A Bootstrapped Interlingua-Based SMT Architecture

Manny Rayner[1]**, Paula Estrella**[2]**, Pierrette Bouillon**[1]
(1) University of Geneva, TIM/ISSCO
40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
{Emmanuel.Rayner,Pierrette.Bouillon}@unige.ch

(2) FaMAF, U. Nacional de Córdoba
5000 - Córdoba, Argentina
pestrella@famaf.unc.edu.ar

Abstract

We describe a simple and general construction, which can be used to bootstrap useful SMT models from an interlingua-based MT system and add non-trivial robustness. As in previous work, the rule-based system is used to generate aligned data, which is then used to train SMTs. The novelty described here is to introduce an "interlingua grammar" which associates interlingua representations with surface text strings in a reversible way, making it possible to factor the induced SMT translation into Source → Interlingua and Interlingua → Target components. We describe several refinements of the basic scheme. If the source and target languages have widely different word-orders, performance can be greatly improved by defining two different surface forms for the interlingua grammar, based on the source and target languages respectively; the interlingua grammar can be used to rescore N-best SMT translation hypotheses; and, finally, one can combine SMT and RBMT modules into a hybrid system, increasing robustness without sacrificing precision. We have implemented these ideas inside English → French and English → Japanese versions of the Open Source MedSLT medical speech translator, and present an evaluation.

1 Introduction

At the moment, the dominant paradigm for machine translation is the statistical one (Statistical Machine Translation; SMT), but rule-based machine translation (RBMT) is far from dead. The advantages and disadvantages of each approach are well known. SMT systems are robust, and can be built quickly if sufficient quantities of bilingual data are available. RBMT systems, on the other hand, can be built without much training data, and appear to be more reliable, at least in limited domains (Seneff et al., 2006; Wilks, 2007). In applications where training data is hard to obtain, and precision is more important than recall, there is still much to recommend them.

To get the best of both worlds — a robust system that can be constructed without a large bilingual corpus — there is a natural way to combine SMT and RBMT: we use the RBMT to create artificial training data for an SMT model. A prominent recent example is (Dugast et al., 2008), which describes an experiment where SYSTRAN was used to translate a monolingual French corpus, creating an aligned corpus which then served as training data to create a French → English SMT model.

The present paper has as its starting point an earlier study, described in (Rayner et al., 2009), which used MedSLT (Bouillon et al., 2008), a medium-vocabulary interlingua-based multilingual Open Source medical speech translator. The goal was to bootstrap a useful SMT from the RBMT. We generated large parallel corpora from English → French and English → Japanese versions of the system, trained SMT models from them, and tested these models on data which was outside the coverage of the RBMT. Our hope that the SMT would be able to add robustness to the RBMT, recovering on some input which the RBMT was unable to process, but the results reported were negative. Although the SMT did produce good translations for about 15% of the out-of-coverage sentences, about as many more were translated incorrectly. We concluded that the major loss of precision rendered the small improvement in recall worthless.

Here, we show, on the contrary, that it is in fact quite possible to achieve the goals we set ourselves in the earlier paper, if we correctly exploit the interlingua-based architecture of the original RBMT system to train separate SMT models for translation from source language to interlingua, and from interlingua to target language. The key technical idea is to define an "interlingua grammar", which associates each interlingua representation with a surface text form, which we will call an "interlingua gloss". We can then construct aligned corpora which pair source or target sentences with interlingua glosses.

Factoring SMT translation through the interlingua turns out to offer several advantages. To begin with, the original RBMT system's ability to offer useful performance on noisy speech input depends crucially on the interlingua; in the live application, each sentence produced by the speech recogniser is first translated into the interlingua, and then "backtranslated" into the source language. The user is given a chance to approve or abort the backtranslation before a target language sentence is produced. The system gives reliable translations for sentences which produce good backtranslations, while the remaining ones are discarded. If SMT is performed using the interlingua as a pivot, it is possible to employ the same basic architecture. As we will show later, a hybrid system can also use SMT to translate into the interlingua and then backtranslate the result before translating to the target, improving robustness without compromising reliability.

Once the interlingua grammar is available, it turns out that we can also exploit it for other purposes. First, if the SMT decoder is set to produce N-best output, we can use the interlingua grammar as a knowledge source to reorder N-best hypotheses, preferring ones which the grammar defines as well-formed. Second, when the source and target languages have widely different word-orders, SMT translation can be made far more accurate when it is broken up into several processing steps. Here, we were partly inspired by (Xu and Seneff, 2008), who address the problem arising from word-order differences when translating from English to Chinese. They first perform RBMT from the English source to an intermediate representation they call "Zhonglish", in which English words are arranged in a Chinese order; they then use an SMT to produce the final Chinese result. For English to Japanese translation, we have a similar set of modules, but connected in a different order: we first use SMT to translate English into an English-like interlingua, then reformulate the interlingua into a Japanese-ordered "Japlish", and finally use RBMT to generate Japanese.

