Unsupervised Translation Disambiguation for Cross-Domain Statistical Machine Translation

Mei Yang
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

Most attempts at integrating word sense disambiguation with statistical machine translation have focused on supervised disambiguation approaches. These approaches are of limited use when the distribution of the test data differs strongly from that of the training data; however, word sense errors tend to be especially common under these conditions. In this paper we present different approaches to unsupervised word translation disambiguation and apply them to the problem of translating conversational speech under resource-poor training conditions. Both human and automatic evaluation metrics demonstrate significant improvements resulting from our technique.

1 Introduction

Word sense errors are a significant source of semantic inadequacy in machine translation. They arise when the system produces a target-language translation that is considered inappropriate given the context. In statistical machine translation (SMT) the problem of word sense disambiguation (WSD) has been investigated as a possible solution; however, the focus has been on supervised WSD approaches that require labeled training data. These are suboptimal when applying an SMT system across domains or styles: when the distributions underlying the training and test data differ strongly, little relevant training material will be available for the word sense classifier. In this study we therefore focus on the problem of *unsupervised word translation disambiguation for SMT*. Our unsupervised approach does not rely on labeled training data but makes use of various word relatedness measures in combination with a graph-based link analysis algorithm. This algorithm produces a probabilistic ranking of different translation candidates, which is then integrated into the SMT system. We compare three monolingual approaches and one bilingual approach to constructing the disambiguation graph and apply our method to the translation of meeting-style conversations from English into German. We evaluate its performance by comparing the ranked word translations against human reference rankings, as well as by applying standard machine translation evaluation procedures.

2 Background and Related Work

The process of automatically selecting the appropriate word sense in a given context is referred to as word sense disambiguation (WSD). In monolingual settings both supervised and unsupervised WSD approaches have been proposed (see (Navigli, 2009) for an overview). Supervised approaches (e.g. (Lee et al., 2004; Joshi et al., 2006)) typically represent a word and its surrounding context as a feature vector. Feature vectors and manually annotated gold labels are then used to train a statistical classifier which is subsequently applied to unlabeled test data to predict the correct sense for each word instance. State-of-the-art unsupervised WSD algorithms mostly rely on graph-based ranking algorithms (Mihalcea, 2005; Navigli and Lapata, 2007), where all possible senses of the words of

interest in a sentence or document are represented as nodes in a graph. Nodes are connected by edges weighted with a similarity or relatedness measure. A link analysis algorithm such as PageRank (Brin and Page, 1998) is used to assign scores to each node; nodes representing mutually exclusive word senses are then ranked by this score and the top node is chosen as the correct word sense.

In machine translation the problem is to identify the most appropriate target-language translation for a source-language word in a given context. The ultimate goal is to discriminate between actual lexical items and not word sense labels per se – although different translation options often correspond to different senses of the source-language word they may also represent alternative translations of the same source-language word sense. We therefore refer to this problem as *word translation disambiguation* (WTD). WTD has been addressed in e.g. (Dagan and Itai, 1994; Kikui, 1999; Li and Li, 2002). However, these studies predate the current SMT framework; they evaluated only small subsets of words and did not use standard machine translation evaluation measures to assess the impact on translation performance. Previous studies addressing the WTD problem in SMT have relied on supervised approaches that involve training statistical classifiers to distinguish between all of the translation options (single words or phrases in a phrase-based SMT system) for the most frequent source phrases. Classifiers were trained based on features of the source phrases and their contexts in the parallel training data. In early experimental investigations (e.g. (Carpuat and Wu, 2005; Cabezas and Resnik, 2005) results based on this approach were inconclusive and showed a lack of improvement or only marginal improvements in standard automatic evaluation scores. Statistically significant improvements to state-of-the-art SMT systems were later reported in (Gimenez and Marquez, 2007; Chang et al., 2007; Carpuat and Wu, 2007).

