Using Artificial Data to Compare the Difficulty of Using Statistical Machine Translation in Different Language-Pairs

Manny Rayner, Paula Estrella, Pierrette Bouillon, Sonia Halimi
University of Geneva, TIM/ISSCO
40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
{Emmanuel.Rayner,Pierrette.Bouillon,Sonia.Halimi}@unige.ch
pestrella@gmail.com

Yukie Nakao
LINA, Nantes University, 2, rue de la Houssinière, BP 92208 44322 Nantes Cedex 03
yukie.nakao@univ-nantes.fr

Abstract

Anecdotally, Statistical Machine Translation works much better in some language pairs than others, but methodological problems mean that it is difficult to draw hard conclusions. In particular, it is generally unclear whether translations in parallel training corpora have been produced using equivalent conventions. In this paper, we report on an experiment where a small-vocabulary multilingual interlingua-based translation system was used to generate data to train SMT models for the 12 pairs involving the languages {English, French, Japanese, Arabic}. By construction, the data can be assumed strongly uniform. As expected, translation between English and French in both directions performed much better than translation involving Japanese. Less obviously, translation from English and French to Arabic performed approximately as well as translation between English and French, and translation to Japanese performed better than translation from Japanese.

1 Introduction

Since its introduction in the early 90s, when it was regarded as a dubious outsider, Statistical Machine Translation (SMT) has rapidly gained ground until it is now considered the mainstream approach. There is, however, general agreement that some language-pairs work much better than others. In the positive direction, the early successes reported by the IBM group (Brown et al., 1990) used French-English, which is now known to be an unusually suitable pair (Koehn and Monz, 2005; Koehn and Monz, 2006). Anecdotally, translating between two European languages is easier than translating between a European and a non-European language, and some languages, like Japanese, are widely assumed to be difficult. Given the steadily increasing importance of MT technology, it is often important to be able to make a reasonable guess at how well SMT will work for a new language-pair. Both SMT and RBMT require a large investment of effort before any evaluable system emerges; when planning a project, both architectures are in principle possible, and it is desirable to be able to make an informed choice between them at an early stage.

A recent large-scale study (Birch et al., 2008), using the 110 language-pairs covered by the Europarl corpus, found that the features most predictive of SMT translation quality were target language vocabulary size, lexicostatical relatedness (measured in terms of proportion of cognate words), and similarity in word order. The same study, however, also highlighted the methodological problems inherent in carrying out this type of comparison. As already noted, target vocabulary size turned out to be the most predictive feature. Vocabulary size, however, depends crucially on how morphology is taken into account. For example, (Birch et al., 2008) considered that Swedish had a much larger vocabulary size than English, but this is almost entirely due to the fact that Swedish, like German, writes compound nominals without intervening spaces. The structure of these nominals, however, is often very similar to that of the corresponding English phrases. The problem becomes more acute in a language like Japanese, which is normally written with no word boundaries at all.

Another set of issues arise from the use of parallel human-translated corpora, where it is generally difficult to know whether the data is truly uniform. Quality and style of translation can vary widely, with translators using different guidelines. In particular, some translators will prefer a more literal style than others. It is also common to mix in low-quality data; a frequent choice is translations taken from the reverse language-pair, with the source and target swapped around. Some recent studies (Ozdowska, 2009) have in fact suggested that this kind of low-quality adulteration can do more harm than good. Conversely, other practitioners of SMT have pointed to the performance gains that can be achieved by careful cleaning of the data.

Without controlling for all these factors, it is hard to know how general the results of comparative studies really are. Although (Birch et al., 2008) is an unusually responsible and careful piece of work, the authors point out that removal of the outlier language (Finnish) substantially changes the overall conclusions; it is probably not a coincidence that Finnish was also the only non-Indo-European language used in the study.

