Toward Integrating Word Sense and Entity Disambiguation into Statistical Machine Translation

Marine CARPUAT Yihai SHEN Xiaofeng YU Dekai WU

Human Language Technology Center
HKUST
Department of Computer Science, University of Science and Technology, Hong Kong
{marine,shenyh,xfyu,dekai}@cs.ust.hk

Abstract

We describe a machine translation approach being designed at HKUST to integrate semantic processing into statistical machine translation, beginning with entity and word sense disambiguation. We show how integrating the semantic modules consistently improves translation quality across several data sets. We report results on five different IWSLT 2006 speech translation tasks, representing HKUST's first participation in the IWSLT spoken language translation evaluation campaign. We translated both read and spontaneous speech transcriptions from Chinese to English, achieving reasonable performance despite the fact that our system is essentially text-based and therefore not designed and tuned to tackle the challenges of speech translation. We also find that the system achieves reasonable results on a wide range of languages, by evaluating on read speech transcriptions from Arabic, Italian, and Japanese into English.

1. Introduction

The role and usefulness of semantic processing for Statistical Machine Translation (SMT) has recently been much debated. In previous work, we reported surprisingly disappointing results when using the predictions of a Senseval word sense disambiguation (WSD) system in conjunction with SMT using an IBM-style model (Carpuat and Wu, 2005b). Nevertheless, error analysis leaves little doubt that the performance of SMT systems still suffers from inaccurate lexical choice. Other empirical studies have shown that SMT systems perform much more poorly than dedicated WSD models, both supervised and unsupervised, on Senseval WSD tasks (Carpuat and Wu, 2005a)—also suggesting that WSD still has a role to play in improving SMT.

In this paper, we describe ongoing work on an approach being designed at HKUST to investigate the effect of semantic handling on current SMT models, using dedicated word sense and entity disambiguation modules. In particular, we propose a new architecture for integrating WSD into SMT architectures, and show that this additional semantic handling consistently improves translation quality across several data sets.

We then turn to the IWSLT 2006 tasks, describing the experimental set-up and evaluation results. This represents a first participation by HKUST in the IWSLT spoken language translation evaluation campaign. For this first participation, we focused on building a baseline system for Chinese to English translation that could be easily ported to different language pairs. We therefore chose to translate additional input languages from different language families. We submitted translations of read and spontaneous speech in the Chinese to English task, as well as read speech translations from Arabic, Italian and Japanese into English. Despite the fact that the system is essentially text-based, and therefore is not designed and tuned to tackle the challenges of speech translation, the system achieves reasonable performance, yielding a BLEU score of 15.45 and a METEOR score of 44.56 for Chinese to English translation, our main language pair of interest. Results on other language pairs suggest that the system can achieve reasonable results with little modification.

2. Machine translation engine

The core MT engine as used in the experiments here is an off-the-shelf phrase-based statistical machine trans-

*This work was supported in part by DARPA GALE contract HR0011-06-C-0023, and by the Hong Kong Research Grants Council (RGC) research grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09.

lation model. This is a useful engine since the approach has been shown to achieve competitive translation quality and is commonly used. Many state-of-the-art systems employ phrase-based approaches (e.g., Zens *et al.* (2005), Koehn *et al.* (2005), Sadat *et al.* (2005)). All phrase-based models make use of a phrasal bilexicon, but essentially differ in the bilexicon extraction and parameter estimation strategies, and the phrase reordering method.

2.1. Decoder

For the experiments here, we used the Pharaoh decoder (Koehn, 2004), which implements a heuristic beam search for phrase based translation. While the phrase reordering model used in Pharaoh is weaker than in other proposed models, Pharaoh was chosen for the advantages of being freely available and widely used, and therefore constitutes an appropriate point of reference.

2.2. Phrasal bilexicon

The core phrasal bilexicon is derived from the intersection of bidirectional IBM Model 4 alignments, obtained with GIZA++ (Och and Ney, 2002). The intersection is augmented using growing heuristics proposed by Och and Ney (2002) in order to improve recall. Following Koehn (2003), each entry in the phrasal bilexicon is scored using phrase translation conditional probabilities for both translation directions, as well as lexical weights which combine word translation probabilities according to the word alignment observed within the phrase pair during training.