When large amounts of in-domain parallel training data are not available, supervised WTD is suboptimal. However, unsupervised WTD has not been investigated as an alternative. The work most closely related to our scenario is the cross-lingual word sense disambiguation benchmark task in the recent 2010 SemEval evaluations (Lefever and Hoste, 2010). Here, systems were required to generate sets of target-language translations for English source words, each of which was embedded in an example sentence. Proposed translations were compared against manually annotated translations and were evaluated by precision and recall. This task is similar to ours in that no gold labels were provided and translations were judged with respect to how context-appropriate they were. The difference is that the SemEval task only included nouns whereas our goal is to disambiguate translation candidates for *all* open-class words in the source language. In addition, there was no mismatch between the test data and the data that provided the translation inventory. By contrast, in this study we have only limited domain-matched parallel training data, which increases the difficulty of the problem. Finally, we also evaluate the effect of WTD on end-to-end translation performance.

3 Data and System

Our study is part of a research effort on translating unconstrained conversational speech. For this purpose we use the translated portion of the the AMI multimodal meeting corpus (Carletta, 2007) described in (Yang and Kirchhoff, 2010). This subset consists of 10 meetings whose audio transcriptions were translated from English into German. The length of a single meeting conversation ranges between 2300 and 5700 words; in total the corpus contains roughly 36K words. Translating meeting conversations is a difficult task and one of the unsolved challenges for SMT – in addition to the wide variety of topics and domains encountered in meetings, conversational speech differs strongly from text in style. Large amounts of relevant training data are non-existent and generally hard to collect since it requires both the transcription and translation of conversational speech. Thus, by necessity, out-of-domain training data needs to be used for system development.

	BLEU (%)	PER
Baseline	22.2	48.1
Google Translator	23.7	50.8

Table 1: BELU(%) and PER scores of the baseline SMT system and Google Translator on the evaluation set.

A phrase-based English-German SMT baseline system was trained using the Moses infrastructure (Hoang and Koehn, 2008). The training data consisted of the `de-news` corpus[1] (1.5M words of English-German newswire text), the Europarl corpus (Koehn, 2005) (24M words of European parliamentary proceedings text), supplemented by two generic English-German machine-readable dictionaries[2]. Three of the 10 translated meetings were added to the training data as in-domain data; two are used for development, and five are used for evaluation. Separate phrase tables and language models were trained for each data source and were then interpolated, optimizing the model weights on the development set. More details on baseline system training can be found in (Yang and Kirchhoff, 2010). Table 1 shows the BLEU and position-independent error rate (PER) score of the baseline system on the evaluation data, in comparison to that of a freely available generic SMT system (Google Translate[3]). Both systems have similar performance; Google Translate achieves better BLEU scores but lower PER. This is explained by the fact that our system achieves more 1-gram matches but has worse word ordering than Google Translate. In addition to a standard phrase-based system, a hierarchical phrase-based system (Chiang, 2005; Li et al., 2009) was trained for comparison; however, its performance only differed insignificantly from that of the standard phrase-based system.

4 Word Translation Disambiguation Approaches

Following monolingual unsupervised WSD approaches, and expanding on (Yang and Kirchhoff, 2010), we utilize a graph-based ranking algorithm for disambiguation. We compare two different graph-based methods, a monolingual method that only uses information from the target language, and a bilingual method combining information from both source and target language.

4.1 Monolingual Disambiguation

For each test document (i.e. each individual meeting) we construct a graph G with a set of nodes V, a set of edges $E \in V \times V$, and a semantic relatedness function W. The set of nodes is determined by extracting all open-class words from the source text along with all their possible single-word target-language translations obtained from any of the training resources listed in Section 3. The part-of-speech (POS) tags of source and target words were obtained by first applying TreeTagger[4] and then collapsing the tags into the four major categories noun, verb, adjective and adverb. Words with tags out of these four categories were discarded. In determining allowable translations, POS matching was enforced (e.g., target-language translations of nouns have to be nouns as well). Though this constraints is by no means universally valid, it helps greatly in removing noisy translation options for this particular language pair. Target-language words are then represented as nodes $v_1, ..., v_n$ in the graph, with their POS tag information associated so that each unique combination of a target-language word and a POS tag is represented only once, irrespective of the number of source-language words that list it as a possible translation. Records are kept of which nodes are mutually exclusive (i.e. derive from the same source-language word), and the locations of all source-language words from which a node is derived from. The latter is used in computing distances between nodes. A function w is then applied to each pair of nodes $(v_i, v_j), i, j \in 1, ..., n$ that measures the relatedness of the associated words. If the result w_{ij} exceeds a given threshold, and if v_i and v_j are located within a window comprising at most 3

