Communicating Unknown Words in Machine Translation

Matthias Eck, Stephan Vogel, Alex Waibel
interACT Research
Carnegie Mellon University, Pittsburgh, PA, USA
matteck@cs.cmu.edu, stephan.vogel@cs.cmu.edu, waibel@cs.cmu.edu

Abstract

A new approach to handle unknown words in machine translation is presented. The basic idea is to find definitions for the unknown words on the source language side and translate those definitions instead. Only monolingual resources are required, which generally offer a broader coverage than bilingual resources and are available for a large number of languages.
In order to use this in a machine translation system definitions are extracted automatically from online dictionaries and encyclopedias. The translated definition is then inserted and clearly marked in the original hypothesis. This is shown to lead to significant improvements in (subjective) translation quality.

1 Introduction

Every automatic machine translation system faces the problem of unknown words. Currently available corpora, especially for lower resource languages do not cover all possible words in a given language and new words are frequently added. It is for example estimated that the English language has a vocabulary of about 500,000 to 600,000 words with about 25,000 new words introduced each year (Kister, 1992).

The unknown word problem is especially severe for small portable translation systems, as here the vocabulary has to be limited to be able to put the translation system on the portable device.

2 Motivation

At first an apparently small domain like tourism seems to be covered reasonably well by available corpora, e.g. the BTEC corpus (Takezawa et al., 2002) as evaluations do show (e.g. Fordyce, 2007), but this is not a real situation.
Every potential user of such a translation system in the tourism domain will have a variety of interests, hobbies, and needs with potentially thousands of technical terms, that are not covered. Even worse is the situation in other domains like the medical domain, as here specialized technical terms are very common.

The following are two example translations from a medical baseline system translating Spanish → English, where one word each is unknown.

Baseline:	**revelan** you have diabetes.
Reference:	they reveal that you have diabetes.
Baseline:	i am sure you will be getting a great **mejoría** in 4 or 5 weeks
Reference:	i am sure you will feel a great improvement in 4 to 5 weeks

Table 1: Example translations with references

In both cases the rest of the sentence is translated very well, but especially in the second sentence the unknown word contains a lot of information as the patient will not know what he has to expect in 4 or 5 weeks. It is not even clear if it will be a positive or a negative event. The first sentence is relatively understandable, but there is the possibility that the unknown word might negate the actual sentence. A background lexicon can ameliorate this situation, but it will not be possible to have a lexicon covering all words.

Our approach is based on how a non-native speaker would handle the situation if he does not know the translation of a certain word he intends to communicate. The natural reaction will be to use a known synonym or generally describe the unknown word in other words. For example the translation of the word “biweekly” might not be known, but the explanations “twice a week” or “two times a week” contain more common words and should be easier to translate for a non-native speaker.
In order to do this in a machine translation system the first goal is to automatically produce these explanations. Monolingual resources like dictionaries and encyclopedias contain this information and online versions are available. Wikipedia (www.wikipedia.org) specifically is an online encyclopedia in a variety of languages but other websites offer similar information and online dictionaries exist for a number of languages.
The advantages of the monolingual sources are mainly an improved coverage. Monolingual resources contain far more vocabulary than any bilingual resource and their actual intention is to cover every word in a language. The most recent Oxford English Dictionary for example contains 616,500 word forms (Simpson and Weiner, 1989). Broad monolingual resources are also available in languages where parallel bilingual data might be very scarce.
For all translation experiments a standard statistical machine translation system was used but this does not limit the applicability of the proposed approach. Any machine translation system is theoretically able to benefit from the approach.