In this paper, we present the results of a novel type of comparative study carried out using MedSLT, a small-vocabulary interlingua-based multilingual speech translation system for a medical domain. We generated parallel corpora for all 12 pairs involving the source languages {English, French, Japanese, Arabic}, first using the source language grammars to generate arbitrary amounts of source-language data, then, for each target language, passing it through the relevant translation rules to generate target language expressions. Use of interlingua-based translation enforces a uniform translation style. The small domain, which we completely control, made it possible to enforce uniform decisions about how morphology is treated. For example, we decided in Arabic to split off the definite article *al*, normally affixed to the following noun, and treat it as a separate word. For similar reasons, we also treated Japanese tense and politeness affixes as separate words. Thus a word like *okorimashita* ("happened") is split up as *okori mashita* ("happen PAST-POLITE"). Once we had created the parallel corpora, we trained SMT models, and evaluated the quality of the translations they produced. As expected, translation between English and French in both directions performed much better than translation involving Japanese. We were, however, interested to discover that translation from English and French to Arabic performed as well as translation between English and French, and that translation to Japanese performed better than translation from Japanese.

The rest of the paper is organised as follows. Section 2 provides background on the MedSLT system. Section 3 describes the experimental framework, and Section 4 the results obtained. Section 5 concludes.

2 The MedSLT System

MedSLT (Bouillon et al., 2008) is a medium-vocabulary interlingua-based Open Source speech translation system for doctor-patient medical examination questions, which provides any-language-to-any-language translation capabilities for all languages in the set {English, French, Japanese, Arabic, Catalan}. Both speech recognition and translation are rule-based. Speech recognition runs on the Nuance 8.5 recognition platform, with grammar-based language models built using the Open Source Regulus compiler. As described in (Rayner et al., 2006), each domain-specific language model is extracted from a general resource grammar using corpus-based methods driven by a seed corpus of domain-specific examples. The seed corpus, which typically contains between 500 and 1500 utterances, is then used a second time to add probabilistic weights to the grammar rules; this substantially improves recognition performance (Rayner et al., 2006, §11.5). Performance measures for speech recognition in the three languages where serious evaluations have been carried out are shown in Table 1.

At run-time, the recogniser produces a source-language semantic representation. This is first translated by one set of rules into an interlingual form, and then by a second set into a target language representation. A target-language Regulus grammar, compiled into generation form, turns this into one or more possible surface strings, after which a set of generation preferences picks one out. Finally, the selected string is realised in spoken form. Robustness issues are addressed by means of a back-up

Language	WER	SemER
English	6%	11%
French	8%	10%
Japanese	3%	4%

Table 1: Recognition performance for English, French and Japanese MedSLT recognisers. "WER" = Word Error Rate for source language recogniser, on in-coverage material; "SemER" = semantic error rate (proportion of utterances failing to produce correct interlingua) for source language recogniser, on in-coverage material.

statistical recogniser, which drives a robust embedded help system. The purpose of the help system (Chatzichrisafis et al., 2006) is to guide the user towards supported coverage; it performs approximate matching of output from the statistical recogniser again a library of sentences which have been marked as correctly processed during system development, and then presents the closest matches to the user.

Examples of typical English domain sentences and their translations into French, Arabic and Japanese are shown in Figure 2.

3 Experimental framework

In the literature on language modelling, there is a known technique for bootstrapping a statistical language model (SLM) from a grammar-based language model (GLM). The grammar which forms the basis of the GLM is sampled randomly in order to create an arbitrarily large corpus of examples; these examples are then used as a training corpus to build the SLM (Jurafsky et al., 1995; Jonson, 2005). We adapt this process in a straightforward way to construct an SMT model for a given language pair, using the source language grammar, the source-to-interlingua translation rules, the interlingua-to-target-language rules, and the target language generation grammar. We start in the same way, using the source language grammar to build a randomly generated source language corpus; as shown in (Hockey et al., 2008), it is important to have a probabilistic grammar. We then use the composition of the other components to attempt to translate each source language sentence into a target language equivalent, discarding the examples for which no translation is produced. The result is an aligned bilingual corpus of arbitrary size, which can be used to train an SMT model.

