Using Artificially Generated Data to Evaluate Statistical Machine Translation

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

Beth Ann Hockey
Mail Stop 19-26, UCSC UARC
NASA Ames Research Center, Moffett Field, CA 94035–1000
bahockey@ucsc.edu

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

Abstract

Although Statistical Machine Translation (SMT) is now the dominant paradigm within Machine Translation, we argue that it is far from clear that it can outperform Rule-Based Machine Translation (RBMT) on small- to medium-vocabulary applications where high precision is more important than recall. A particularly important practical example is medical speech translation. We report the results of experiments where we configured the various grammars and rule-sets in an Open Source medium-vocabulary multi-lingual medical speech translation system to generate large aligned bilingual corpora for English → French and English → Japanese, which were then used to train SMT models based on the common combination of Giza++, Moses and SRILM. The resulting SMTs were unable fully to reproduce the performance of the RBMT, with performance topping out, even for English → French, with less than 70% of the SMT translations of previously unseen sentences agreeing with RBMT translations. When the outputs of the two systems differed, human judges reported the SMT result as frequently being worse than the RBMT result, and hardly ever better; moreover, the added robustness of the SMT only yielded a small improvement in recall, with a large penalty in precision.

1 Introduction

When Statistical Machine Translation (SMT) was first introduced in the early 90s, it encountered a hostile reception, and many people in the research community were unwilling to believe it could ever be a serious competitor to symbolic approaches (cf. for example (Arnold et al., 1994)). The pendulum has now swung all the way to the other end of the scale; right now, the prevailing wisdom within the research community is that SMT is the only truly viable architecture, and that rule-based machine translation (RBMT) is ultimately doomed to failure. In this paper, one of our initial concerns will be to argue for a compromise position. In our opinion, the initial scepticism about SMT was not groundless; the arguments presented against it often took the form of examples involving deep linguistic reasoning, which, it was claimed, would be hard to address using surface methods. Proponents of RBMT had, however, greatly underestimated the extent to which SMT would be able to tackle the problem of robustness, where it appears to be far more powerful than RBMT. For most machine translation applications, robustness is the central issue, so SMT's current preeminence is hardly surprising.

Even for the large-vocabulary tasks where SMT does best, the situation is by no means as clear as one might imagine: according to (Wilks, 2007), purely statistical systems are still unable to outperform SYSTRAN. In this paper, we will however be more concerned with limited-domain MT tasks, where robustness is not the key requirement, and accuracy is paramount. An immediate exam-

Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks, ACL-IJCNLP 2009, pages 54–62,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

ple is medical speech translation, which is establishing itself as an an application area of some significance (Bouillon et al., 2006; Bouillon et al., 2008a). Translation in medical applications needs to be extremely accurate, since mistranslations can have serious or even fatal consequences. At the panel discussion at the 2008 COLING workshop on safety-critical speech translation (Rayner et al., 2008), the consensus opinion, based on input from practising physicians, was that an appropriate evaluation metric for medical applications would be heavily slanted towards accuracy, as opposed to robustness. If the metric is normalised so as to award 0 points for no translation, and 1 point for a correct translation, the estimate was that a suitable score for an incorrect translation would be something between –25 and –100 points. With these requirements, it seems unlikely that a robust, broad-coverage architecture has much chance of success. The obvious strategy is to build a limited-domain controlled-language system, and tune it to the point where accuracy reaches the desired level.

For systems of this kind, it is at least *conceivable* that RBMT may be able to outperform SMT. The next question is how to investigate the issues in a methodologically even-handed way. A few studies, notably (Seneff et al., 2006), suggest that rule-based translation may in fact be preferable in these cases. (Another related experiment is described in (Dugast et al., 2008), though this was carried out in a large-vocabulary system). These studies, however, have not been widely cited. One possible explanation is suspicion about methodological issues. Seneff and her colleagues trained their SMT system on 20 000 sentence pairs, a small number by the standards of SMT. It is *a priori* not implausible that more training data would have enabled them to create an SMT system that was as good as, or better than, the rule-based system.