The rest of the paper is organised as follows. Sections 2 and 3 present background on the MedSLT system, and the way it uses interlingua; Section 4 describes the experimental framework, and Section 5 the experiments themselves; Section 6 gives the results; and Section 7 concludes.

2 Background: 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 in AFF (Almost Flat Functional Semantics; (Rayner et al., 2008)). This is first translated by one set of rules into an interlingual form, and then by a second set into a target language representation. The interlingua and target representation are also in AFF form. A target-language Regulus grammar, compiled into generation form, turns the target representation into one or more possible surface strings, after which a set of generation preferences picks one out. Finally, the selected string is realised as spoken output.

Language	Vocab	WER	SemER
English	447	6%	11%
French	1025	8%	10%
Japanese	422	3%	4%

Table 1: Recognition performance for English, French and Japanese headache-domain recognisers. "Vocab" = number of surface words in source language recogniser vocabulary; "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. Differences in vocabulary size are mainly related to differences in inflectional morphology.

3 Interlingua and interlingua grammars

The space of well-formed interlingua representations in MedSLT is defined by yet another Regulus grammar (Bouillon et al., 2008); this grammar is designed to have minimal structure, so checking for well-formedness can be performed very quickly. During speech understanding, the well-formedness check is used as a knowledge source to enhance the language model for the source language. The speech recogniser is set to generate N-best recognition hypotheses, and hypotheses which give rise to non-well-formed interlingua can safely be discarded. Use of this "highest-in-coverage" rescoring algorithm is found to reduce semantic error rate during speech understanding by about 10% relative.

The interlingua grammar is built in such a way that the surface forms it defines can also be used as human-readable glosses. We will make heavy use of these glosses in what follows. The usual form of the "interlingua gloss language" is modelled on English. It is, however, straightforward to parametrize the grammar so that glosses can also be generated with word-orders based on those occurring in other languages; here, we have created one based on Japanese.

Table 2 shows examples of English domain sentences together with translations into French and Japanese, and interlingua glosses in English-based and Japanese-based format. Note the very simple structure of the interlingua gloss, which is in most cases just a concatenation of text representations for the underlying AFF representation; since AFF representations are unordered lists, they can be presented in any desired order. Thus the AFF for the first example, "does the pain usually last for more than one day" is[1]

```
[null=[utterance_type,ynq],
 arg1=[symptom,pain],
 null=[state,last],
 null=[tense,present],
 null=[freq,usually],
 duration=[spec,[more_than,1]],
 duration=[timeunit,day]]
```

The English-format interlingua gloss, "YN-QUESTION pain last PRESENT usually duration more-than one day" presents these elements in the order given here, which is approximately that of a normal English rendition of the sentence. In contrast, the Japanese-format gloss, "more-than one day duration pain usually last PRESENT YN-QUESTION" makes concessions to standard Japanese word-order, in which the sentence normally ends with the verb (here, *tsuzuki masu*), followed by the interrogative particle *ka*.

Similarly, in the second example from Table 2, we see that the English-format gloss puts "sc-when" ("subordinating-conjunction when") before the representation of the subordinate clause; the Japanese-format gloss puts "sc-when" after, mirroring the fact that the corresponding Japanese particle, *to*, comes after the subordinate clause *tabemono wo taberu*. This is literally "food OBJ eat", i.e. "(you) eat food"; note that the Japanese-format interlingua suppresses the personal pronoun "you", again following normal Japanese usage.

In the next section, we explain how we use the interlingua, and in particular the interlingua gloss forms, to create a bootstrapped SMT framework much more powerful than the one from (Rayner et al., 2009). We first review their construction, and then explain what we have added to it.

4 Experimental framework

We start with a well-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

[1] AFF representations and glosses have been slightly simplified for presentational reasons.