We used this method to generate aligned corpora for 12 MedSLT language pairs with source and target languages taken from the set {English, French, Japanese, Arabic}. For each language pair, we first generated one million source-language utterances; we next filtered them to keep only examples which were full sentences, as opposed to elliptical phrases, and used the translation rules and target-language generators to attempt to translate each sentence. This created between 260K and 310K aligned sentence-pairs for each language-pair. In order to make coverage uniform for each source language, we kept only the pairs for which the source sentence had translations in all target languages. This makes it possible to compare fairly between language-pairs with the same source-language. In contrast, it appears to us that it is less straightforward to compare across language-pairs with different source-languages, since there is no obvious way to ascertain that the two source-language corpora are of comparable difficulty.

The sizes of the final source language corpus for each of the three source languages is shown in Table 3. We randomly held out 2.5% of each of these sets as development data, and 2.5% as test data. Using Giza++, Moses and SRILM (Och and Ney, 2000; Koehn et al., 2007; Stolcke, 2002), we trained SMT models from increasingly large subsets of the training portion, using the development portion in the usual way to optimize parameter values. Finally, we used the resulting models to translate the test portion. We performed the tests with subcorpora of different sizes in order to satisfy ourselves that performance had topped out, and that generation of further training data would not improve performance. Full details are presented in (Rayner et al., 2009).

Language	#Sentences	#Words	Vocab.
Eng	236340	1441263	364
Fre	179758	1205308	557
Ara	233141	1509594	253
Jap	207717	1169106	336

Table 3: Statistics for final auto-generated source language corpora for source languages: number of sentences, number of words, and size of vocabulary

English	Have you had the pain for more than a month?
French	Avez-vous mal depuis plus d'un mois?
Arabic	Hal tahus bi al alam moundhou akthar min chahr wahid?
Japanese	Ikkagetsu ijou itami wa tsuzuki mashita ka?
English	When do the headaches usually appear?
French	Quand avez-vous habituellement vos maux de tête?
Arabic	Mataa tahus bi al soudaa adatan?
Japanese	Daitai itsu atama wa itami masu ka?
English	Is the pain associated with nausea?
French	Avez-vous des nausées quand vous avez la douleur?
Arabic	Hal tourid an tataqaya indama tahus bi al alam
Japanese	Itamu to hakike wa okori masu ka?
English	Does bright light make the headache worse?
French	Vos maux de tête sont ils aggravés par la lumière?
Arabic	Hal yachtaddou al soudaa fi al dhaw?
Japanese	Akarui hikari wo miru to zutsu wa hidoku nari masu ka?

Table 2: Examples of English domain sentences, with system translations into French, Arabic and Japanese.

Our metrics measure the extent to which the derived versions of the SMT were able to approximate the original RBMT on data which was within the RBMT's coverage. The most straightforward way to do this is simply to count the sentences in the test set which receive different translations from the RBMT and the SMT. A variant is to define a non-standard version of the BLEU metric (Papineni et al., 2001), with the RBMT's translation taken as the reference. This means that perfect correspondence between the two translations would yield a non-standard BLEU score of 1.0.

For all these metrics, it is important to bring in human judges at some point, using them to evaluate the cases where the SMT and RBMT differ. If, in these cases, it transpired that human judges typically thought that the SMT was as good as the RBMT, then the metrics would not be useful. We need to satisfy ourselves that human judges typically ascribe differences between SMT and RBMT to shortcomings in the SMT rather than in the RBMT.

Concretely, we collected all the different ⟨Source, SMT-translation, RBMT-translation⟩ triples produced during the course of the experiments, and extracted those triples where the two translations were different. We randomly selected triples for selected language pairs, and asked human judges to classify them into one of the following categories:

- **RBMT better:** The RBMT translation was better, in terms of preserving meaning and/or being grammatically correct;
- **SMT better:** The SMT translation was better, in terms of preserving meaning and/or being grammatically correct;
- **Similar:** Both translations were about equally good OR the source sentence was meaningless in the domain.

In order to show that our metrics are intuitively meaningful, it is sufficient to demonstrate that the frequency of occurrence of **RBMT better** is both large in comparison to that of **SMT better**, and accounts for a substantial proportion of the total population.

In the next section, we present the results of the various experiments we have just described.