2.3. Language model

The language model is a standard trigram model with Kneser-Ney smoothing trained using the SRI language modeling toolkit (Stolcke, 2002).

3. Word sense disambiguation for translation lexical choice

We now present a new architecture integrating a state-of-the-art WSD model into phrase-based SMT, and show that WSD produces small but consistent gains across several test sets.

3.1. WSD classifiers

The model consists of an ensemble of four voting models combined by majority vote.

The first voting model is a naïve Bayes model, since Yarowsky and Florian (2002) found this model to be the most accurate classifier in a comparative study on a subset of Senseval-2 English lexical sample data.

The second voting model is a maximum entropy model (Jaynes, 1979), since Klein and Manning (2002) found that this model yielded higher accuracy than naïve Bayes in a subsequent comparison of WSD performance.

The third voting model is a boosting model (Freund and Schapire, 1997), since has consistently turned in very competitive scores on related tasks such as named entity classification, as described in Section 4.1.1. We also use the Adaboost.MH algorithm for WSD, just like for NER.

The fourth voting model is a model based on Kernel PCA (Wu *et al.*, 2004). Kernel Principal Component Analysis (*KPCA*) is a nonlinear kernel method for extracting nonlinear principal components from vector sets where, conceptually, the n-dimensional input vectors are nonlinearly mapped from their original space R^n to a high-dimensional feature space F where linear PCA is performed, yielding a transform by which the input vectors can be mapped nonlinearly to a new set of vectors (Schölkopf *et al.*, 1998). WSD can be performed by a Nearest Neighbor Classifier in the high-dimensional KPCA feature space. We have showed that KPCA-based WSD models achieve close accuracies to the best individual WSD models, while having a significantly different bias (Carpuat *et al.*, 2004).

All these classifiers have the ability to handle large numbers of sparse features, many of which may be irrelevant. Moreover, the maximum entropy and boosting models are known to be well suited to handling features that are highly interdependent.

3.2. WSD features

The WSD classifier employs much richer features than IBM-style statistical MT systems. The feature set consists of position-sensitive, syntactic, and local collocational features, since these features yielded the best results when combined in a naïve Bayes model on several Senseval-2 lexical sample tasks (Yarowsky and Florian, 2002).

All these WSD models were extensively evaluated on a wide range of monolingual and multilingual lexi-

Table 1: Evaluation results: integrating the WSD translation predictions improves BLEU and NIST scores across 4 different Chinese-English test sets.

Test Set	Experiment	BLEU	NIST
DevTest 1	Baseline	40.76	7.9388
	+ WSD for SMT	41.28	7.9814
DevTest 2	Baseline	39.81	8.1533
	+ WSD for SMT	39.85	8.1753
DevTest 3	Baseline	49.26	9.1172
	+ WSD for SMT	49.81	9.1522
DevTest 4	Baseline	16.13	5.7258
	+ WSD for SMT	16.27	5.7569

Table 2: Translation examples with and without WSD for SMT

	Example 1
Input	让 我 看看 菜单 好吗 ？
Ref.	May I see the menu ?
Baseline	Let me see the menu ?
+ WSD	May I see the menu ?
	Example 2
Input	能 把 我的 座位 指 给 我 吗 ？
Ref.	Would you show me to my seat ?
Baseline	Can you change my seat finger for me ?
+ WSD	Can you direct me to my seat ?

cal sample disambiguation tasks both on Senseval-2 and Senseval-3 data (e.g., Carpuat *et al.* (2004), Wu *et al.* (2004), Su *et al.* (2004)).