English	does the pain usually last for more than one day
Eng-Interlingua	YN-QUESTION pain last PRESENT usually duration more-than one day
French	la douleur dure-t-elle habituellement plus d'un jour
Jap-Interlingua	more-than one day duration pain usually last PRESENT YN-QUESTION
Japanese	daitai ichinichi sukunakutomo itami wa tsuzuki masu ka
English	does it ever appear when you eat
Eng-Interlingua	YN-QUESTION you have PRESENT ever pain sc-when you eat PRESENT
French	avez-vous déjà eu mal quand vous mangez
Jap-Interlingua	eat PRESENT sc-when ever pain have PRESENT YN-QUESTION
Japanese	koremadeni tabemono wo taberu to itami mashita ka
English	is the pain on one side
Eng-Interlingua	YN-QUESTION you have PRESENT pain in-loc head one side-part
French	avez-vous mal sur l'un des côtés de la tête
Jap-Interlingua	head one side-part in-loc pain have PRESENT YN-QUESTION
Japanese	atama no katagawa wa itami masu ka

Table 2: English MedSLT examples: English source sentence, English-format interlingua gloss, RBMT translation into French, Japanese-format interlingua gloss and RBMT translation into Japanese

(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 use 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 corpus of arbitrary size, which can be used to train an SMT model. In (Rayner et al., 2009), the corpus was a bilingual one, consisting of ⟨Source, Target⟩ pairs. In the present paper, our corpora also contain the intermediate interlingua steps, and thus consist of ⟨Source, Interlingua-Gloss, Target⟩ triples.

We used this method to generate aligned corpora for English → Interlingua → French and English → Interlingua → Japanese. Each aligned corpus started with one million randomly generated English sentences. After discarding sentences which received no translation, we were left with about 310K triples. 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 for the following six pairs: English → English-Interlingua; English → French; English → Japanese; English-Interlingua → French; Japanese-Interlingua → Japanese; English-Interlingua → Japanese. The models were tuned in the standard way using MERT. As reported in (Rayner et al., 2009), the quantity of training data available appears easily sufficient to ensure that translation performance tops out.

The resulting models were combined in the ways described in Section 5 to translate the test portion of the English corpus. Again following (Rayner et al., 2009), our primary evaluation metric quantifies agreement between the translations produced by the SMT and those produced by the RBMT. We use the most straightforward measure: we take those sentences in the test set which do not also occur in the training material (since both sets are independently randomly generated, overlap is inevitable), and count the proportion for which the SMT translation is the same as the RBMT translation. As demonstrated in the earlier paper, evaluation by human judges indicates that differences frequently favour the RBMT and hardly ever favour the SMT. This shows that the metric has intuitive significance, and that scores of less than 100% represent real deficiencies in the SMT's performance. Finally, we tested the best configurations on the out-of-coverage MedSLT dataset from (Rayner et al., 2009), using human judges to evaluate the results.

5 Experiments

We combined the resources described in the previous sections to compare the performance of several different translation pipelines, for both English → French and English → Japanese:

5.1 Plain RBMT

Translation using the baseline RBMT system.

5.2 Plain SMT

Translation using a Source → Target SMT model.

5.3 SMT + SMT

Translation using a Source → English-interlingua SMT model composed with an English-interlingua → Target SMT model.

5.4 SMT + interlingua-reformulation + SMT

For translation to Japanese, the Japanese-interlingua → Japanese SMT model is much better than the English-interlingua → Japanese SMT model, since the word-orders are closer. It thus makes sense to perform the sequence Source → English-Interlingua, using SMT; English-Interlingua → Japanese-Interlingua, using rule-based reformulation of the interlingua gloss; and finally Japanese-Interlingua → Japanese, using SMT.

5.5 SMT + rescoring + SMT

Another possible refinement is to use the interlingua grammar to rescore Source → Interlingua SMT results. Just as in the case of speech recognition (cf. Section 2), we can set the SMT decoding engine to produce a list of N-best hypotheses; we rescore this list by selecting the highest hypothesis that is well-formed according to the interlingua grammar, or the first hypothesis if no well-formed hypothesis exists. The result is then passed through the Interlingua-gloss → Target SMT model.

5.6 SMT + rescoring + interlingua-reformulation + SMT

A combination of 5.5 and 5.4; in the case of translation to Japanese, we can perform SMT and rescoring as in 5.5 to get English-Interlingua, then reformulate to Japanese-Interlingua and perform Japanese-Interlingua → Japanese SMT as in 5.4.

5.7 SMT + RBMT

We use SMT to perform Source → English-Interlingua translation, then do English-Interlingua → Target using RBMT if the interlingua is well-formed. Ill-formed interlingua representations fail to produce a translation.

5.8 SMT + rescoring + RBMT

As in 5.7, but setting the Source → English-Interlingua to create N-best output, and rescoring it using the interlingua grammar before performing RBMT.

6 Results

Table 3 presents the results of running the different configurations described in the previous section on randomly generated in-coverage data, evaluated by measuring the proportion of not-in-training sentences for which translation matches the RBMT gold standard. As previously reported in (Rayner et al., 2009), English → French scores much better than English → Japanese with plain SMT (65.8% versus 26.8%).