[1] www.iccs.inf.ed.ac.uk/~pkoehn/publications/de-news

[2] www.dict.cc and www-user.tu-chemnitz.de/ding

[3] www.google.com/translate

[4] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

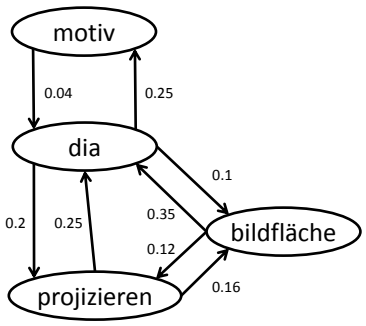

Figure 1: An example monolingual graph. The relatedness scores in the graph are not symmetric because of the normalization factor in Equation 3.

sentences, an edge from v_i to v_j is created and labeled with w_{ij}. The windowing constraint is applied in order to eliminate noise and to reduce the size of the graph. If multiple locations are associated with v_i and/or v_j, the windowing constraint applies to the locations that lead to the minimal distance between v_i and v_j. Figure 1 shows an example graph.

The edge weights express the semantic relatedness of words. Traditionally, three main sources of information have been used to determine semantic relatedness: lexical semantic resources like WordNet (Pedersen et al., 2004), dictionary glosses (Lesk, 1986), and word co-occurrence counts (Gabrilovich and Markovitch, 2005; Mihalcea and Corley, 2006). Wordnet-style resources provide an explicit account of semantic relationships between words, such as hyponymy and hyperonymy but they only exist for a small number of languages and typically have low coverage. The gloss-based approach computes the degree of word overlap in dictionary glosses for different words. While traditional dictionaries may also suffer from the low coverage problem, more recent approaches use Wikimedia (Wiktionary and/or Wikipedia) resources for this purpose (Mihaelcea, 2007; Ponzetto and Strube, 2007; Zesch et al., 2008a; Zesch et al., 2008b). The advantages of Wikimedia are: (a) larger (and still growing) coverage than either WordNet or standard dictionaries; (b) public availability; and (c) inclusion of not only sense glosses but also explicit links to translations of word senses into other languages in Wiktionary.

Finally, word co-occurrence based approaches count the relative frequency of two words occurring in the same document, or some measure derived from such counts (such as mutual information). We compare three monolingual disambiguation approaches exploiting each type of relatedness measure:

Wordnet-based relatedness

We use GermaNet (Henrich and Hinrichs, 2010) to compute a path-based word relatedness score, expressed as the reciprocal of the average distance between two words in the GermaNet database across all combinations of synsets to which either word belongs.

$$w(i,j) = \frac{1}{1 + \sum_{s_i,s_j} d(s_i,s_j) / \sum_{s_i,s_j} 1} \quad (1)$$

$$\forall(s_i,s_j) \in \{s_i \in S_i, s_j \in S_j, d(s_i,s_j) < \inf\}$$

where S_i is the set of synsets containing word i and $d(s_i,s_j)$ is the shortest-path distance between s_i and s_j as computed by the `get_shortest_path` routine in the GermaNet perl API. When no path is found, the shortest-path distance is infinite. For example, synsets of different POS classes are disconnected from each other in GermaNet and their shortest-path distances are thus infinite. In order to avoid penalty on words belonging to multiple synsets, synset pairs of infinite distance are not accumulated in Equation 1.