3.3. Repurposing the WSD models for SMT

Table 1 shows that our method of integrating a state-of-the-art WSD model into phrase-based SMT produces small but consistent gains across all Chinese-English development test sets. The main difference between this approach and our earlier experiments (Carpuat and Wu, 2005b) lies in the fact that we focus on repurposing the WSD system for SMT. Rather than using a generic Senseval WSD model, both the WSD training and the WSD predictions are integrated into the SMT framework. Specifically:

- Instead of using a Senseval system, we redefine the WSD task to be as close as possible to the translation disambiguation task faced by the SMT system.
- Instead of using predefined senses drawn from manually constructed sense inventories such as HowNet (Dong, 1998), our WSD for SMT system directly disambiguates between all translation candidates seen during SMT training.
- Instead of learning from manually annotated training data, our WSD system is trained on the same corpora as the SMT system.

Thus, in a given SMT input sentence, for every word that was seen in the training data, we have a WSD model and a context-dependent distribution over the possible translation candidates of the word. This distribution is used to augment the baseline bilexicon. With Pharaoh, we use the provided XML markup scheme to specifiy translation candidates and their corresponding probabilities. At decoding time, these externally generated translation candidates are considered as if they were additional bilexicon entries, and are used to build translation hypotheses that compete with other translation hypotheses build from within the traditional SMT phrasal translation lexicon.

Analysis shows that the WSD translation probabilities give better rankings and are more discriminative than the baseline translation probabilities, yielding improved translations as can be seen in Table 2.

4. Named-entity translation

Recognizing, disambiguating, and translating entities is a special case of word sense disambiguation for translation lexical choice, where the words or phrases in question are entities of various sorts. Translating names correctly is particularly important to translation quality and usefulness, but does present some distinct challenges from regular phrase translation. First, the vast majority of names are rare and often never seen in training, and, with the exception of names of well-known persons or other entities, are typically not recorded in lexicons. Second, whether a phrase is a named-entity (NE) depends on context and is therefore ambiguous. Third, names have specific translation patterns. For instance, the translation of a person name usually cannot be inferred from the translation of each of its components.

The first step in handling NE translation consists in identifying NE boundaries and their type. In this system, we are focusing on identifying the PERSON, LOCATION and ORGANIZATION entity types. For the purpose of translation, identifying NE boundaries is not sufficient, since the type of a NE affects the translation

Table 3: IWSLT-06 Training data statistics computed for the 4 language pairs

Training Data Statistics	Chinese-English	Arabic-English	Italian-English	Japanese-English
Number of bisentences	39953	19972	19972	39953
Vocabulary size (input lang)	11178	25152	17917	12535
Vocabulary size (English)	18992	13337	13337	18992

patterns: for instance, many location and person names can typically be transliterated, while some components of organization names should be translated with a standard bilexicon instead.

After identifying NE boundaries and types, a rule-based translation approach based on name gazetteers and transliteration schemes is used to obtain one or more translations for each identified NE.

The decoder integrates the NE translation candidates as additional translation candidates for the NE phrase, using the Pharaoh XML markup scheme for translation input, as for the integration of the WSD predictions.

4.1. Identifying named entities

The named-entity recognition (NER) system is based on a multilingual NER system initially developped for several European languages, and subsequently adapted to Chinese.

4.1.1. NER classifiers

As NER can be framed as a classification task, we use an ensemble of three relatively high performing machine learning classifiers:

Boosting: The main idea behind boosting algorithms is that a set of many weak classifiers can be effectively combined to yield a single strong classifier. Each weak classifier is trained sequentially, increasingly focusing more heavily on the instances that the previous classifiers found difficult to classify. Our system uses AdaBoost.MH (Freund and Schapire, 1997), an n-ary classification variant of the original binary AdaBoost algorithm. As demonstrated by Wu *et al.* (2002) and Carreras *et al.* (2002), boosting can be used to build language independent NER models that perform exceptionally well.

Support Vector Machines: Support Vector Machines (SVMs) have gained a considerable following in recent years (Boser *et al.*, 1992). Sassano and Utsuro (2000) and McNamee and Mayfield (2002) have demonstrated that SVMs show promise when applied to named entity recognition, though performance appears quite sensitive to parameter choices.