4 Results

Tables 4, 5 and 6 present the main results, summarising the extent to which SMT and RBMT translations differ for the 12 language-pairs. Since the training and test data are independently sampled from the source grammar, and the domain is quite constrained, they overlap. This is natural, since, in this limited domain, it is to be expected that some

training sentences will also occur in test data; basic questions like "Where is the pain?" will be generated with high frequency by the probabilistic source language model, and will tend to occur in any substantial independently generated set, hence both in test and training. When counting divergent translations in RBMT and SMT output, we none the less present separately results for test data that does not overlap with training data (Table 4) and for test data that does overlap with training data (Table 5), on the grounds that the figures are, as usual, very different for the two kinds of material. These two tables thus summarise agreement between SMT and RBMT at the sentence level. Table 6 shows the non-standard BLEU scores, where the RBMT translations have been used as the reference; these give a picture of agreement between the two types of translation at the n-gram level.

Looking in particular at Table 4, we see that the figures fall into three distinct groups. For language-pairs involving only languages in the group {English, French, Arabic}, SMT and RBMT agree on about 70% to 80% of the sentences. For translation from English and French to Japanese, the two types of translation agree on about 27% of the sentences. For translation from Japanese into the other three languages, and for Arabic into Japanese, we only get agreement on about 13% to 16% of the sentences. These divisions appear to show clear qualitative differences.

Source	Target			
	Eng	Fre	Ara	Jap
Eng	xxx	69.6	76.5	27.9
Fre	77.1	xxx	72.4	26.9
Ara	76.7	79.1	xxx	13.9
Jap	15.7	14.7	12.7	xxx

Table 4: Percentage of translations where SMT translation coincides with RBMT translation, over test sentences **not occurring** in training data.

As discussed in the previous section, simply counting differences between SMT and RBMT says nothing on its own; it is also necessary to establish what these differences mean in terms of human judgements. We performed evaluations of this kind for two representative language-pairs where we found it easy to locate bilingual judges. First, Table 8 shows the categorisation, according to the criteria outlined at the end of Section 3, for 500 English → French pairs randomly selected from the set of examples where RBMT and SMT gave different results; we asked three judges to evaluate them independently, and combined their judgments by majority decision where appropriate. We observed a very heavy bias towards the RBMT, with unanimous agreement that the RBMT translation was better in 201/500 cases, and 2-1 agreement in a further 127. In contrast, there were only 4/500 cases where the judges unanimously thought that the SMT translation was preferable, with a further 12 supported by a majority decision. The rest of the table gives the cases where the RBMT and SMT translations were judged the same or cases in which the judges disagreed; there were only 41/500 cases where no majority decision was reached. Our overall conclusion is that we are justified in evaluating the SMT by using the BLEU scores with the RBMT as the reference. Of the cases where the two systems differ, only a tiny fraction, at most 16/500, indicate a better translation from the SMT, and well over half are translated better by the RBMT. Table 7 shows some examples of bad SMT translations in the English → French pair, contrasted with the translations

Source	Target			
	Eng	Fre	Ara	Jap
Eng	xxx	87.9	92.4	77.8
Fre	94.7	xxx	94.4	74.4
Ara	95.2	90.8	xxx	64.0
Jap	79.1	81.4	76.6	xxx

Table 5: Percentage of translations where SMT translation coincides with RBMT translation, over test sentences **occurring** in training data.

Source	Target			
	Eng	Fre	Ara	Jap
Eng	xxx	0.91	0.92	0.79
Fre	0.93	xxx	0.92	0.76
Ara	0.97	0.98	xxx	0.74
Jap	0.80	0.83	0.85	xxx

Table 6: Non-standard BLEU scores (RBMT translations used as reference), all data.

produced by the RBMT. The first two are grammatical errors (a superfluous extra verb in the first, and agreement errors in the second). The third is an bad choice of tense and preposition; although grammatical, the target language sentence fails to preserve the meaning, and, rather than referring to a 20 day period ending now, instead refers to a 20 day period some time in the past.