Dictionary-based relatedness

In our second approach we follow (Zesch et al., 2008a) and compute semantic relatedness from the degree of lexical overlap in pseudo-glosses derived from Wiktionary. For each target-language word in the graph we retrieve its entries from German Wiktionary[5] and concatenate all glosses to a single pseudo-gloss. This method merges the glosses for different senses of the target word, which may decrease the reliability of this approach; however, it also provides richer glosses for words whose individual sense glosses are very short (consisting of two or three words only). Words in the pseudo-gloss are lowercased and lemmatized using TreeTagger. Punctuations and function words are removed. Given

[5]Wiktionary dump up to the date of 2012/02/04

German: **Ordner**
Wiktionary descriptions:
[1] ein grosser, robuster Umschlag, in den man Blätter (meist Dokumente) einheftet [2] (Computer) ein Ablagefach in einem Speichermedium, in dem Dateien, Dokumente und Unterordner abgelegt werden können. Ordner sind in eine Verzeichnisstruktur eingebettet [3] ordnende Person bei Massenveranstaltungen
Pseudo-gloss:
Ordner: gross robust umschlag man blatt meist dokument einheften computer ablagefach speichermedium datei dokument unterordner ablegen werden können ordner sein verzeichnisstruktur einbetten ordnen person massenveranstaltung

Table 2: German Wiktionary definitions of *Ordner* and the resulting pseudo-gloss.

two words i and j and their associated pseudo-glosses g_i and g_j, their relatedness $R(i,j)$ is calculated by Equation 2, where $C(g_i, g_j)$ is the number of common words in the pseudo-glosses g_i and g_j. A word occurring multiple times in both glosses is counted only once.

$$R(i,j) = \begin{cases} C(g_i,g_j) & \text{if} \quad C(g_i,g_j) > 1 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

The final relatedness function $w(i,j)$ is then determined as follows:

$$w(i,j) = R(i,j) * min(1, exp^{1-|g_j|/L}) \qquad (3)$$

where L is the average length of all pseudo-glosses constructed. The last term in this equation serves as a normalization factor accounting for the fact that a word with a longer pseudo-gloss is more likely to have a higher degree of overlap with another word.

Co-occurrence Based Relatedness

Our third approach measures the co-occurrence count of words within Wikipedia documents. We use Wikipedia as one of the largest publicly available corpora for German, totalling 700K documents and 470M words[6]. Again, word forms are lemmatized using TreeTagger and word forms occurring only once in a document are removed. Let $C_d(i)$ be the number of documents containing word form i and $C_d(i,j)$ be the number of documents containing both i and j. The relatedness measure is then computed as point-wise mutual information:

$$\begin{aligned} w(i,j) &= \log \frac{p(i,j)}{p(i) * p(j)} \\ &= \log \frac{C_d(i,j) * N}{C_d(i) * C_d(j)} \end{aligned} \qquad (4)$$

where N is the total number of documents in Wikipedia. The three procedures achieve different coverage rates on the German word translations, as shown in Table 3. The count-based method obtains the highest coverage, followed by GermaNet and Wiktionary respectively.

Set	**# words**	A	B	C
dev set	2056	72%	60%	84%
eval set	3954	69%	55%	81%

Table 3: Coverage rates of translations in the development and evaluation sets for Methods A (GermaNet-based shortest path), B (Wiktionary-based gloss overlap), and C (Wikipedia-based co-ocurrence counts). Column 2 shows the total number of unique word forms to be disambiguated. POS tag restrictions were observed in computing coverage rates.

Scoring

To score all target-language translations for a given source word we use *personalized* PageRank (Haveliwala, 2002; Agirre and Soroa, 2009), which exploits prior weights on nodes of interest along with the properties of the graph structure. For each node $v_i \in V$, let $In(v_i)$ be the set of nodes that point to v_i and $Out(v_i)$ the set of nodes that v_i points to. Although they are al-

[6]Wikipedia dump up to the date of 2010/10/13

ways the same in monolingual graphs, $In(v_i)$ can be different from $Out(v_i)$ in bilingual graphs as described in Section 4.2. $R(v_i)$, the PageRank score of v_i, is computed iteratively as

$$R(v_i) = (1-d)u_i + d\sum_{v_j \in In(v_i)} R(v_j)\frac{w_{ji}}{\sum_{k \in Out(v_j)} w_{jk}} \quad (5)$$

where the damping factor d is set to 0.85 and the vector $\mathbf{u} = (u_1, u_2, \ldots, u_{|V|})$ varies with the source-language word. Given a source-language word w, the vector $\mathbf{u}(w)$ assigns stronger probabilities to the nodes $Sense(w)$ that represent the word senses of w (only in bilingual graphs) and the nodes $Trans(w)$ that represent the target-language translations of w in the phrase-based translation table:

$$u_i^{'}(w) = \begin{cases} 1 & i \in Sense(w) \\ 0.1 & i \in Trans(w) \\ 0 & otherwise \end{cases}$$

$$\mathbf{u}(w) = \frac{\mathbf{u}^{'}(w)}{|\mathbf{u}^{'}(\mathbf{w})|} \quad (6)$$

The personalized PageRank algorithm ranks all nodes with $\mathbf{u}(w)$ and the translations of w are then ranked by their PageRank scores.