In this paper, our primary goal is to take this kind of objection seriously, and develop a methodology designed to enable a tight comparison between rule-based and statistical architectures. In particular, we wish to examine the widely believed claim that SMT is now inherently better than RBMT. In order to do this, we start with a limited-domain RBMT system; we use it to automatically generate a large corpus of aligned pairs, which is used to train a corresponding SMT system. We then compare the performance of the two systems.

Our argument will be that this situation essentially represents an upper bound for what is possible using the SMT approach in a limited domain. It has been widely remarked that quality, as well as quantity, of training data is important for good SMT; in many projects, significant effort is expended to clean the original training data. Here, since the data is automatically generated by a rule-based system, we can be sure that it is already completely clean (in the sense of being internally consistent), and we can generate as large a quantity of it as we require. The application, moreover, uses only a smallish vocabulary and a fairly constrained syntax. If the derived SMT system is unable to match the original RBMT system's performance, it seems reasonable to claim that this shows that there are types of applications where RBMT architectures are superior.

The experiments described have been carried out using MedSLT, an Open Source interlingua-based limited-domain medical speech translation system. 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., 2005; Bouillon et al., 2008b) 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). Vocabulary sizes and performance measures for speech recognition in the three languages where serious evaluations have been carried out are shown in Figure 1.

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 MedSLT 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.

At run-time, the recogniser produces a source-langage 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 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 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 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 the two MedSLT language pairs English → French and English → Japanese. 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 finally used the translation rules and target-language generators to attempt to translate each sentence. This created approximately 305K aligned sentence-pairs for English → French (1901K words English, 1993K words French), and 311K aligned sentence-pairs for English → Japanese (1941K words English, 2214K words Japanese). We held out 2.5% of each set 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.

Our primary goal was to 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. There is a simple and natural way to perform this measurement: we apply 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 BLEU score of 1.0.

This raises an important point. The BLEU scores we are using here are non-standard; they measure the extent to which the SMT approximates the RBMT, rather than, as usual, measuring

English	Is the pain above your eye?
French	Avez-vous mal au dessus des yeux?
Japanese	Itami wa me no ue no atari desu ka?
English	Have you had the pain for more than a month?
French	Avez-vous mal depuis plus d'un mois?
Japanese	Ikkagetsu ijou itami wa tsuzuki mashita ka?
English	Is the pain associated with nausea?
French	Avez-vous des nausées quand vous avez la douleur?
Japanese	Itamu to hakike wa okori masu ka?
English	Does bright light make the pain worse?
French	La douleur est-elle aggravée par une lumière forte?
Japanese	Akarui hikari wo miru to zutsu wa hidoku nari masu ka?

Table 2: Examples of English domain sentences, and the system's translations into French and Japanese.

the extent to which it approximates human translations. It is important to bring in human judgement, 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 difference would be purely academic. 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 where the two translations were different. We randomly selected a set of examples for each language pair, 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.

Finally, we consider the question of whether the SMT, which is capable of translating out-of-grammar sentences, can add useful robustness to the base system. We collected, from the set used in the experiments described in (Rayner et al., 2005), all the English sentences which failed to be translated into French. We used the best version of the English → French SMT to translate each of these sentences, and asked human judges to evaluate the translations as being clearly acceptable, clearly unacceptable, or borderline.

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

4 Results

We begin with Figure 1, which shows non-standard BLEU scores for versions of the English → French SMT system trained on quantities of data increasing from 14 287 to 285 740 pairs. As can be seen, translation performance improves up to about 175 000 pairs. After this, it levels out at around BLEU = 0.90, well below that of the RBMT system with which it is being compared. A more direct way to report the result is simply to count the proportion of test sentences that are not in the training data, which are translated similarly by the SMT and the RBMT. This figure tops out at around 68%.