Transformation-based learning: Transformation-based learning (TBL) is a rule-based machine learning algorithm that was first introduced by Brill (1995) and used for part-of-speech tagging. The central idea of transformation-based learning is to learn an ordered list of rules which progressively improve upon the current state of the training set. An initial assignment is made based on simple statistics, and then rules are greedily learned to correct the mistakes, until no net improvement can be made. Our system uses the fnTBL toolkit (Ngai and Florian, 2001), which implements several optimizations in rule learning to drastically speed up the time needed for training.

4.1.2. NER features

We use a set of primary features which can be easily obtained across languages, and require little linguistic analysis.

For European languages, features are defined as follows:

- Lexical (words and lemmas) and syntactic (part-of-speech) information within a window of 2 words surrounding the current word
- Prefixes and suffixes of up to a length of 4 characters from the current word
- Capitalization: whether the word starts with a capital letter and/or the entire word is capitalized
- A small set of conjunctions of POS tags and words within a window of 2 words of the current word
- Previous history: the chunk tags (gold standard during training; assigned for evaluation) of the previous two words.
- Gazetteer features: whether the current word is within a NE occuring in a given gazetteer.

For Chinese, the feature set must be adapted to tackle several additional challenges. First, unlike European

Table 4: IWSLT-06 Evaluation test data statistics computed for the correct speech transcriptions (text) and the read speech transcriptions (read)

Test Data Statistics	**Chinese-English**	**Arabic-English**	**Italian-English**	**Japanese-English**
Number of sentences	500	500	500	500
Vocabulary size (text)	1328	1950	1467	1330
Vocabulary size (read)	1361	1890	1552	1383
unknown words (text)	150	727	399	154
unknown words (read)	124	763	340	105

languages, Chinese lacks capitalization information which plays a very important role in identifying named entities. Second, there is no space between words in Chinese, so ambiguous segmentation interacts with NER decisions. Consequently, segmentation errors will affect the NER performance, and vice versa. Third, unlike European languages, Chinese allows an open vocabulary for proper names of persons, eliminating another major source of explicit clues used by European language NER models. Based on these observations, we use character-level features instead of word-level features; this prevents committing to a given word segmentation, which might be incorrect at NE boundaries.

Several versions of this NER system were extensively evaluated on NER shared tasks for Chinese at SIGHAN 2006 (Yu *et al.*, 2006) and for several European languages at CoNLL 2002 (Wu *et al.*, 2002) and 2003 (Wu *et al.*, 2003).

5. IWSLT experimental set-up and results

5.1. Data description

Training and evaluation data are drawn from the multilingual *Basic Travel Expression Corpus (BTEC*)*, which contains relatively short sentences used in simple conversations in the travel domain, and their translations in several languages.

We participated in the open track of the evaluation campaign, where we were allowed to use only the BTEC* data given for each translation task, plus any other external resources. The training and evaluation data statistics are given in Table 3 and 4 respectively. The Chinese-English and Japanese-English tasks were provided with twice as many training bisentences as the Arabic-English and Italian-English tasks. Taking advantage of the fact that BTEC* is a multilingual parallel corpus, all training sets share the same English side. Similarly, the evaluation test sets are composed of Arabic, Chinese, Italian and Japanese sentences that can all translate to the same English sentence.

All training data was clean text, representing a mismatch to the test data used in the evaluation, which was noisy output from automatic speech recognition. In addition to recognition errors, automatic speech transcriptions do not contain punctuation, and use digits to represent numbers. Performance could be improved by eliminating the mismatch between training and test data.

For each Chinese sentence, we are given correct speech transcriptions as well as automatic read speech transcriptions and automatic spontaneous speech transcriptions. For the other languages, we only translated the correct and the read speech transcriptions. For this first IWSLT participation, we did not take advantage of the availability of n-best lists, and only made use of the 1-best transcription, as if the input were text.

5.2. Language-specific data preprocessing

For all language pairs, sentence pairs containing multiple segments are split and re-aligned to provide cleaner parallel training data. After this common processing step, each language followed a minimal language-specific tokenization scheme.

English: The English was simply tokenized and case-normalized in the same manner for all languages.

