Virtual Babel: Towards Context-Aware Machine Translation in Virtual Worlds

Ying Zhang and Nguyen Bach
Carnegie Mellon University
23 S. Akron Rd.
Moffett Field, CA
{joy+,nbach+}@cs.cmu.edu

Abstract

In this paper, we describe our ongoing research project of Virtual Babel, a context-aware machine translation system for Second Life, one of the most popular virtual worlds. We augment the Second Life viewer to intercept the incoming/outgoing chat messages and reroute the message to a statistical machine translation server. The returned translations are appended to the original text message to help users to understand the foreign language. Virtual Babel provides a platform to study cross-lingual conversations facilitated by machine translation in virtual worlds and we observe interesting phenomena that are not present in document translations. Virtual Babel is aware of the non-verbal context of the conversation. Language model and translation models are trained from collected conversations and are used to generate translations according to observed non-verbal context of the conversation.

1 Introduction

Virtual world is fast becoming a favorite venue for online learning, collaborating, and networking. Just as the real world, users in the virtual world come from different background and speak different languages. Even if some users know more than one languages, it is still more natural for them to use their native languages to communicate. Thus, the language barrier hinders the communication in the virtual world just as it does in the real world.

In recent years, machine translation (MT) technologies have been greatly improved. In particular, phrase-based statistical machine translation systems (Och et al., 1999; Koehn et al., 2003) have advanced to a state that the translation quality for certain domains (e.g. broadcasting news) and certain language pairs (e.g. Spanish-English) is acceptable to users. The publicly available translation services such as the Google Translation API make it possible to develop translation plug-ins for various applications including Skype, MSN and Google Talk.

These translation services create great interests in the community of using machine translation to bridge the language gaps in cyber communication. These general purpose machine translation services are usually trained from bilingual text like broadcasting news and parliamentary debate data which is of different genre compared to the online chat. Such mismatch in genre usually results in sub optimal translation quality. Another problem of using generic translation services is that these services are context independent. In other words, a sentence has the same translation no matter where it occurs. As shown in the following sections, ambiguities in natural language can only be resolved with the context.

In this paper, we describe the Virtual Babel project, an ongoing research effort of developing a context-aware machine translation system for virtual worlds. On one hand, machine translation can greatly help the multilingual communication inside the virtual world due to the non-critical nature of online chats. On the other hand, virtual worlds allow MT systems to explore the non-verbal context of the conversation in a much easier way than in real world. This makes it feasible for us to study how context influences the language usage and how MT system

could make use of context information to improve the translation quality.

2 Virtual Worlds

Virtual worlds, such as Second Life (SL), World of Warcraft and Kaneva are computer-based environments where real-life users inhabit and interact via avatars. Virtual worlds are close-to-real simulation of the real world and users can experience telepresence to a certain degree. Users can use both instant text message (IM) and real-time voice chat to communicate with other users in the 3D environment. Users can participate many virtual activities in virtual worlds including sight-seeing, talk to other people, dancing, listen to live music and attend live concert, attend lectures in open university, building and creating things, doing business, shopping and role-playing games.

We are particularly interested in providing translation services for educational activities in virtual worlds. Virtual worlds provide an alternative, more engaging platform for education in cyber-space. A 3D virtual world provides students with a supportive community. Feeling part of a community of learners has a direct impact, not only on retention, but also on students' perception of successful university experiences (Wellman and Kahne, 1993; Wehlage et al., 1998).

Many educators have discovered the unique possibilities offered by utilizing these virtual worlds to develop new forms of education. A report funded by the Eduserv Foundation estimates that some three-quarters of UK universities are actively developing or using SL, at the institutional, departmental and/or individual academic level[1]. Harvard University, Texas State University, and Stanford University have set up virtual campuses where students can meet, attend classes, and create content together. Using virtual world applications, students around the globe can "sit" in the same classroom without the need to build physical campuses thousands of miles away.

According to the Key Metrics report published by Second life on June 11, 2007 [2], there are 507,844 active users coming from 100 countries and regions. On average each active user spends 36 hours in the virtual world. Table 1 shows the 20 countries with the most active users in Second Life. Though the statistics shown here include all activities in SL, we assume that the educational activities have a similar distribution among users' from different parts of the world. The need of a universal translation system is obvious given the fact that not all users speak English and even they do, a user would prefer to use his/her native language.

3 Virtual Babel Translation Service

Just as in real life, users are most comfortable speaking their native languages. The demand for automatic translation in virtual worlds is as strong as in the real world. Several text translation tools have been developed based on existing online translation services such as Babel Fish and Google Language Translator. Users type in messages in their own languages and the machine translation results in the foreign language are overlayed in the chat display.