The results strongly suggest that the SMT is unable to replicate the RBMT's performance at all closely even in an easy language-pair, irrespective of the amount of training data available. Out of curiosity, and to reassure ourselves that the automatic generation procedure was doing something useful, we also tried training the English → French SMT on pairs derived from the 669 ut-

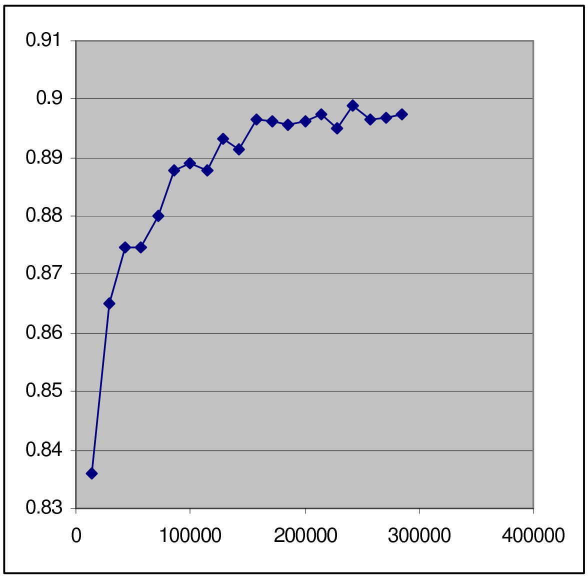

Figure 1: Non-standard BLEU scores against number of pairs of training sentences for English → French; training and test data both independently generated, hence overlapping.

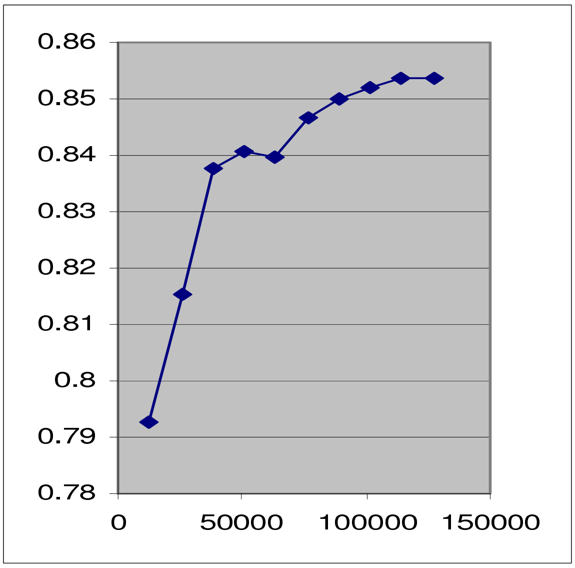

Figure 2: Non-standard BLEU scores against number of pairs of training sentences for English → French; training and test data both independently generated, then uniqued to remove duplicates and overlapping items.

terance "seed corpus" used to generate the grammar (cf. Section 2). This produced utterly dismal performance, with BLEU = 0.52. The result is more interesting than it may first appear, since, in speech recognition, the difference in performance between the SLMs trained from seed corpora and large generated corpora is fairly small (Hockey et al., 2008).

It seemed possible that the improvement in performance with increased quantities of training data might, in effect, only be due to the SMT functioning as a translation memory; since training and test data are independently generated by the same random process, they overlap, with the degree of overlap increasing as the training set gets larger. In order to investigate this hypothesis, we repeated the experiments with data which had been uniqued, so that the training and test sets were completely disjoint, and neither contained any duplicate sentences[1]. In fact, Figure 2 show that the graph for uniqued English → French data are fairly similar to the one for the original non-uniqued data shown in Figures 1. The main difference is that the non-standard BLEU score for the uniqued data, unsurprisingly, tops out at a lower level, reflecting the fact that a "translation memory" effect does indeed occur to some extent.