4.2 Bilingual disambiguation

English Wiktionary provides not only dictionary glosses but also translations into other languages for some of the English word senses. In order to make full use of this information we construct bilingual WTD graphs – these consist of two loosely coupled graphs, one that represents source-language word senses and the other that represents target-language translation candidates, as before. The links between these two graphs are the translation associations between source-language senses and target-language words extracted from Wiktionary, as demonstrated in Figure 2. Unlike monolingual graphs, where translation options are distinguished solely based on their relationships with other words in the target language, bilingual graphs simultaneously rank the source-language word senses and their target-language translations

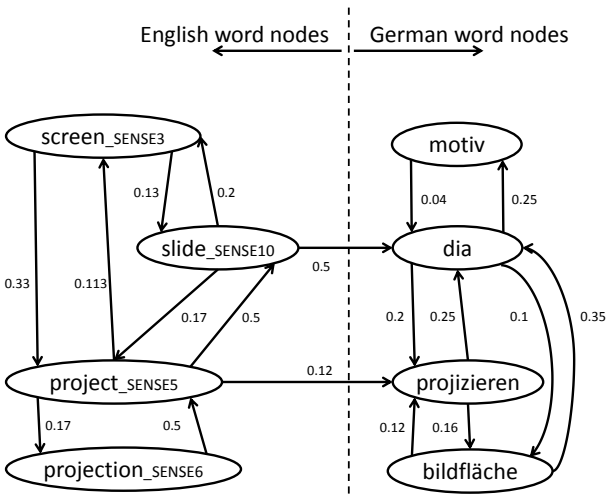

Figure 2: An example bilingual graph. For illustration purpose, nodes representing English senses and German words are placed on the left and right sides of the graph respectively. Each English node is labeled by the word form and the sense ID. For example, the node "slide_SENSE10" represents the 10-th sense of the English word "slide".

using the same graph-based disambiguation algorithm. They are constructed as follows: the target-language words and corresponding nodes in the graph are established as described above. Additionally, the entries of the corresponding source-language words in English Wiktionary[7] are extracted. English Wiktionary entries explicitly distinguish different senses of English words and provide links between word senses and translations into other languages, including German. Among the English word forms of interest, 82% have at least one sense that contains links into German translations, though the overall percentage of English senses with German translations is lower at 35%, because such information tends to be missing for less commonly used word senses. For each English word sense, a graph node on the source side is established, and the mappings from word senses to translations are used to define links between source graph and target graph nodes. As previously, we incorporate the constraint that words must occur within a window of 3 sentences, both on the source-language and

[7]Wiktionary dump up to 2012/02/03

target-language side. Formally, the set of nodes V in the graph G is now partitioned into two subsets, S for the source-language senses and T for the target-language words. $V = S \cup T$. The set of edges E is partitioned into three subsets, edges connecting nodes from S only ($E_S = S \times S$), those connecting target nodes only ($E_T = T \times T$), and the bilingual edges ($E_B = S \times T$) that connect a source and a target node. Note that bilingual edges are determined solely by the translation associations in the Wiktionary database (considered reliable information), without regard to the phrase table of our baseline MT system; therefore source and target nodes may not necessarily be connected even though the corresponding words are translations of each other in the phrase table.

For each English word sense, we construct a pseudo-gloss by retrieving its description from the English Wiktionary. Unlike the procedure used for creating pseudo-glosses for our monolingual graph, we treat every English word sense separately; thus, descriptions of different senses for the same English word are not merged. Table 4 shows the senses and corresponding glosses of the English word *folder*.