We had expected that performance on English → Japanese would improve when we split up SMT translation into two pieces, with an interlingua-reformulation phase in between. SMT's problems with English → Japanese stem from the very different word-orders in the two languages, and interlingua-reformulation levels the playing-field, ensuring that SMT translation always takes place between languages with similar word-orders. We had not anticipated, however, that the improvement would be so large that factored English → Japanese would outscore plain English → French (74.1% versus 65.8%), and we were also surprised to find that factored English → French was considerably better than plain English → French (76.6% versus 65.8%). It is evident that factoring only helps if the interlingua formats are appropriately chosen; factored English → Japanese without interlingua reformulation is in fact much worse than plain English → Japanese (10.5% versus 26.8%).

Rescoring helps to improve performance on factored SMT; English → French increases from 76.6% to 78.5%, and English → Japanese from 74.1% to 78.5%. Finally, we look at the hybrid system, which combines SMT translation from source to interlingua with RBMT translation from interlingua to target. This is noticeably better than factored SMT: 83.5% versus 76.6% for English

Configuration	Eng → Fre	Eng → Jap
Plain RBMT	(100%)	(100%)
Plain SMT	65.8%	26.8%
SMT + SMT	76.6%	10.5%
SMT + interlingua-reformulation + SMT	—	74.1%
SMT + rescoring + SMT	78.5%	10.8%
SMT + rescoring + interlingua-reformulation + SMT	—	78.5%
SMT + RBMT	83.5%	81.9%
SMT + rescoring + RBMT	87.0%	87.1%

Table 3: Translation performance of different versions of the translation pipeline on randomly generated in-coverage test sentences not in training data. The figures show the proportion of translations which agree with the RBMT translation.

→ French, and 81.9% versus 74.1% for English → Japanese. Rescoring also combines well with the hybrid SMT + RBMT configurations, since the RBMT-based interlingua → target phase requires that the interlingua is well-formed. The hybrid configurations including rescoring have almost identical performance, at around 87%.

In order to investigate whether the new architecture was potentially capable of adding robustness to the speech translation system, we ran three configurations of the pipeline which involved use of the interlingua on the 358 out-of-coverage English sentences from (Rayner et al., 2009); these are transcriptions of spoken utterances from a real data collection exercise. The intention was to simulate normal use of the system, where the user would be given a backtranslation of the source, and allowed to abort sentences which had been unsuccessfully rendered into Interlingua.

To this end, we used SMT to translate the English source sentences into interlingua in N-best mode, and rescored using the interlingua grammar to pick the highest in-coverage translation. The SMT decoder was set to discard out-of-vocabulary words, after some preliminary experiments showed that this was the most effective strategy. Then, using the Interlingua → English RBMT component, we translated all the well-formed interlingua utterances produced by this process back into English, and asked an English native speaker to judge the resulting English → English translations for correctness. Finally, using both RBMT and SMT, we translated into French and Japanese the well-formed interlingua translations marked as having correct backtranslations. For SMT translation into Japanese, the original English-format interlingua was first reformulated into Japanese-format interlingua. The results are summarised in Table 4; Table 5 gives some examples of robust translations produced using the combination of SMT up to interlingua and RBMT from interlingua to target.

Of the 358 sentences, 81 (23%) produced an English backtranslation that was judged to be correct, and would thus not have led to the user aborting translation. When RBMT was used to translate these 81 sentences into the target language, 6 (7%) failed to produce a French translation, with no incorrect translations; for Japanese, there were no failed translations, and 4 (5%) translations judged incorrect. Three of the four English sentences which produced incorrect Japanese translations were occurrences of “does the pain last a long time”, backtranslated as “does the pain last” and judged as acceptable; “a long time” is a vague expression which does not clearly add anything to “last”. The French translation, “vos maux de tête durent-ils” is acceptable for similar reasons; however, the Japanese translation, *zutsu wa tsuzuki masu ka* (“pain TOPIC last PRESENT Q”) is incorrect, since *tsuzuki masu* with no temporal modifier has the meaning “continue (since the last time we talked)” rather than “last”. We find this an interesting example illustrating how difficult it is to provide very high quality translation, even in a limited domain.

When SMT was used for the interlingua → target phase, a translation was always produced, but there were more mistakes; 5 sentences (6%) were judged incorrect for French, and 10 (12%) for Japanese. Given the importance of precision to the application, it seems clear that one would in practice prefer the hybrid (SMT + RBMT) configuration, but factored SMT is not enormously worse.