3 Related Work

Various other approaches to translate or deal with unknown words have been introduced before.
One idea relies on morphological similarities to map the unknown words to actually known words as for example used in Mermer et al. (2007). This can give very good

results, especially if the source language is morphologically richer than the target language. A problem can arise if the inflectional difference indicates a different part of speech. The approach will then produce the translation in an incorrect part of speech in the target language.
This approach is comparable to the proportional analogies in Lepage and Denoual (2005). They used the technique of Analogical Learning that was later also directly applied to the translation of unknown words in Langlais and Patry (2007).
Other groups investigated comparable corpora as a source for unknown word translations (e.g. Fung and Yee, 1998; Takaaki and Matsuo, 1999). In these cases a seed lexicon is used to be able to compare contexts in both languages of the bilingual comparable corpora. If the contexts in which an unknown word pair occurs are often similar in both languages the pair is assumed to be a translation candidate.
Some other approaches mainly target named entities as here often special transliteration rules are applied. This is primarily relevant if the character sets of source and target languages are different. Techniques are for example described in Huang (2005) and Zhao et al. (2007). Generally transliteration models are introduced that learn character and character group transliteration from given examples and can produce a transliteration hypothesis. It can be beneficial to spell-check this hypothesis against a large vocabulary to correct small transliteration errors.
Closer related to the approach presented here are the paraphrasing methods described in Callison-Burch et al. (2006) and Cohn and Lapata (2007). Their idea is to generate paraphrases for unknown words and n-grams from bilingual texts. These paraphrases retain the same meaning but can be translated as they contain other, potentially more common, words as the original phrase. Additional bilingual corpora are necessary, but they can vary in their second language. That means English paraphrases can be extracted from an English-German corpus and can later be applied to an English → Spanish translation task.
It should be pointed out that the presented approach is related to these approaches, but should be seen as an additional step to handle the remaining unknown words after these approaches have been applied. Especially approaches based on morphological similarities have shown very good results, but they can only be used if reasonably close words are in the translation lexicon. The other approaches also have inherent limitations, e.g. the transliteration approaches are only applicable to certain named entities or need additional bilingual or at least comparable corpora.

4 Proposed Process

Figure 1 shows an overview of the proposed process. After a regular translation run the hypothesis still contains unknown words. For these unknown words definitions have to be extracted and translated. Finally the translated definitions are introduced in the original translation hypothesis.

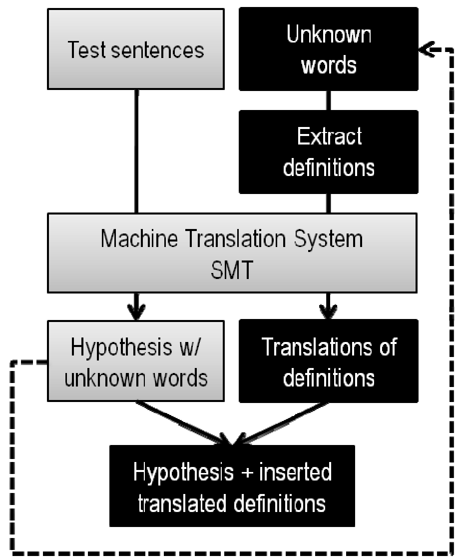

Figure 1: Process Overview

4.1 Extract Definition for Unknown Word

Given an unknown word the first step is to find the definition for this word in the source language. A useful definition of an unknown word should be unambiguous, short and as easy to understand as the original word. It should also not use any words that are unknown to the translation system, as it would again not be possible to translate the definition.

Concerning the source for the definitions we identified two main cases, depending on the type of unknown word (Examples from the later experiments)

Regular content words (that are not named entities):
e.g. Workforce, thrice, ascertain, biweekly, ancestry

Named entities:
e.g. Cairo, Pennsylvania, McDonalds, BMW

For regular content words that are not named entities standard online dictionaries usually provide concise and short explanations.
For our experiments in English we used dictionary.com (http://www.dictionary.com). Dictionary.com merges results from various dictionaries and showed a very good coverage. For Spanish we used wordreference.com (http://www.wordreference.com) (Kellogg, 1999).
Both dictionaries list multiple meanings per word. If necessary a word sense disambiguation component could help to find the best fitting one at this stage but it was shown that rare words are usually less ambiguous than common words (Twilley, 1994).
Figure 2 shows the results from dictionary.com for the search term “biweekly”.