The weights of monolingual edges connecting nodes in the same language are then determined as shown in Equation 3. The weights of bilingual edges connecting nodes of English sense to German words are uniformly set to 1. Finally, the PageRank algorithm is run as described above.

Multilingual graphs were used for cross-lingual WSD in (Silberer and Ponzetto, 2010), where graphs were constructed over nodes representing source words and their translations into different target languages, provided by word-aligned parallel corpora. Differing from our work, the use of multilingual graphs there was to disambiguate the sense of a given source word. Translation candidates corresponding to the selected sense were subsequently ranked by their translation frequencies in the parallel corpora but not by the scores induced from the graphs. In addition, bilingual edges in the graphs, which encode translation relationships, were only used to select disambiguating evidence in the target languages and did not contribute to the graph-based score computation.

5 Evaluation

We evaluate the performance of our disambiguation algorithms in a stand-alone fashion as well as in an integrated end-to-end system evaluation. The first method measures disambiguation performance only and compares the rankings obtained by various WTD systems against reference human rankings. The second method analyzes the impact on overall translation performance under the standard MT evaluation measures BLEU and PER.

5.1 Human Evaluation

Human judgments for the stand-alone evaluate were collected using Amazon Mechanical Turk. Each human intelligence task (HIT) consisted in ranking suggested target-language translations for a given source word in its context. We randomly selected a subset of 61 source words that have multiple possible translations. For each source word we collected all single-word translations from the top 20 translations in the baseline phrase table (20 is also the cut-off limit used during decoding). The total number of translation options per source word varied between 3 and 11. In total, 359 different HITs were created, each of which was processed by three different workers, generating three reference rankings for each word. Workers were instructed to assign consecutive ranks from 1 through N to the N translation options. All translations provided had to be assigned a unique rank; results with missing or tied ranks were automatically rejected. We did not require workers to pass a qualification language test; however the HIT description was entirely in German, and results were cross-checked by a native German speaker.

Based on the reference rankings we evaluate the automatically generated rankings produced by our baseline system (no disambiguation, phrase table probabilities only) and the 4 different WTD systems. We use two metrics commonly used for comparing rankings: normalized discounted cumulative gain (NDCG)

English: **folder**
Wiktionary descriptions
[1] An organizer that papers are kept in, usually with an index tab, to be stored as a single unit in a filing cabinet. I keep all my schoolwork in a yellow folder. [2] (computing) A virtual container in a computer's file system, in which files and other folders may be stored. The files and subfolders in a folder are usually related. My essays are in the folder marked "Essays" [3] A machine or person who folds things.
Pseudo-gloss
folder_SENSE1 organizer paper be keep usual index tab store single unit filing cabinet **folder_SENSE2** virtual container computer file system file other folder may be store file sub-folders folder be usual relate **folder_SENSE3** machine person fold thing

Table 4: German/English Wiktionary descriptions and resulting pseudo-glosses. Lemmatization and lowercasing steps were applied. Punctuations and function words were removed.

and mean reciprocal rank (MRR). NDCG measures the accumulated correctness of a ranked list, with the correctness of each item discounted by its position in the list. MRR, by contrast, assumes that only the first item in the reference ranking is correct and measures the rank of this item only in the predicted ranking, ignoring the ranks of others. Given a source word and a predicted ranking of its translations, let $w_1, w_2, \ldots, w_n$ be the translations in order in the predicted ranking and $H(w_i)$ be their ranks in the reference ranking. Then, the NDCG score of the predicted ranking at position $p = 3$ is computed as in Equation 7. Note that the ideal DCG (IDCG) is the DCG score of the ranking where $w_1, w_2, \ldots, w_n$ are re-sorted by their ranks in the reference ranking, instead of the predicted ranking.

$$DCG_3 = \frac{1}{H(w_1)} + \sum_{i=2}^{3} \frac{1/H(w_i)}{log_2 i}$$

$$NDCG_3 = \frac{DCG_3}{IDCG_3} \quad (7)$$

Let w^* be the top-1 translation in the reference ranking and $M(w^*)$ its rank in the predicted ranking. Then, the MRR score of the predicted ranking is computed as in Equation 8. If w^* does not exist in the predicted ranking, then $M(w^*)$ is infinite and the MRR score becomes 0.