Chinese: The Chinese side of the parallel corpora was word segmented using the LDC segmenter.

Arabic: In contrast with the 4 other languages considered, Arabic is a morphologically rich language and requires more sophisticated processing. The Arabic text is first converted to the Buckwalter romanization scheme. Tokenization and lemmatization are performed using the ASVMT Arabic morphological analysis toolkit (Diab, 2005). An Arabic word is typically formed of a stem, and possibly affixes and clitics. Affixes are inflectional markers for tense, gender and/or number, while the clitics include some prepositions, conjunctions, determiners, etc. Tokenization, which consists of separating those

Table 5: Evaluation results on the Chinese-English translation task, on correct speech transcriptions (text), read speech transcriptions (read) and spontaneous speech (transcriptions)

Evaluation Metric	**HKUST Result (text)**	**Result Range (text)**	**HKUST Result (read)**	**Result Range (read)**	**HKUST Result (spontaneous)**	**Result Range (spontaneous)**
BLEU	18.04	12.84-24.23	15.45	10.37-21.11	14.41	03.44-18.98
NIST	5.3615	4.0658-6.4004	4.7769	3.6384-5.5858	4.6365	2.7374-5.1513
METEOR	49.15	41.64-51.82	44.56	37.29-45.96	42.38	31.78-41.98
WER	68.99	74.42-65.06	71.16	77.64-69.10	71.87	87.12-70.60
PER	54.87	58.21-49.80	58.20	61.62-54.73	59.11	74.30-57.05

Table 6: Examples of Chinese translations for different input conditions: correct speech transcription (text), read speech transcription (read), and spontaneous speech transcription (spontaneous).

	Example 1
Input (text):	可以 请 把 你 在 日本 的 地址 写 下来 好吗
Output:	Could you please write down the address in Japan, please.
Input (read):	可以 请问 您 在 日本 的 地址 写 下来 好吗
Output:	Could you please write down the address in Japan, please.
Input (spontaneous):	可以 请 办理 给 人 的 地址 写 了 好吗
Output:	May handle, deal with the address of the please.
	Example 2
Input (text):	你 晚上十点 以前 必须 登记
Output:	You must check in by ten o'clock in the evening.
Input (read):	您 玩儿十点 以前 必须 登记
Output:	You must check in at ten before.
Input (spontaneous):	您 晚上十点 以前 必须 登记
Output:	You must check in by ten o'clock in the evening.

syntactic units, is the first step of processing in ASVMT. This is followed by lemmatization which, in ASVMT, refers to a normalization step where the tokens coming from stems that were modified when agglutinated are converted back to their original form.

Italian: We preprocessed the Italian corpus just like the English corpus: it was simply tokenized, using the same rules as for English, and case-normalized. This is obviously not optimal, as Italian presents more morphological inflexions than English, as suggested by the larger vocabulary size on the Italian side of the training data than on the English side (Table 3).

Japanese: We used the provided word segmentation and did not perform any additional processing.

5.3. Chinese-English task

Table 5 shows the evaluation of translation quality for the Chinese-English translation task, using the most common automatic evaluation metrics: BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), as well as word error rate (WER) and position-independent word error rate (PER) (Tillmann *et al.*, 1997). The HKUST system achieves reasonable performance, with evaluation scores situated in the middle range, compared to all systems evaluated on the open track.

As expected, translation quality degrades for all evaluation metrics when moving from correct transcriptions to read and spontaneous speech. Table 6 shows how differences in the accuracy of speech transcription affects the final translation quality. In the first example, the spontaneous speech transcription contains a sequence of four incorrect characters ("办理 给 人" instead of the correct "把 你 在 日本"), which makes the translation meaningless. In contrast, the read speech translation contains only one error: the speech recognizer confuses the more formal word "您" with the correct word

Table 7: Evaluation results on the Arabic, Italian and Japanese translation tasks, for both correct speech transcriptions (text), and read speech trancriptions (read).