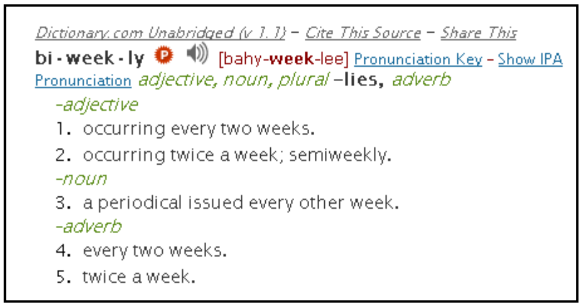
Dictionary.com Unabridged (v 1.1) - Cite This Source - Share This
bi·week·ly [bahy-week-lee] Pronunciation Key - Show IPA Pronunciation adjective, noun, plural -lies, adverb
-adjective
1. occurring every two weeks.
2. occurring twice a week; semiweekly.
-noun
3. a periodical issued every other week.
-adverb
4. every two weeks.
5. twice a week.

Figure 2: Dictionary.com results for the search term "biweekly"

Named entities are a special case as regular dictionaries do not contain many actual named entities or specific brand names. An encyclopedia offers a better coverage here. For our experiments we used automatic extraction from Wikipedia articles. A problem is that Wikipedia articles can be extremely long and it is not useful to translate a long article just to communicate a single unknown word. However we found empirically that the first sentence of the Wikipedia article usually gives a good definition of the term if the term can be clearly defined.

4.2 Translation

The next step is the actual translation of the extracted definition. Major issues here are potentially unknown words occurring in the definition. Especially named entity explanations tend to use further named entities that might also not be known to the translation system.

For other content words the general impression is that the explanations tries to use more common words, which generally helps the translation system. Some dictionaries are specifically designed with this goal in mind, e.g. the Oxford 3000 is a list of 3000 common words that are used in the definitions in the Oxford Advanced Learners Dictionary (with few exceptions) (Wehmeier, 2007).

On rare occasions a word is explained by using it again in a sentence.

Another problem is that the style of the definitions can be very different from the domain of the actual translation system. In our example the translation system was trained on dialogs, a completely different style than the short and concise definitions. It could be valuable to further optimize or specifically design a translation system dedicated to translating this type of data.

4.3 Insert translated definition

To finally produce the improved translation we have to introduce the translated definition into the baseline translation. Just replacing the unknown word with the definition is questionable as this might make the sentence unclear and confusing in the target language.

For that reason we propose to clearly mark the definition as such and leave the decision if that definition clearly defines a word or a short phrase to the speaker of the target language. This way we avoid affecting the coherence of the rest of the sentence.

Table 2 shows the two previous example sentences with translated and inserted definitions. Both sentences are clearly improved and are can be easily understood.

Improved hypothesis:	(UNK: **revelan** # undiscovered it secret) you have diabetes
Spanish definition:	revelan: descrubrir lo secreto
Baseline:	**revelan** you have diabetes.
Reference:	they reveal that you have diabetes.
Improved:	i am sure you will be getting a great (UNK: **mejoría** # getting better) in 4 or 5 weeks
Spanish definition:	mejoria: mejora
Baseline:	i am sure you will be getting a great **mejoría** in 4 or 5 weeks
Reference:	i am sure you will feel a great improvement in 4 to 5 weeks

Table 2: Sentences with inserted and translated definitions compared to baseline and reference

5 Experimental Results

5.1 Monolingual Experiments

In our first experiment we wanted to see for how many monolingual English words we could extract a meaningful definition that would nicely explain the unknown word.

For this experiment we chose the 16 reference translations of the IWSLT 2004 test set (Akiba et al., 2004) and determined all unknown words in those references compared to the English Full BTEC corpus (Takezawa et al., 2002).

Overall 236 words out of the references are unseen and definitions for those were extracted automatically from dictionary.com. For this experiment the first definition was always chosen.

One human subject (native English speaker) judged the adequacy of the extracted definitions on a scale of 1(worst) to 5(best) (compare Fordyce, 2007). The subject was asked to judge the adequacy as if the definitions were translations. The subject was also asked to assume the most common meaning for each word.

Figure 3 shows the distribution of the different adequacy scores.