$$MRR = \frac{1}{M(w^*)} \quad (8)$$

5.2 Automatic Evaluation

To conduct end-to-end SMT experiments we add the WTD score to the other feature functions in the log-linear model in the SMT system. Words not covered by the WTD system receive default scores. The SMT system is then re-tuned by optimizing all feature function weights jointly on the development set. Note that the WTD feature affects single-word translations only; multiword phrases are not affected. We did experiment with integrating the WTD score into the scores for phrasal translations as well; however this has not yet resulted in significantly different results.

6 Results

Table 5 shows the results from the human evaluation. Out of the four different unsupervised WTD methods, the bilingual system achieves the best results overall: it shows significant improvements in NDCG and only non-significant changes in MRR. The three monolingual systems perform closely. While all monolingual systems show improvements in NDCG, they also tend to deteriorate MRR significantly. Monolingual methods globally rank the set of context-appropriate options higher than context-inappropriate options but often fail to discriminate between the top two or three translation options, especially if they are near equivalents. MRR focuses on the ranking of the top reference translation only and will therefore de-

	IB4003	IB4004	IB4005	IS1008c	TS3005a	Eval (5 meetings)
Baseline	0.70/0.66	0.71/0.69	0.80/0.77	0.73/0.62	0.81/0.62	0.75/0.67
Gloss	**0.87**/0.52	**0.89**/0.53	0.73/**0.49**	0.78/0.46	**0.90**/0.58	**0.83**/**0.51**
GermaNet	**0.85**/0.56	0.81/**0.42**	0.84/**0.52**	0.80/0.65	0.86/0.60	**0.83**/**0.55**
Counts	0.82/0.54	**0.85**/0.62	0.76/0.59	0.79/0.59	**0.91**/0.54	**0.82**/**0.58**
Bilingual	**0.90**/0.59	**0.93**/0.60	0.85/**0.54**	0.88/0.64	**0.95**/0.78	**0.90**/0.62

Table 5: Performance of automatically generated rankings (baseline SMT system without WTD, and different WTD-based systems) compared against human rankings on the individual meetings, and the entire evaluation set (5 meetings combined). Numbers are reported as "NDCG/MRR". Significant differences compared to the baseline are highlighted in bold.

Meeting set	IB4003	IB4004	IB4005	IS1008c	TS3005a	Eval (5 meetings)
Baseline	21.8/47.9	22.5/48.8	25.7/45.7	16.1/52.5	22.7/45.4	22.2/48.1
Gloss	22.2/**47.1**	**23.5**/48.2	24.9/45.0	15.9/52.2	**24.0**/44.3	22.4/**47.4**
GermaNet	21.8/**47.0**	**23.5**/**47.8**	25.2/45.0	16.3/52.3	23.4/45.2	22.4/**47.4**
Counts	22.0/47.5	23.0/48.3	25.6/45.9	15.8/53.0	23.3/45.5	22.4/48.0
Bilingual	22.1/**47.1**	**23.7**/**48.1**	25.0/**45.0**	15.8/52.3	**24.0**/**44.2**	22.5/**47.4**

Table 6: BLEU(%)/PER scores of baseline SMT system without WTD, and SMT systems with different WTD methods on the individual meetings and the entire evaluation set (5 meetings combined). Significant differences compared to the baseline are highlighted in bold.

teriorate even if that translation comes in second (as opposed to a much lower rank) after WTD. The results from the end-to-end evaluation (based on the two manual reference translations for the meeting data) is shown in Table 6. Since we only disambiguate individual words rather than phrases we did not expect to see a large effect on BLEU, which is primarily sensitive to changes to higher-order n-grams in the translation output. As can be seen from Table 6, it is indeed the PER score that is affected more strongly. While BLEU on the total evaluation set is increased only slightly, 0.3 points by our best method, PER is reduced by 0.7%, which is statistically significant at the 0.05 level. Again, it is the bilingual disambiguation method that produces the best results. The effect of WTD also differs depending on the meeting. While significant improvements in both BLEU and PER are obtained in *IB4004*, other meetings are harder to process; in particular, *IS1008c* – this meeting exhibits more technical vocabulary than the others and is dominated by one speaker who is not a native English speaker.