Table 10 shows a similar evaluation for the English → Japanese. Here, the difference between the SMT and RBMT versions was so pronounced that we felt justified in taking a smaller sample, of only 150 sentences. This time, 92/150 cases were unanimously judged as having a better RBMT translation, and there was not a single case where even a majority found that the SMT was better. Agreement was good here too, with only 8/150 cases not yielding at least a majority decision. Unsurprisingly, the main problem with this language-pair was inability to handle correctly the differences between English and Japanese word-order. Table 9 again shows some typical examples. The errors are much more serious than in French, and the SMT translations are only marginally comprehensible.

Result	Agreement	Count
RBMT better	all judges	201
RBMT better	majority	127
SMT better	all judges	4
SMT better	majority	12
Similar	all judges	34
Similar	majority	81
Unclear	disagree	41
Total		500

Table 8: Comparison of RBMT and SMT performance on 500 randomly chosen English → French translation examples, evaluated independently by three judges.

Cursory examination of the remaining language-pairs strongly suggested that the same patterns obtained there as well, with very few cases where SMT was better than RBMT, and numerous cases in the opposite direction. Since other evaluations of the MedSLT system (e.g. (Rayner et al., 2005)) show that over 98% of in-coverage translations produced by the RBMT system are acceptable in terms of preserving meaning and being grammatically correct, our overall conclusion is that differences between SMT and RBMT can plausibly be interpreted as reflecting errors produced by the SMT.

Result	Agreement	Count
RBMT better	all judges	92
RBMT better	majority	32
SMT better	all judges	0
SMT better	majority	0
Similar	all judges	2
Similar	majority	16
Unclear	disagree	8
Total		150

Table 10: Comparison of RBMT and SMT performance on 150 randomly chosen English → Japanese translation examples, evaluated independently by three judges.

5 Summary and Conclusions

We have presented an experiment in which we generated uniform artificial data for 12 language pairs in a multilingual small-vocabulary interlingua-based translation system. Use of the interlingua enforced a uniform translation standard, so we feel justified in claiming that the results provide harder evidence than usual about the relative suitability of different language-pairs for SMT.

As expected, translation between English and French in both directions is much more reliable than translation in language pairs involving Japanese. To our surprise, we also found that translation between English or French and Arabic worked about as well as translation between English and French, despite the fact that Arabic typologically belongs to a different language family. Informal conversations with colleagues who have worked on Arabic suggest, however, that the result is not as unexpected as we first imagined.

Table 4 appears to suggest that translation *from* Japanese works substantially less well than translation *to* Japanese. The explanation is most probably the usual problem of zero-anaphora, which is very common in Japanese, with words that can clearly be inferred from context generally deleted. In the from-Japanese direction, it is necessary to generate a translation of a zero anaphor (most often the implicit second-person pronoun), while in the to-Japanese

English	does a temperature change cause the headache
RBMT French	vos maux de tête sont-ils causés par des changements de température
	(your headaches are-they caused by changes of temperature)
SMT French	**avez-vous** vos maux de tête sont-ils causés par des changements de température
	(**have-you** your headaches are-they caused by changes of temperature)
English	are headaches relieved in the afternoon
RBMT French	vos maux de tête diminuent-ils l'après-midi
	(your headaches (MASC-PLUR) decrease-MASC-PLUR the afternoon)
SMT French	vos maux de tête **diminue-t-elle** l'après-midi
	(your headaches (MASC-PLUR) decrease-**FEM-SING** the afternoon)
English	have you had them for twenty days
RBMT French	avez-vous vos maux de tête depuis vingt jours
	(have-you your headaches since twenty days)
SMT French	avez-vous **eu** vos maux de tête **pendant** vingt jours
	(have-you **had** your headaches **during** twenty days)

Table 7: Examples of incorrect SMT translations from English into French. Errors are highlighted in bold.

direction it is only a question of deleting material. Although, as pointed out earlier in Section 3, we need to be careful when comparing between language-pairs with different source-language, the gap in performance here is large enough that we can expect it to reflect a real trend.

Simple as the idea is, we hope that the methodology described in this paper will make it possible to evaluate the relative suitability of SMT for different language pairs in a more quantitative way than has so far been possible. In general, the construction used could equally well be implemented in the context of any other high-performance multilingual RBMT system. The idea of statistically "relearning" an RBMT system has recently begun to acquire some popularity (Seneff et al., 2006; Dugast et al., 2008), and it should be easy to check whether our results are generally reproducible.