For 46 words no definition could be extracted and the lowest score of 1 was assigned. These words include typographical errors, exclamations ("Yum") and certain slang terms. 9 other words also received the lowest score, mainly due to definitions for unusual word meanings.

88 words overall received a score of 2 or 3. This was mainly due to incorrect conjugations or referring to an incorrect part of speech e.g.:

- summoning: To call together; to convene
- locations: The act or process of locating

Most online dictionaries automatically project inflected word forms to the base form, which leads to these issues, but we expect these problems to be less severe than a score of 2 or 3 might suggest. 8 definitions received an adequacy score of 4, and 85 times the best score of 5 was assigned.

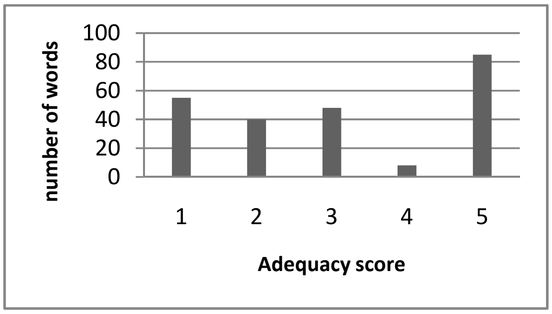

Figure 3: Monolingual adequacy score overview

These results also indicate that the potentially different meanings are not a very significant problem as only 9 times a definition was judged incorrect because of an unusual meaning.
The overall average score was 3.12.

5.2 Bilingual Experiments

The actual question is however if it is possible to extract these definitions in a source language and translate the definitions to a target language while conserving the complete meaning of the extracted definition. This was investigated in this bilingual experiment.

Experimental Setup

For training data we used the English-Spanish BTEC corpus. As mentioned the BTEC corpus contains dialog style tourism phrases.
Overall 123,416 lines of bilingual data were available. A 500 line test set consisting of medical dialogs was used to test the approach translating from Spanish to English.
The translation system used for the baseline translations and also the translation of the extracted definitions is a standard statistical machine translation system using an online phrase extraction method and a 6-gram language model trained on the English part of the bilingual training corpus (Vogel, 2003; Vogel, 2005; Eck et al., 2006).

Extracting definitions

This test sets contains 289 unknown words in Spanish for which we tried to extract Spanish definitions from wordreference.com.
For the first experiment the first definition was chosen as it is most likely the most common meaning. In a second experiment the definition with the lowest number of unknown words was chosen. The argument for this is simply as pointed out in section 4.2 that the definition has to be translated.
For 86 words no definition could be extracted. As in the previous experiment these words are mainly typos, named entities and brand names that are not available in wordreference.com.
For the remaining 203 words definitions were extracted.
Those definitions contained on average 2.50 unknown words if the first definition was extracted and 1.71 unknown words if the definition with the lowest number of unknown words was chosen. Table 3 compares how often definitions with 0 to 2 unknown words could be extracted in both approaches.

# Unknown words	First definition	Lowest number of unknown words
0	33x	57x
1	46x	61x
2	49x	46x
>2	75x	39x
Average/definition	2.50	1.71

Table 3: Unknown words in extracted definitions

The definitions were again subjectively judged for adequacy according to the scale in Table 4. Here the translations were inserted in the respective hypothesis sentences as described in section 4.3.

1	Worse than unknown word, misleading
2	No change compared to unknown word
3	Clear improvement
4	Good translation
5	Perfect translation

Table 4: Adequacy judgments for bilingual experiments

A score of 1 was assigned if the translation became actually misleading and was clearly worse than the unknown word. This means the sentence had to make reasonable sense, but was also misleading. A score of 2 was assigned if the inserted definition did not give any benefit over the unknown word. This also means that the baseline score with all unknown words would be a score of 2. Scores 3 to 5 were assigned for improvements over the unknown word. Figure 4 illustrates the adequacy results.