Evaluation Metric	**Arabic (text)**	**Arabic (read)**	**Italian (text)**	**Italian (read)**	**Japanese (text)**	**Japanese (read)**
BLEU	16.63	14.77	29.64	23.74	15.60	15.23
NIST	3.8863	3.3318	7.1816	6.0956	0.1560	0.1523
METEOR	42.88	39.20	62.39	54.03	45.79	42.83
WER	67.57	69.16	58.08	63.07	72.48	72.39
PER	56.47	59.48	43.40	49.38	57.86	58.18

Table 8: Translations of test sentences from Arabic (ar), Chinese (zh), Italian (it) and Japanese (jp) into English

Input	**Translation**
Ref.	*It is about twenty kilometers away from here.*
ar	On in about twenty kilometers from here.
zh	About twenty kilometers from here.
it	It's about twenty kilometers far from here.
jp	About two - kilometers from here.
Ref.	*This wine is from France. It's very famous.*
ar	This wine from France and is very popular.
zh	This is very famous French made wine.
it	This wine comes from France is very popular.
jp	This is 's very famous.
Ref.	*Yes. We also have blue, red, yellow and pink.*
ar	Yes, we have a red and my.
zh	Do you have any blue red yellow and pink.
it	Yes, we have red yellow blue and pink.
jp	Yes, we have red green yellow pink.

"你". However, they both translate to the same English word ("*you*") yielding an acceptable sentence translation despite the speech recognizer error. The second set of sentences gives an example of a less common case, where the spontaneous speech translation is better than the read speech translation. The read speech transcription wrongly recognizes the word "晚上" ("*evening*") as "玩儿", which is meaningless and cannot be translated.

5.4. Other language pairs

Translation results for the other language pairs are reported in Table 7. Despite the smaller amount of training data available, translating from Italian yields the best performance, since Italian is closer to English than the three other input languages considered.

Table 8 shows sentence translations obtained for all the input languages for a common reference translation. In these examples, the translation from Italian is usually the best of the four, as shown by the evaluation scores. Japanese translations seem to be the hardest for the system, with many input words that are not or incorrectly translated, despite a phrasal bilexicon learned on twice as much data as the Italian phrasal bilexicon. In Chinese, the phrasal lexicon coverage seems better on these sentences, but our phrase-based model fails to accurately capture differences in syntax: in the third example, the Chinese system translates most words correctly but fails to correctly disambiguate the use of the Chinese verb in assertion vs. interrogation.

6. Conclusion

We have described the design of an approach at HKUST to integrating semantic processing into statistical machine translation, with specific modules for word sense and entity disambiguation and translation, and showed how repurposing the semantic analysis modules for the translation task yields improvements in translation quality. We discussed results obtained on four different languages in the IWSLT 2006 speech translation tasks, in HKUST's first participation in the IWSLT evaluation campaign. On the Chinese to English translation task, the system achieved reasonable performance as measured by a set of automatic evaluation metrics. We also reported results on the Arabic, Italian and Japanese read speech translation tasks, showing that the system is easily portable to other language pairs.

7. References

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgement. In *Pro-*

ceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005.

Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In David Haussler, editor, *Proceedings of the 4th Workshop on Computational Learning Theory*, pages 144–152, San Malteo, CA, 1992. ACM Press.

Eric Brill. Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics*, 4:543–565, 1995.

Marine Carpuat and Dekai Wu. Evaluating the word sense disambiguation performance of statistical machine translation. In *Second International Joint Conference on Natural Language Processing (IJCNLP)*, pages 122–127, Jeju Island, Korea, Oct 2005.

Marine Carpuat and Dekai Wu. Word sense disambiguation vs. statistical machine translation. In *43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 387–394, Ann Arbor, Jun 2005.

Marine Carpuat, Weifeng Su, and Dekai Wu. Augmenting ensemble classification for word sense disambiguation with a Kernel PCA model. In *Third International Workshop on Evaluating Word Sense Disambiguation Systems (Senseval-3)*, Barcelona, Jul 2004. SIGLEX, Association for Computational Linguistics.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. Named entity extraction using AdaBoost. In *Computational Natural Language Learning (CoNLL-2002), at COLING-2002*, pages 171–174, Taipei, Sep 2002.