Translation research continues to make constant progress. This progress is driven by projects like GALE and TransTac or the recently completed TC-STAR in Europe, but also by open, competitive MT evaluations organized by NIST, by the C-Star consortium (IWSLT evaluation campaign), or in connection with the Workshop on Machine Translation. With automatic MT evaluation metrics (i.e., BLEU (Papineni et al., 2001), TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005)) and a well-defined significance testing method (Zhang et al., 2004), results have become more meaningful, supporting both the day-to-day research in individual research groups and the comparison of techniques and ideas across them.

Crucial is also the availably of open source toolkits, such as GIZA++ (Al-Onaizan et al., 1999; Och, 2003) and more recently mGIZA++ (Bach et al., 2008), for training word alignment models, the SRI LM toolkit (Stolcke, 2002) and the Suffix Array LM toolkit (Zhang, 2006) for building language models, or the Moses package (Koehn et al., 2007) which provides phrase pair extraction scripts and a widely used decoder.

These elements have recently helped make

[1] http://www.eduserv.org.uk/foundation/studies/slsnapshots

[2] http://blog.secondlife.com/2007/06/12/may-2007-key-metrics-published

Country	Active Avatars	% of Avatar	Total hours spent	Hours per avatar
United States	130033	25.60%	6358494	48.9
Germany	59610	11.74%	2187171	36.7
France	39727	7.82%	1362788	34.3
United Kingdom	29831	5.87%	1211925	40.6
Not specified	26357	5.19%	515524	19.6
Spain	25819	5.08%	677353	26.2
Italy	24690	4.86%	673627	27.3
Brazil	24470	4.82%	530939	21.7
Japan	18778	3.70%	574893	30.6
Netherlands	17130	3.37%	890727	52.0
Canada	12234	2.41%	625776	51.1
Australia	9779	1.93%	359919	36.8
Portugal	7655	1.51%	118662	15.5
Belgium	6330	1.25%	221821	35.0
Switzerland	5801	1.14%	188008	32.4
Sweden	5187	1.02%	138801	26.8
Denmark	4622	0.91%	183789	39.8
Mexico	4346	0.86%	65221	15.0
Argentina	4108	0.81%	85112	20.7

Table 1: Top 20 countries with most active avatars and total/average hours users spent in SecondLife (data through May 2007).

statistical machine translation a quickly maturing/developing field, and promise to yield substantial improvements and impact in the near future. This growth in MT research is not only quantitative; it has also led to diversity. Within the dominant paradigm of data-driven statistical MT, we currently observe a wide variety of ideas being explored. String-to-tree, tree-to-tree, and tree-to-string alignment models capture syntactic relations and divergences between languages. Language models based on dependency grammars, or continuous LMs are explored. A variety of decoding algorithms are being investigated, as is the interaction and better combination of different components. The increasing number of different MT systems has also triggered new interest in system combination.

These and similar efforts have resulted in better understanding of the translation process and in better-performing translation systems, leading to a platform on which research into more difficult tasks, such as automatic translation of spontaneous multiparty speech, may be based.

Figure 1 shows the system architecture of the Virtual Babel system. Users connect to Second Life server through a client software referred to as "viewer". The viewer client has been modified such that text message from and into user A's client is redirected to the translation server together with the context information including user's ID, location and local time. Based on the map provided by SecondLife, we can derive point-of-interests given user's current location. Figure 2 shows an example of the modified SL viewer with integrated translation function.

The translation server is based on the PanDoRA statistical machine translation system (Zhang and Vogel, 2007). It is augmented to be context-aware such that the decoder translates a sentence conditioned on the context information.

The translation results are sent back to user A's viewer to be displayed on his/her screen. The original message and the translated message is later sent from A's viewer to the Second Life server. User B receives the message in A's native language and translated form on his/her viewer. The same process is repeated when B sends message to A. The