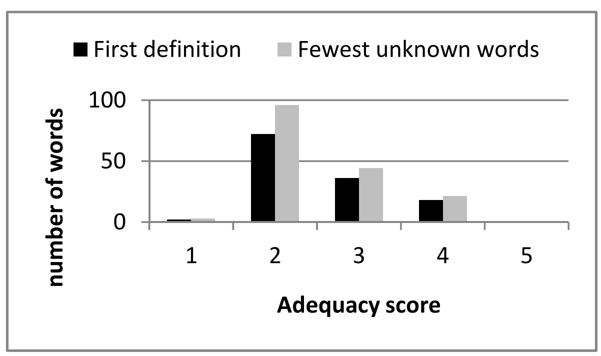

Figure 4: Bilingual Adequacy score overview

Generally if more than 2 unknown words are in the definition the translations were useless and received a score of 2, which leaves 128 first definitions and 164 definitions with the fewest number of unknown words. The average adequacy score for those definitions is 2.55 for the first definitions and 2.51 for the definitions with the fewest number of unknown words, which is only slightly lower than the score for the first definition but it produced definitions for more words. A score of 1 was only assigned two times and three times respectively while a score of 5 (perfect translation) was never assigned. If only the definitions with 0 unknown words are considered the averages are 2.96 for the first definition (33 instances) and 2.67 (57 instances) for the definition with the fewest unknown words.

5.3 Translation Examples

Table 5 shows some examples for translated definitions with 0 unknown words in the first definition (so it was also the definition with the fewest unknown words). Most examples show clear improvements.

Spanish word:	innecesario
Translated definition:	is not it necessary
Reference:	unnecessary
Spanish word:	repetitivos
Translated definition:	to repeat
Reference:	repeat
Spanish word:	acidos
Translated definition:	you have taste sour
Reference:	acidic
Spanish word:	energías
Translated definition:	power be able
Reference:	energy
Spanish word:	intranquilas
Translated definition:	eager nervous
Reference:	stressful

Table 5: Example translations with 0 unknown words in the definitions

Table 6 compares the translations of the first definition with the translations of the definition with the fewest unknown words.

Spanish word:	brota
First definition:	quite UNK the floor the UNK
Fewest unknown:	get out to the surface disease
Reference:	outbreak (disease)
Spanish word:	dolían
First definition:	quite UNK pain in a part of the body
Fewest unknown:	causing consumers' pain
Reference:	sore
Spanish word:	asquerosa
First definition:	4x UNK
Fewest unknown:	disgusting to have UNK
Reference:	disgusting
Spanish word:	radiante
First definition:	UNK bright
Fewest unknown:	very happy, or satisfied for something
Reference:	radiating (pain)

Table 6: Comparison - Translations of the first definition and the definition with the fewest unknown words

This clearly illustrates the correlation between the number of unknown words and the quality of the translation. The last example shows one of the instances where the translated definition was judged worse than the unknown word. “radiante” is translated correctly but in this context describing “radiating pain” the incorrect and misleading meaning was chosen.

5.4 Extracting definitions from Wikipedia

To extract definitions for named entities we used Wikipedia in preliminary experiments. Bilingual experiments were not done yet.

However it is reasonable to assume that similar results as before can be achieved if concise definitions for the unknown words are available. This is also the main issue in Wikipedia as articles tend to be very long and not concise at all. As mentioned before we found empirically that the first sentence of an article tends to give a good definition if a concise definition is possible.

Table 7 shows some examples. It is clear that the dialog partner has to have additional world knowledge to understand what is being defined to get to the actual term. However there could also be situations where the more general term e.g. “city” or “city in Egypt” for “Cairo” could be better than not having any translations.

The last example shows an instance where the definition is not unambiguous (Bolivia is not the only landlocked country in South America).

Unknown word	First Wikipedia sentence
Cairo	is the capital city of Egypt.
Kilimanjaro	is an inactive stratovolcano in north-eastern Tanzania.
Tempura	is a classic Japanese dish of deep fried lightly-battered vegetables or seafood.
Nile	is a major north-flowing river in Africa, generally regarded as the longest river in the world.
Bolivia	is a landlocked country in South America.