Mona Diab. Documentation for the Arabic SVM Toolkit. http://www.cs.columbia.edu/mdiab/, 2005.

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

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

Yoram Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Journal of Computer and System Sciences, 55(1)*, pages 119–139, 1997.

Edwin Thompson Jaynes. Where do we stand on maximum entropy? In Raphael D. Levine and Myron Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118. MIT Press, Cambridge, MA, 1979.

Dan Klein and Christopher D. Manning. Conditional structure versus conditional estimation in NLP models. In *Proceedings of EMNLP-2002, Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Philadelphia, July 2002. SIGDAT, Association for Computational Linguistics.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of IWSLT-2005*, Pittsburgh, PA, 2005.

Philipp Koehn. *Noun-Phrase Translation*. PhD thesis, University of Southern California, 2003.

Philipp Koehn. Pharaoh: a beam search decoder forphrase-based statistical machine translation models. In *6th Conference of the Association for Machine Translation in the Americas (AMTA)*, Washington, DC, September 2004.

Paul McNamee and James Mayfield. Entity extraction without language specific resources. In *Proceedings of CoNLL-2002*, pages 183–186, Taipei, Taiwan, 2002.

Grace Ngai and Radu Florian. Transformation-based learning in the fast lane. In *Proceedings of HLT/NAACL-2001*, Pittsburgh, PA, 2001. ACL.

Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, 2002.

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

Fatiha Sadat, Howard Johnson, Akakpo Agagbo, George Foster, Roland Kuhn, Joel Martin, and Aaron Tikuisis. Portage: A phrase-based machine translation system. In *Proceedings of the ACL Worskhop on Building and Using Parallel Texts*, Ann Arbor, MI, June 2005. ACL.

Manabu Sassano and Takehito Utsuro. Named entity chunking techniques in supervised learning for Japanese named entity recognition. In *19th International Conference on Computational Linguistics (COLING-2000)*, pages 705–711, 2000.

Bernhard Schölkopf, Alexander Smola, and Klaus-Rober Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5), 1998.

Andreas Stolcke. SRILM - an extensible language modeling toolkit". In *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, September 2002.

Weifeng Su, Marine Carpuat, and Dekai Wu. Semi-supervised training of a Kernel PCA-based model for word sense disambiguation. In *20th International Conference on Computational Linguistics (COLING-2004)*, Geneva, Switzerland, August 2004.

Christoph Tillmann, Stefan Vogel, Hermann Ney, Alex Zubiaga, and Hassan Sawaf. Accelerated DP-based search for statistical translation. In *Proceedings of Eurospeech'97*, pages 2667–2670, Rhodes, Greece, 1997.

Dekai Wu, Grace Ngai, Marine Carpuat, Jeppe Larsen, and Yongsheng Yang. Boosting for named entity recognition. In *Computational Natural Language Learning (CoNLL-2002), at COLING-2002*, pages 195–198, Taipei, Sep 2002.

Dekai Wu, Grace Ngai, and Marine Carpuat. A stacked, voted, stacked model for named entity recognition. In *Computational Natural Language Learning (CoNLL-2003), at Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL-2003).*, Edmonton, Canada, May 2003.

Dekai Wu, Weifeng Su, and Marine Carpuat. A Kernel PCA method for superior word sense disambiguation. In *42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, Barcelona, Jul 2004.

David Yarowsky and Radu Florian. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310, 2002.

Xiaofeng Yu, Marine Carpuat, and Dekai Wu. Boosting for Chinese named-entity recognition. In *Fifth SIGHAN Workshop of the Special Interest Group for Chinese Language Processing (SIGHAN5) at COLING/ACL 2006*, Sydney, Australia, Jul 2006.

Richard Zens, Oliver Bender, Sasa Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuqi Zhang, and Hermann Ney. The RWTH phrase-based statistical machine translation system. In *Proceedings of IWSLT-2005*, Pittsburgh, PA, 2005.