Original sentences	358
Well-formed interlingua translation produced	245
English RBMT backtranslation produced	213
Backtranslation judged correct	81
French RBMT translation produced	75
French RBMT translation judged correct	75
French SMT translation produced	81
French SMT translation judged correct	76
Japanese RBMT translation produced	81
Japanese RBMT translation judged correct	77
Japanese SMT translation produced	81
Japanese SMT translation judged correct	

Table 4: Results of simulating the speech translation system on out-of-coverage data. Sentences are translated into interlingua using SMT and rescoring, backtranslated into English using RBMT, and judged. Sentences with correct backtranslations are translated into the target language using both RBMT and SMT.

7 Summary and conclusions

We have defined a simple and general construction which can be used to bootstrap SMT models from an interlingua-based RBMT system, and evaluated it concretely in the context of English → French and English → Japanese versions of the MedSLT medical speech translator. The central idea is to define grammars that associate interlingua representations with surface forms, which we call "interlingua glosses". This makes it possible to generate aligned corpora of source/interlingua-gloss or interlingua-gloss/target pairs, and induce a factored SMT system, with separate SMT modules for source → interlingua and interlingua → target translation.

By defining two versions of the interlingua gloss form, tailored to the word-orders of the source and target languages, we can address the problems that arise when using SMT between languages with very different word-orders. In MedSLT, we have shown how this allowed us to improve SMT performance in the difficult pair English → Japanese to the point where it was approximately as good as in the easy pair English → French. We have also shown how the interlingua grammar can be used as a knowledge source to rescore N-best SMT translation hypotheses, significantly improving translation quality.

Finally, we described a hybrid architecture which combines SMT and RBMT modules. This uses SMT to translate from source to interlingua, while RBMT is used both to translate from interlingua to target, and also to produce a "backtranslation" into the source language. In a safety-critical application like MedSLT, this adds useful robustness without seriously compromising precision. The backtranslation allows the user to abort unsuccessful translations produced by the SMT-based source → interlingua module, and be confident that the remaining ones are accurately translated using the RBMT-based interlingua → target module.

In an initial evaluation using text transcriptions of English MedSLT data, 21% of the out-of-coverage sentences were judged as having correct backtranslations, 97% of the sentences with correct backtranslations produced a target language translation, and 98% of the target language translations were judged correct. We find these figures distinctly encouraging. In the next phase of the project, we will attempt to tune performance further, and experiment with speech input data.

References

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.

L. Dugast, J. Senellart, and P. Koehn. 2008. Can we relearn an RBMT system? In *Proceedings of the*

English	has the intensity of your headaches increased
Interlingua	YN-QUESTION headache become-worse PRESENT-PERFECT
B/translation	have the headaches been worse
French	vos maux de tête ont-ils empiré
Japanese	zutsu wa hidoku nari mashita ka
(Jap. gloss)	headache TOPIC bad become PAST Q
English	is the pain related to stress
Interlingua	YN-QUESTION you have PRESENT pain sc-when you experience PRESENT stress
B/translation	do you experience the pain when you feel stressed
French	avez-vous mal quand vous êtes stressé
Japanese	sutoresu wo kanjiru to itami masu ka
(Jap. gloss)	stress OBJ feel-PLAIN if hurt POLITE-PRESENT Q
English	have the headaches become more frequent
Interlingua	YN-QUESTION frequency increase PRESENT headache
B/translation	is the frequency of the headaches increasing
French	la fréquence de vos maux de tête augmente-t-elle
Japanese	zutsu no hindo wa fuete imasu ka
(Jap. gloss)	headache GEN frequency TOPIC increase POLITE-PRESENT Q

Table 5: Examples of robust translation with the hybrid SMT/RBMT architecture. The out-of-coverage English source sentence is translated to interlingua using SMT with rescoring, and then (back-)translated into English, French and Japanese using RBMT.

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.

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, 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.

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

M. Rayner, P. Bouillon, B.A. Hockey, and Y. Nakao. 2008. Almost flat functional semantics for speech translation. In *Proceedings of COLING-2008*, Manchester, England.

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 2009 Workshop on Grammar Engineering Across Frameworks*, pages 54–62, Singapore. Association for Computational Linguistics.

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.

Y. Wilks. 2007. Stone soup and the French room. In K. Ahmad, C. Brewster, and M. Stevenson, editors, *Words and Intelligence I: Selected Papers by Yorick Wilks*, pages 255–265.

Y. Xu and S. Seneff. 2008. Two-Stage Translation: A Combined Linguistic and Statistical Machine Translation Framework. In *Proceedings of The Eighth Conference of the Association for Machine Translation in the Americas*, Waikiki, Hawaii.