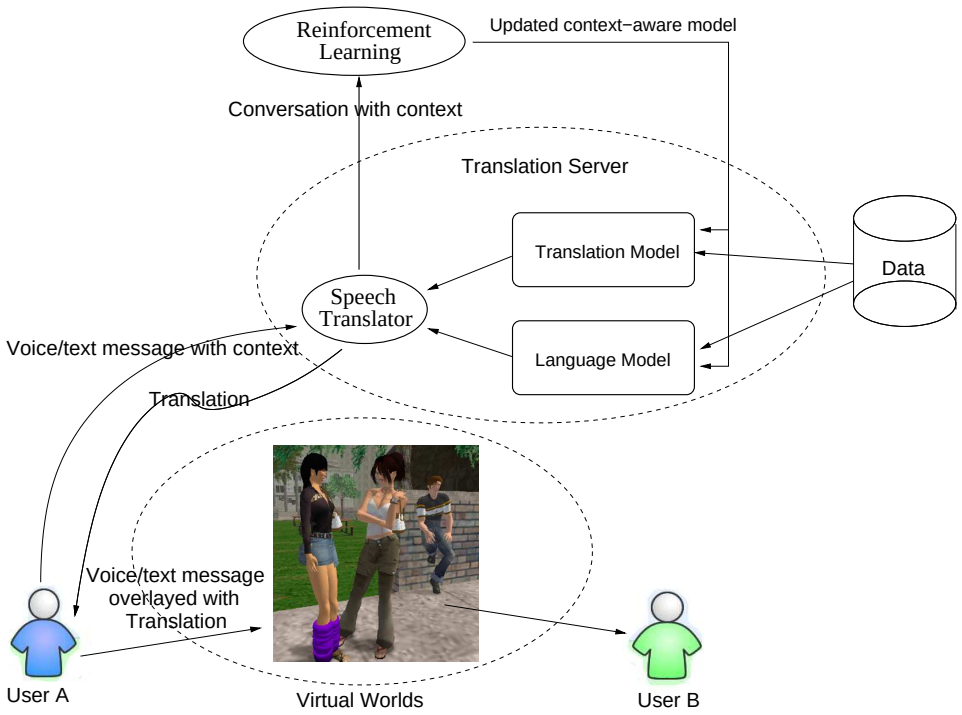

Figure 1: Context-aware machine translation for virtual worlds.

modified viewer client can be downloaded at `http://mlt.sv.cmu.edu/secondlife`

The initial translation system is trained on bilingual data from the travel domain which is closest to the conversation domain in Second Life. Multiple instances of translation decoder run on a small computer cluster to provide real-time translation services to users. The conversation data associated with the context information is logged on the server for reinforcement learning.

4 Context-aware Machine Translation

Natural language is full of ambiguities. Almost all words, phrases and sentences in natural language can be interpreted in more than one ways. Ambiguities in conversation occur more frequently than in the written text as people tend to skip certain information to make the conversation more efficient. People assume that the other party has the same background knowledge and can choose the right interpretation based on the context. Without the right context information, it is difficult for human beings to get the right meaning.

The surface form of the language does not convey all the information. We have to rely on non-verbal contextual information and background knowledge to reason and search for the right underlining meaning of the sentence. For example, we can't tell if the speaker is looking for a river bank or a financial bank from sentence "how can I get to the bank?" itself. We can figure out the actual meaning if we know he is walking on street fully dressed or driving a Jeep with a canoe on the roof rack.

There are three levels of context information in a conversation: domain, topic and 5w context. *Domains* are broad range context such as travel, politics, sports, finance etc. For each domain, speech conversations can be classified by their *topics*. For example, a travel domain conversation can be in topics such as hotel reservation, travel arrangement, ask for direction etc. *Five Ws* are more fine-grained context information regarding *who* the users are, *what* are users doing, *where* they are located, *when* does the conversation happen and *why* users talk. What the user has just talked about gives verbal context to what he is talking now. Non-verbal context such as when and where the conversation occurs is also needed for correct meaning understanding.

Figure 2: Screen shot of the modified Second Life viewer where conversations

Similar to human beings, statistical machine translation (SMT) systems (Brown et al., 1990; Och et al., 1999) also need to know the context to translate sentences correctly. Some verbal context can be encapsulated in phrases. Translating the phrase as one unit thus can resolve some ambiguities of words inside the phrase. The standard n-gram language model, a crucial component in statistical machine translation system, estimates the probability of generating the next word given the history/context of already translated words. In a sense, the phrase-based SMT systems improve greatly over word-based systems because they use phrases to encapsulate the verbal context.

However, all current SMT systems do not model non-verbal context. This is mainly due to the unavailability non-verbal context information and lack of training data that is labeled with context information. The impact on translation quality is obvious. As an example, we use the state-of-the-art Google statistical machine translation system[3] to translate the Chinese testing data from the evaluation campaign of the 2007 International Workshop on Spoken Language Translation (IWSLT 2007). IWSLT testing data is in the travel domain where travelers ask for directions, order food, make hotel reservations, visit doctors for medical problems etc. Table2 lists examples where translations from Google MT are wrong and could have been better translated if the context is known. In example 1, if we know the conversation happens at the ticketing booth of an opera house, the system should choose "tickets" instead of "votes" for the ambiguous Chinese word 票. Similarly for example 2, knowing that ordering drink in a restaurant is the context, the translation system can fill in the missing information "wine" to make the translation more comprehensible. Example 3 needs the context to disambiguate word 点 and also fill in the missing information of what the speaker is referring to depends on if this conversation happens at an airport, a train station, a bus station or a ferry.