7 Conclusions

We have presented different approaches to unsupervised translation disambiguation and have evaluated their performance within a SMT system for meeting-style speech. Our results show that significant improvements can be obtained from unsupervised WTD; however, improvements are strongly dependent on the nature of the test data, in this case the meeting. Out of the different approaches we investigated the bilingual approach exploiting both source-language and target-language information yielded the best results. We intend to further test the different methods using different test data, especially more homogeneous written-style text.

References

E. Agirre and A. Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of EACL*, pages 33–41.

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW7*.

C.P. Cabezas and P. Resnik. 2005. Using WSD techniques for lexical selection in statistical machine

translation. Technical Report UMIACS-TR-2005-42, University of Maryland.

J. Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, 41(2):181–190.

M. Carpuat and D. Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *ACL*, pages 387–394.

M. Carpuat and D. Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of EMNLP-CoNLL*, pages 61–72.

Y.S. Chang, H. Tou Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of ACL*, pages 33–40.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*, pages 263–270.

I. Dagan and A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

E. Gabrilovich and S. Markovitch. 2005. Feature generation for text categorization using world knowledge. In *Proceedings of IJCAI*, pages 1048–1053.

J. Gimenez and L. Marquez. 2007. Context-aware discriminative phrase selection for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 159–166.

T. H. Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of WWW*.

V. Henrich and E. Hinrichs. 2010. GernEdiT : A graphical tool for GermaNet development. In *Proceedings of the ACL 2010 System Demonstrations*.

H. Hoang and P. Koehn. 2008. Design of the Moses decoder for statistical machine translation. In *Proceedings of SETQA-NLP*, pages 58–65.

M. Joshi, S. Pakhomov, T. Pedersen, R. Maclin, and C. Chute. 2006. An end-to-end supervised target-word sense disambiguation system. In *Proceedings of AAAI*.

G. Kikui. 1999. Resolving translation ambiguity using non-parallel bilingual corpora. In *Proceedings of the Workshop On Unsupervised Learning In Natural Language Processing*.

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

Y.K. Lee, H.T. Ng, and T.K. Chia. 2004. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Proceedings of Senseval-3*, pages 137–140.

E. Lefever and V. Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the Workshop on Semantic Evaluation*, pages 15–20.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC*, pages 24–26.

C. Li and H. Li. 2002. Word translation disambiguation using bilingual bootstrapping. In *Proceedings of ACL*, pages 343–351.

Z. Li, C. Callison-burch, C. Dyer, J. Ganitkevitch, S. Khudanpur, L. Schwartz, W. N. G. Thornton, J. Weese, and O. F. Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*.

R. Mihaelcea. 2007. Using Wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL*, pages 196–203.

R. Mihalcea and C. Corley. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of AAAI*, pages 775–780.

R. Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of EMNLP*, pages 411–418.

R. Navigli and M. Lapata. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of IJCAI*, pages 1683–1688.

R. Navigli. 2009. Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.

T. Pedersen, S. Pathwardan, and J. Michelizzi. 2004. WordNet::Similarity - measuring the relatedness of concepts. In *Proceedings of AAAI*, pages 1024–1025.

S.P. Ponzetto and M. Strube. 2007. Knowledge derived from Wikipedia for computing semantic relatedness. *JAIR*, 30:181–212.

C. Silberer and S. P. Ponzetto. 2010. UHD: Cross-lingual Word Sense Disambiguation using multilingual co-occurrence graphs. In *Proceedings of SemEval-2010*, pages 134–137.

M. Yang and K. Kirchhoff. 2010. Contextual modeling for meeting translation using unsupervised word sense disambiguation. In *Proceedings of Coling*, pages 1227–1235.

T. Zesch, C. Müller, and Iryna Gurevych. 2008a. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of LREC*, pages 1646–1652.

T. Zesch, C. Müller, and Iryna Gurevych. 2008b. Using Wiktionary for computing semantic relatedness. In *Proceedings of AAAI*, pages 861–866.