References

A. Birch, M. Osborne, and P. Koehn. 2008. Predicting success in machine translation. In *Proceedings of EMNLP 2008*, Waikiki, Hawaii.

P. Bouillon, G. Flores, M. Georgescul, S. Halimi, B.A. Hockey, H. Isahara, K. Kanzaki, Y. Nakao, M. Rayner, M. Santaholma, M. Starlander, and N. Tsourakis. 2008. Many-to-many multilingual medical speech translation on a PDA. In *Proceedings of The Eighth Conference of the Association for Machine Translation in the Americas*, Waikiki, Hawaii.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

N. Chatzichrisafis, P. Bouillon, M. Rayner, M. Santaholma, M. Starlander, and B.A. Hockey. 2006. Evaluating task performance for a unidirectional controlled language medical speech translation system. In *Proceedings of the HLT-NAACL International Workshop on Medical Speech Translation*, pages 9–16, New York.

L. Dugast, J. Senellart, and P. Koehn. 2008. Can we relearn an RBMT system? In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 175–178, Columbus, Ohio.

B.A. Hockey, M. Rayner, and G. Christian. 2008. Training statistical language models from grammar-generated data: A comparative case-study. In *Proceedings of the 6th International Conference on Natural Language Processing*, Gothenburg, Sweden.

R. Jonson. 2005. Generating statistical language models from interpretation grammars in dialogue systems. In *Proceedings of the 11th EACL*, Trento, Italy.

English	have you had the attacks in the evening
RBMT Japanese	ban ari mashita ka
	evening be POLITE-PAST Q
SMT Japanese	ari mashita ka **ban**
	be POLITE-PAST Q **evening**
English	does noise typically give you your headaches
RBMT Japanese	souon ga kini naru to daitai atama wa itami masu ka
	noise SUBJ experience WHEN often head TOP hurt POLITE Q
SMT Japanese	souon ga kini naru to **itami masu ka** daitai atama wa itami masu ka
	noise SUBJ experience WHEN **hurt POLITE Q** often head TOP hurt POLITE Q
English	is the pain usually caused by sudden head movements
RBMT Japanese	daitai totsuzen atama wo ugokasu to itami masu ka
	usually quickly head OBJ move WHEN hurt POLITE Q
SMT Japanese	daitai **itami wa** totsuzen atama wo ugokasu to ***[Missing: itami masu ka]***
	usually **pain SUBJ** quickly head OBJ move WHEN ***[Missing: hurt POLITE Q]***

Table 9: Examples of incorrect SMT translations from English into Japanese. Errors are highlighted in bold.

A. Jurafsky, C. Wooters, J. Segal, A. Stolcke, E. Fosler, G. Tajchman, and N. Morgan. 1995. Using a stochastic context-free grammar as a language model for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 189–192.

P. Koehn and C. Monz. 2005. Shared task: Statistical machine translation between european languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, MI.

P. Koehn and C. Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, New York City, NY.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 2.

F.J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

S. Ozdowska. 2009. Données bilingues pour la tas français-anglais : impact de la langue source et direction de traduction originales sur la qualité de la traduction. In *Proceedings of TALN 2009*, Senlis, France.

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

M. Rayner, P. Bouillon, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, H. Isahara, K. Kanzaki, and Y. Nakao. 2005. A methodology for comparing grammar-based and robust approaches to speech understanding. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP)*, pages 1103–1107, Lisboa, Portugal.

M. Rayner, B.A. Hockey, and P. Bouillon. 2006. *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. CSLI Press, Chicago.

M. Rayner, P. Estrella, P. Bouillon, B.A. Hockey, and Y. Nakao. 2009. Using artificially generated data to evaluate statistical machine translation. In *Proceedings of the Third Workshop on Grammar Engineering Across Frameworks*, Singapore.

S. Seneff, C. Wang, and J. Lee. 2006. Combining linguistic and statistical methods for bi-directional English Chinese translation in the flight domain. In *Proceedings of AMTA 2006*.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*. ISCA.