Results for English → Japanese showed the same trends as English → French, but were more pronounced. Table 3 compares the performance of the best versions of the SMTs for the two language-pairs, using both plain and artificially uniqued data. We see that, with plain data, the English → Japanese SMT falls even further short of replicating the performance of the RBMT than was the case for English → French; BLEU is only 0.76. The difference between the plain and uniqued versions is also more extreme. BLEU (0.64) is considerably lower for the version trained on uniqued data, suggesting that the SMT for this language pair is finding it harder to generalise, and is in effect closer to functioning as a translation memory. This is confirmed by counting the sentences in test data and not in training data which were translated similarly by the SMT and the RBMT; we find that the figure tops out at the very low value of 26%.

As noted in our discussion of the experimental framework, the non-standard BLEU scores only address the question of whether the performance of the SMT and RBMT systems is the same. It is

[1] Our opinion is that this is *not* a realistic way to evaluate the performance of a small-vocabulary system; for example, in MedSLT, one expects that at least some training sentences, e.g. "Where is the pain?", will also occur frequently in test data.

Training data	Test data	BLEU
English → French		
Generated	Generated	0.90
Gen/uniqued	Gen/uniqued	0.85
English → Japanese		
Generated	Generated	0.76
Gen/uniqued	Gen/uniqued	0.64

Table 3: Translation performance, in terms of non-standard BLEU metric, for different configurations, training on all available data of the specified type. "Generated" = data randomly generated; "Gen/uniqued" = data randomly generated, then uniqued so that duplicates are removed and test and training pairs do not overlap.

necessary to establish what the differences mean in terms of human judgements. We consequently turn to evaluation of the pairs for which the SMT and the RBMT systems produced different translation results.

Table 4 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 among the judges 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 5 presents typical examples of bad SMT translations in the English → French pair, contrasted with the translations 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.

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

Table 6 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.

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

Finally, we look at the performance of the SMT on material which the RBMT is not able to translate. This would seem to be a situation where

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

the SMT could have an advantage; robustness is generally a strength of statistical approaches. We return to English → French in Table 7, which presents the result of running the best SMT model on the 357 examples from the test set in (Rayner et al., 2005) which failed to be translated by the RBMT. We divide the set into categories based on the reason for failure of the RBMT.

In the most populous group, translations that failed due to out of vocabulary items, the SMT was, more or less by construction, also unable to produce a translation. For the 110 items that were out of grammar coverage for the RBMT, the SMT produced 38 good translations, and another 4 borderline translations. There were 50 items that were within the source grammar coverage of the RBMT, but failed somewhere in transfer and generation processing. Of those, the majority (32) represented "bad" source sentences, considered as ill-formed for the purposes of this experiment. Out of the remaining items that were within RBMT grammar coverage, the SMT managed to produce 5 good translations and 1 borderline translation. In total, on the most lenient interpretation, the SMT produced 48 additional translations out of 357. While this improvement in recall is arguably worth having, it would come at the price of a substantial decline in precision.

Result	Count
Out of vocabulary	
Bad translation	187
Out of source grammar coverage	
Good translation	38
Bad translation	44
Borderline translation	4
Bad source sentence	34
In source grammar coverage	
Good translation	5
Bad translation	12
Borderline translation	1
Bad source sentence	32
Total	357

Table 7: English → French SMT performance on examples from the test set which failed to be translated by the RBMT, evaluated by one judge.

5 Discussion and Conclusions

We have presented a novel methodology for comparing RBMT and SMT, and tested it on a specific pair of RBMT and SMT architectures. Our claim is that these results show that the version of SMT used here is *not* in fact capable of reproducing the output of the RBMT system. Although there has been some interest in attempting to train SMT systems from RBMT output, the evaluation issues that arise when comparing SMT and RBMT versions of a high-precision limited-domain system are different from those arising in most MT tasks, and necessitate a correspondingly different methodology. It is easy to gain the impression that it is unsound, and that the experiment has been set up in such a way that only one result is possible. This is not, in fact, true.