Table 7: Example definitions extracted from Wikipedia

6 Conclusions & Future Work

The experiments show that the proposed approach can give considerable improvements in communicating unknown words. The main limiting issues are remaining unknown words in the extracted definitions and the projection of inflected words to a base form, which can lead to differences concerning the part of speech.

It could be shown that selecting a the definition with the lowest number of unknown words can improve this situation while the translation quality still improves.

It might be valuable as noted before to develop specialized translation systems to translate the definitions as the domain mismatch in the experiments clearly influenced the translations. Further experiments with definitions for named entities extracted from Wikipedia articles will be necessary. It might also be valuable to investigate summarization approaches to improve the extraction of concise and unambiguous definitions from the long articles.

The question how this can be included in a complete speech to speech translation system remains as well. It will most likely be necessary to type the unknown word in as it cannot be assumed that it is part of the speech recognition vocabulary. At that point the user could also be asked to select the most fitting definition from a number of presented options.

7 Acknowledgements

This work is in part supported by grants from the US DARPA (TransTac project) and the National Science Foundation (STR-DUST project).

8 References

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Junichi Tsujii (2004). Overview of the IWSLT 2004 Evaluation Campaign. *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2004)*, Kyoto, Japan.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne (2006). Improved statistical machine translation using paraphrases. *In Proceedings of the Human Language Technology Conference and the Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*, New York City, NY, USA.

Trevor Cohn and Mirella Lapata (2007). Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic.

Matthias Eck, Ian Lane, Nguyen Bach, Sanjika Hewavitharana, Muntsin Kolss, Bing Zhao, Almut Silja Hildebrand, Stephan Vogel, and Alex Waibel (2006). The UKA/CMU Statistical Machine Translation System for IWSLT 2006. *In Proceedings of International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, Japan.

Cameron Shaw Fordyce (2007), Overview of the IWSLT 2007 Evaluation Campaign. *In Proceedings of International Workshop on Spoken Language Translation (IWSLT 2007)*, Trento, Italy.

Pascale Fung and Lo Yuen Yee (1998). An IR approach for translating new words from non parallel, comparable texts. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 1998)*, San Francisco, CA, USA.

Fei Huang. Cluster-specific Name Transliteration (2005). *In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005)*, Vancouver, BC, Canada.

Michael Kellogg (1999), Wordreference.com, http://www.wordreference.com

Ken Kister. Dictionaries defined. *Library Journal*, 6/15/92, Vol. 117 Issue 11, p43, 4p, 2bw

Philippe Langlais and Alexandre Patry (2007). Translating Unknown Words by Analogical Learning. *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, Czech Republic.

Yves Lepage and Etienne Denoual (2005). ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages. *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005)*, Pittsburgh, PA, USA.

Lexico Publishing Group, LLC, Dictionary.com, http://dictionary.reference.com.

Coskun Mermer, Hamza Kaya, and Mehmet Ugur Dogan (2007). The TÜBITAK-UEKAE Statistical Machine Translation System for IWSLT 2007. *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2007)*, Trento, Italy.

John Simpson and Edmund Weiner (Editors) (1989) Oxford English Dictionary. second edition, Clarendon Press, 1989, twenty volumes, hardcover, ISBN 0-19-861186-2.

Tanaka Takaaki and Yoshihiro Matsuo. Extraction of translation equivalents from non-parallel corpora. *In Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1999)*, Chester, England.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.

Leslie C. Twilley, Peter Dixon, Dean Taylor, and Karen Clark (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory&Cognition*, 22(1):111-126.

Stephan Vogel (2003). SMT Decoder Dissected: Word Reordering. *In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2003)*, Beijing, China.

Stephan Vogel (2005). PESA: Phrase Pair Extraction as Sentence Splitting. *In Proceedings of MTSummit X*, Phuket, Thailand.

Sally Wehmeier (editor) (2007). Oxford Advanced Learners Dictionary. Oxford University Press, 7th edition.

Bing Zhao, Nguyen Bach, Ian Lane, and Stephan Vogel (2007). A Log-linear Block Transliteration Model based on Bi-Stream HMMs. *In Proceedings of the Human Language Technology Conference and the Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007)*, Rochester, NY, USA.