State-of-the-art statistical machine translation systems are trained from bilingual corpus. We refer the unseen testing data that is to be translated as the testing data following the statistical machine translation terminology of training-tuning-testing developing cycle.

MT systems work best if the sentences to be translated are from the same domain/genre/style as the training data. When the testing data is from a different domain as the training data, the translation

[3] http://www.google.com/language_tools

Example 1	Chinese sentence:	在这儿能买到歌剧的**票**/tickets, admission, votes/ 吗？
	Reference translation:	Can I buy **tickets** for the opera here?
	MT output:	Here's opera can buy **votes**?
Example 2	Chinese sentence:	有**红**/red/ 的吗？
	Reference translation :	Do you have red wine?
	MT output:	There are **red** it?
Example 3	Chinese sentence:	下一班是几 **点**/points, hours, dots/ ？
	Reference translation:	When is the next one (flight/train/bus/ferry) ?
	MT output:	The next is a few **points**?

Table 2: Examples of incorrect machine translations generated by Google's online MT system.

quality suffers dramatically. In Virtual Babey system we explore explicit context information to build a context-aware translation system.

Denote explicit context information as a feature vector $\vec{C}$, where C_i is the value of a context feature function such as location, time and user ID. A context feature function f_i maps the context into a numerical value, for example,

$$f_i = \begin{cases} 1 & \text{if user is male} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We use a context-free translation system to bootstrap the crosslingual conversation and retrain the LM and TM to adjust their probabilities for specific context.

4.1 Context-aware Language Model

Language model plays an important role in speech translation. Both speech recognizer and the translation system use the language model to select the hypothesis with the highest probabilities. Language model estimates how likely a sequence of words is a correct sentence given the training corpus. The probability distribution in a trained language model reflects the language usage in the domain of the training data. We tend to speak about certain topics in certain domains. For example, we are more likely to say "I would like a number 2 meal" or "a large cheese burger and a medium fries with a coke" at a McDonald rather than at a hospital. Language model probabilities can be better estimated if the context is known. Domain-specific language models have been studied in speech recognition and shown minor improvements over the generic language model. Previous work of domain-adaptive language model predicts the domain of the testing data by looking inside the testing data itself. In other words, these approaches adjusts the langauge model probabilities by implicitly induce the domain from the verbal context. For example, if word "university" occurs a lot in previous recognition results, then the topic is very likely to be about education and thus the probabilities of "students", "professors" should be increased.

In this work, we are provided with explicit context, in particular, non-verbal context information. To integrate this knowledge source, we use the maximum entropy language model framework . A maximum entropy language model considers different knowledge sources as constrains, it chooses the probability distribution that satisfy all these constraints and has the highest entropy (Equation 2).

$$P(\mathbf{e}) = \frac{\exp \sum_{\mathbf{i}} \lambda_{\mathbf{i}} \mathbf{f_i}(\mathbf{e})}{\mathbf{Z}}. \quad (2)$$

In equation 2 Z is a normalization factor in order to set the value of $P(\mathbf{e})$ in the proper range between 0 and 1 and λ_is are weights for feature functions f_i. As shown in (Rosenfeld, 1994), conventional n-gram history, self-trigger, class-triggers, long distance n-grams can all be converted into feature functions and integrated into the maximum entropy language model.

The context features can be naturally integrated into the maximum entropy language model framework as additional knowledge sources. Instead of using equation 2 to estimate the probability of a sentence $\mathbf{e}$ where all information is from the sentence itself, we estimate the probability of $\mathbf{e}$ given the known context vector $\vec{C}$ as:

$$P(\mathbf{e}|\vec{\mathbf{C}}) = \frac{\exp \sum_{\mathbf{i}} \lambda_{\mathbf{i}} \mathbf{f_i}(\mathbf{e}, \vec{\mathbf{C}})}{\mathbf{Z}}. \quad (3)$$

where feature functions $f_i(\mathbf{e}, \vec{\mathbf{C}})$ also takes the explicit context information into accounts.

In this work, users' conversation data is collected together with explicit context information. Training a context-aware maximum entropy language model is a straight-forward task.