When we have discussed the methodology with people who work primarily with SMT, we have heard two main objections. The first is that the SMT is being trained on RBMT output, and hence can only be worse; a common suggestion is that a system trained on human-produced translations could yield better results. It is not at all implausible that an SMT trained on this kind of data might perform better on material which is outside the coverage of the RBMT system. In this domain, however, the important issue is precision, not recall; what is critical is the ability to translate accurately on material that is within the constrained language defined by the RBMT coverage. The RBMT engine gives very good performance on in-coverage data, as has been shown in other evaluations of the MedSLT system, e.g. (Rayner et al., 2005); over 97% of all in-coverage sentences are correctly translated. Human-generated translations would often, no doubt, be more natural than those produced by the RBMT, and there would be slightly fewer outright mistranslations. But the primary reason why the SMT is doing badly is not that the training material contains bad translations, but rather that the SMT is incapable of correctly reproducing the translations it sees in the training data. Even in the easy English → French language-pair, the SMT often produces a different translation from the RBMT. It could *a priori* have been conceivable that the differences were uninteresting, in the sense that SMT outputs different from RBMT outputs were as good, or even better. In fact, Table 4 show that this is not true; when the two translations differ, although the SMT translation can occasionally be better, it is usually worse. Table 6 shows that this problem is considerably more acute in English → Japanese. Thus the SMT system's inability to model the RBMT system points to a real limitation.

If the SMT had instead been trained on human-generated data, its performance on in-coverage material could only have improved substantially if the SMT for some reason found it easier to learn to reproduce patterns in human-generated data than in RBMT-generated data. This seems unlikely. The SMT is being trained from a set of translation pairs which are guaranteed to be completely consistent, since they have been automatically generated by the RBMT; the fact that the RBMT system only has a small vocabulary should also work in its favour. If the SMT is unable to reproduce the RBMT's output, it is reasonable to assume it will have even greater difficulty reproducing translations present in normal human-generated training data, which is always far from consistent, and will have a larger vocabulary.

The second objection we have heard is that the non-standard BLEU scores which we have used to measure performance use the RBMT translations as a reference. People are quick to point out that, if real human translations were scored in this way, they would do less well on the non-standard metrics than the RBMT translations. This is, indeed, absolutely true, and explains why it was essential to carry out the comparison judging shown in Tables 4 and 6. If we had compared human translations with RBMT translations in the same way, we would have found that human translations which differed from RBMT translations were sometimes better, and hardly ever worse. This would have shown that the non-standard metrics were inappropriate for the task of evaluating human translations. In the actual case considered in this paper, we find a completely different pattern: the differences are one-sided in the opposite direction, indicating that the non-standard metrics do in fact agree with human judgements here.

A general objection to all these experiments is that there may be more powerful SMT architectures. We used the Giza++/Moses/SRILM combinination because it is the *de facto* standard. We have posted the data we used at `http://www.bahrc.net/geaf2009`; this will allow other groups to experiment with alternate architectures, and determine whether they do in fact yield significant improvements. For the moment, however, we think it is reasonable to claim that, in domains where high accuracy is required, it remains to be shown that SMT approaches are capable of achieving the levels of performance that rule-based systems can deliver.

References

D. Arnold, L. Balkan, S. Meijer, R.L. Humphreys, and L. Sadler. 1994. *Machine Translation: An Introductory Guide*. Blackwell, Oxford.

P. Bouillon, M. Rayner, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, Y. Nakao, K. Kanzaki, and H. Isahara. 2005. A generic multilingual open source platform for limited-domain medical speech translation. In *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT)*, pages 50–58, Budapest, Hungary.

P. Bouillon, F. Ehsani, R. Frederking, and M. Rayner, editors. 2006. *Proceedings of the HLT-NAACL International Workshop on Medical Speech Translation*, New York.

P. Bouillon, F. Ehsani, R. Frederking, M. McTear, and M. Rayner, editors. 2008a. *Proceedings of the COLING Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*, Manchester.

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. 2008b. 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.

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.

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.

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. Bouillon, G. Flores, F. Ehsani, M. Starlander, B. A. Hockey, J. Brotanek, and L. Biewald. 2008. A small-vocabulary shared task for medical speech translation. In *Proceedings of the COLING Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*, Manchester.

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.