environment (such as objects in the room) did not show an effect on translation performance. Second, it demonstrates the application of *unsupervised, global* WSD to SMT, whereas previous work has focused on supervised, local WSD. Third, it explores definitions derived from collaborative Wiki sources (rather than WordNet or existing dictionaries) for use in machine translation. We demonstrated small but consistent improvements even though word coverage was incomplete. Future work will be directed at improving word coverage for the WSD algorithm, investigating alternative word similarity measures, and exploring the combination of global and local WSD techniques.

Acknowledgments

This work was funded by the National Science Foundation under Grant IIS-0840461 and by a grant from the University of Washington's Provost Office. 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 funding organizations.

References

S. Brin and L. Page. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". *Proceedings of WWW7*.

C. Callison-Burch, P. Koehn and M. Osborne. 2006. "Improved Statistical Machine Translation Using Paraphrases". *Proceedings of NAACL*.

M. Carpuat and D. Wu. 2005. "Word sense disambiguation vs. statistical machine translation". *Proceedings of ACL*.

M. Carpuat and D. Wu. 2007. "Improving statistical machine translation using word sense disambiguation". *Proceedings of EMNLP-CoNLL*.

Y.S. Chan and H.T. Ng and D. Chiang 2007. "Word sense disambiguation improves statistical machine translation". *Proceedings of ACL*.

P. Chen, W. Ding, C. Bowes and D. Brown. 2009. "A fully unsupervised word sense disambiguation method using dependency knowledge". *Proceedings of NAACL*.

E. Gabrilovich and S. Markovitch. 2007 "Computing semantic relatedness using Wikipedia-based explicit semantic analysis". *Proceedings of IJCAI*.

R. Haque, S.K. Naskar, Y. Ma and A. Way. 2009. "Using supertags as source language context in SMT". *Proceedings of EAMT*.

A.S. Hildebrandt, M. Eck, S. Vogel and A. Waibel. 2005. "Adaptation of the Translation Model for Statistical Machine Translation using Information Retrieval". *Proceedings of EAMT*.

H. Hoang and P. Koehn. 2008. "Design of the Moses decoder for statistical machine translation". *Proceedings of SETQA-NLP*.

P. Koehn. 2005. "Europarl: a parallel corpus for statistical machine translation". *Proceedings of MT Summit*.

V. Kumar, R. Sridhar, S. Narayanan and S. Bangalore. 2008. "Enriching spoken language translation with dialog acts". *Proceedings of HLT*.

Z. Li et al.. 2009. "Joshua: An Open Source Toolkit for Parsing-based Machine Translation". *Proceedings of StatMT*.

E. Matusov, M. Popović, R. Zens and H. Ney. 2004. "Statistical Machine Translation of Spontaneous Speech with Scarce Resources". *Proceedings of IWSLT*.

A. McCowan. 2005. "The AMI meeting corpus",

H. Mima, O. Furuse and H. Iida. 1998. "Improving Performance of Transfer-Driven Machine Translation with Extra-Linguistic Information from Context, Situation and Environment". *Proceedings of Coling*. *Proceedings of the International Conference on Methods and Techniques in Behavioral Research*.

D.S. Munteanu and D. Marcu. 2005. "Improving machine translation performance by exploiting non-parallel corpora". *Computational Linguistics*.

R. Navigli and M. Lapata. 2007. "Graph Connectivity Measures for Unsupervised Word Sense Disambiguation", *Proceedings of IJCAI*

S. Rao and I. Lane and T. Schultz. 2007. "Improving spoken language translation by automatic disfluency removal". *Proceedings of MT Summit*. 31(4).

R. Sinha and R. Mihalcea. 2007. "Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity", *Proceedings of IEEE-ICSC*

N. Stroppa, A. Bosch and A. Way. 2007. "Exploiting Source Similarity for SMT using Context-Informed Features". *Proceedings of TMI*.

S. Stüker, C. Fügen, F. Kraft and M. Wölfel. 2007. "The ISL 2007 English Speech Transcription System for European Parliament Speeches". *Proceedings of Interspeech*.

A. Waibel and C. Fügen. 2008. "Spoken Language Translation – Enabling cross-lingual human-human communication". *Proceedings of Coling)*.

W. Wang, G. Tur, J. Zheng and N.F. Ayan. 2010. "Automatic disfluency removal for improving spoken language translation". *Proceedings of ICASSP*. *IEEE Signal Processing Magazine*

H. Wu, H. Wang and C. Zong. 2008. "Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora",

T. Zesch, C. Müller and Iryna Gurevych. 2008. "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary". *Proceedings of LREC*.

T. Zesch, Christof Müler and Iryna Gurevych. 2008. "Using Wiktionary for Computing Semantic Relatedness".

J. Zheng, W. Wang and N.F. Ayan. 2008. "Development of SRI's translation systems for broadcast news and broadcast conversations". *Proceedings of Interspeech*. *Proceedings of AAAI*.