4.2 Context-aware Translation Model

Similar to the context-aware language model, we use the log-linear model to integrate the context information into the translation model. Translation model estimates the probability of a source word f given its translation e in the target language $P(f|e)$. The model is usually trained from a sentence-aligned bilingual corpus as introduced by (Brown et al., 1990). To integrate the context information, we condition the translation probability on context $\vec{C}$ and estimate the probability by a log-linear model:

$$P(f|e,\vec{C}) = \frac{\exp(\lambda_0 P(f|e) + \sum_i \lambda_i f_i(f,e,\vec{C}))}{Z}. \quad (4)$$

Notice that context-free translation model probability $P(f|e)$ now becomes one of the features in the context-aware model.

4.3 Context-aware Model Training via Conversation Analysis

In order to train the context-aware translation model and the language model , we need sufficient amount of data that contains context information. To our best knowledge, there is no such data available. Almost all monolingual and bilingual corpus are plain text and have no context information for each sentence. In Virtual Babel, we use a translation system trained on general-domain data to bootstrap the multilingual conversation and adapt the translation/language model through conversation analysis.

Inspired by the emotion detection in intelligent spoken dialogue systems, we believe that conversation analysis can give reliable prediction of the translation quality. Intelligent dialogue systems are widely used for call-center, tutoring and information services where users speak to a computer program to accomplish certain tasks. User emotion reflects the effectiveness of the conversation. It is important to detect user's emotion along the conversation to manage the dialogue. For example, when the speech recognition continuous to fail understanding the user, user could become quite emotional in both language and speech. (Litman and Forbes-Riley, 2004) show that acoustic-prosodic and lexical features can successfully predicting students (users) emotions in a computer-human spoken tutoring system.

The multilingual conversation facilitated by a machine translation system is similar to the computer-human dialogue system. For example, user B may ask "what do you mean?" if the MT output of user A's message causes too much confusion, or "sorry, I don't know what you are saying" when the MT output is totally not understandable. We propose to use lexical features and conversational features to predict the effectiveness of the conversation. Lexical features such as "pardon me" are explicit indicators of negative feedback of the MT quality. Conversational features such as the duration of the conversation, interval between turns are implicit indicators of the conversation effectiveness. Through intelligent conversation analysis we can estimate how well the MT output is and how effective it is to bridge the communication. Using the estimated communication effectiveness, we can adjust the translation model and the language model so that the probabilities of incorrect translations are decreased for this context. Alternative translations will have relatively higher probabilities and may be selected as system output in the next iteration.

Subjective analysis on collected dialogues from Second Life indicate that users indeed provide conversational cues when the translation is wrong. Depends on the seriousness of the translation error, the reaction can range from mild ("can you say it again?") to very strong (avatar walks away and terminate the conversation).

5 Future Work

We are improving the performance of the Virtual Babel system to make the translation faster and more reliable. With more conversation and the associated context information collected from the Virtual Babel system, we can build context-aware language models and apply it in the context-aware translation system. We will compare its performance of the context independent system to justify the need of context-

aware machine translation.

6 Acknowledgement

We would like to thank the reviewers for their comments and suggestions which are very insightful and inspiring.

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, I. Dan Melamed, Franz Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. 1999. Statistical machine translation: Final report for johns hopkins university 1999 summer workshop on language engineering. Technical report, Center for Language and Speech Processing, Baltimore, MD.

Nguyen Bach, Qin Gao, and Stephan Vogel. 2008. Improving word alignment with language model based confidence scores. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 151–154, Columbus, Ohio, June. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edomonton, Canada, May 27-June 1.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Diane J. Litman and Kate Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 351, Morristown, NJ, USA. Association for Computational Linguistics.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

Franz Josef Och. 2003. Minimum classification error training for statistical machine translation. In *ACL 2003: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Ronald Rosenfeld. 1994. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. Ph.D. thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, April. TR CMU-CS-94-138.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings AMTA*, pages 223–231, August.

A. Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver, CO.

G. G. Wehlage, R. A. Rutter, and G. A. Smith. 1998. *Reducing Risk: School as Communities of Support*. Falmer Press.

B. Wellman and J. Kahne, 1993. *Networks in the Global Village*, chapter The Network Basis of Social Support: A network is More than the Sum of its Ties., pages 1–48. Westview Press, Boulder, CO.

Ying Zhang and Stephan Vogel. 2007. Pandora: A large-scale two-way statistical machine translation system for hand-held devices. In *Proceedings of MT Summit XI*, Copenhagen, Denmark, September.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, May. The European Language Resources Association (ELRA).

Ying Zhang. 2006. Suffix array and its applications in empirical natural language processing. Technical Report CMU-LTI-06-010, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Dec.