Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, February 2012

Bootstrapping a Statistical Speech Translator From a Rule-Based One

Manny Rayner
Pierrette Bouillon
Paula Estrella
Yukie Nakao
Gwen Christian

Submitted, May 2011
Revised, February 2012
Published by CSLI Publications

LiLT volume 5, issue 3 February 2012

Bootstrapping a Statistical Speech Translator From a Rule-Based One

Manny Rayner, *Geneva University,* Pierrette Bouillon, *Geneva University,* Paula Estrella, *FaMAF, U. Nacional de Córdoba,* Yukie Nakao, *Département Communication, Langues & Entreprise, Ecole Centrale de Nantes,* Gwen Christian

Abstract

We describe a series of experiments in which we start with English to French and English to Japanese versions of a rule-based speech translation system for a medical domain, and bootstrap corresponding statistical systems. Comparative evaluation reveals that the statistical systems are still slightly inferior to the rule-based ones, despite the fact that considerable effort has been invested in tuning both the recognition and translation components; however, a hybrid system is able to deliver a small but significant improvement in performance. In conclusion, we suggest that the hybrid architecture we describe potentially allows construction of limited-domain speech translation systems which combine substantial source-language coverage with high-precision translation.

LiLT Volume 5, Issue 3, February 2012.
Bootstrapping a Statistical Speech Translator From a Rule-Based One.

1 Introduction

This paper describes a series of experiments centered around MedSLT (Bouillon et al., 2008a), a small/medium-vocabulary medical speech translator designed for doctor-patient communication which uses a rule-based architecture; the purpose of the experiments has been to compare this architecture with more mainstream statistical ones. The original motivation for using rule-based methods comes from considerations regarding the tradeoff between precision and recall. Specifically, medical speech translation is a safety-critical domain, where precision is much more important than recall. It is also important to note that this is a domain where substantial quantities of training data are unavailable. The question is how to use the very limited amounts of data at our disposal to best effect. This is by no means an uncommon scenario in limited-domain speech translation, and could in fact be regarded as the norm rather than the exception.

It is intuitively not unreasonable to believe that rule-based methods are better suited to the requirements outlined above, but the well-known methodological problems involved in performing comparisons between rule-based and statistical systems have made it hard to establish this point unambiguously. In an earlier study (Rayner et al., 2005), we presented head-to-head comparisons between MedSLT and an alternative which combined statistical recognition and an ad hoc translation mechanism based on hand-coded surface patterns, showing that the rule-based system performed comfortably better. It was, however, clear from informal comments we received that other researchers in the field viewed these results sceptically. The basic criticism was that the robust processing components were too much of a straw-man: more powerful recognition or translation engines might conceivably have reversed the result.

In the new series of experiments, our basic goal has been to start with the rule-based components and the corpus data used to construct them, and then use the same resources, together with mainstream tools, to bootstrap statistical processing components. In (Hockey et al., 2008), we adapted and improved methods originally described in (Jurafsky et al., 1995) to bootstrap a statistical recogniser from the original rule-based one. More recently, in (Rayner et al., 2010) we used similar methods to bootstrap statistical machine translation models.

In the current paper, we combine the results of the previous two sets of experiments to build a fully bootstrapped statistical speech translation system, which we then compare with the original rule-based one, and also with a hybrid system which combines rule-based and statisti-

cal processing. The rest of the paper is organised as follows. Section 2 presents background on the MedSLT system; Section 3 summarises the earlier experiments on bootstrapped statistical recognition, and Section 4 those on machine translation; Section 5 describes the new experiments; and Section 6 concludes.

2 Background: the MedSLT System

MedSLT (Bouillon et al., 2008a) is a medium-vocabulary interlingua-based Open Source[1] 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}. In what follows, however, we will only be concerned with the pairs English → French and English → Japanese, which we take, respectively, as representative of a close and distant language-pair. All the experiments described were carried out using the 870-utterance recorded speech corpus from (Rayner et al., 2005); this was collected using a protocol in which subjects played the doctor role in simulated medical examinations carried out using the MedSLT prototype. A transcribed version of the data can be found online at `http://medslt.cvs.sourceforge.net/viewvc/*checkout*/medslt/MedSLT2/corpora/acl_2005_transcriptions.txt?revision=1.1`. A brief examination of the corpus shows that it is fairly noisy. We estimate that about 65–70% of it consists of clearly in-domain and well-formed sentences, depending on the exact definitions of these terms[2], with much of the remaining portion being out-of-domain or dysfluent.

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.

[1] LGPL license; `https://sourceforge.net/projects/medslt/`

[2] 61% of the corpus is within the coverage of the current English grammar.

Both speech recognition and translation are rule-based. Speech recognition runs on the commercial Nuance 8.5 recognition platform, with grammar-based language models built using the Open Source[3] 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.

The run-time architecture is summarised in Figure 1. At the start of the processing chain, the recogniser produces a set of N-best speech hypotheses. Each of these is parsed, using the source-language grammar, into a source-language semantic representation in AFF (Almost Flat Functional Semantics; (Rayner et al., 2008)), a type of key-value formalism that we will shortly describe in more detail. Since the source-language grammar has also been used to build the recogniser's language model, all recognition hypotheses are guaranteed to be within its coverage.

The set of AFF source-language representations is translated by a set of rules into corresponding interlingual forms, again represented in AFF. The space of well-formed interlingua representations in MedSLT is also defined by a Regulus grammar (Bouillon et al., 2008a); this grammar is designed to have minimal structure, so checking for well-formedness can be performed very quickly, and hypotheses which give rise to non-wellformed 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 (Bouillon et al., 2008b). After rescoring, the top interlingua representation in the rescored list is translated by a second set of rules into a target language AFF representation. 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.

In parallel, the interlingua is also translated, using the same methods, into the source-language ("backtranslated"). The backtranslation is shown to the source-language user, who has the option of aborting processing if they consider that speech understanding has produced

[3]LGPL license; `https://sourceforge.net/projects/regulus/`

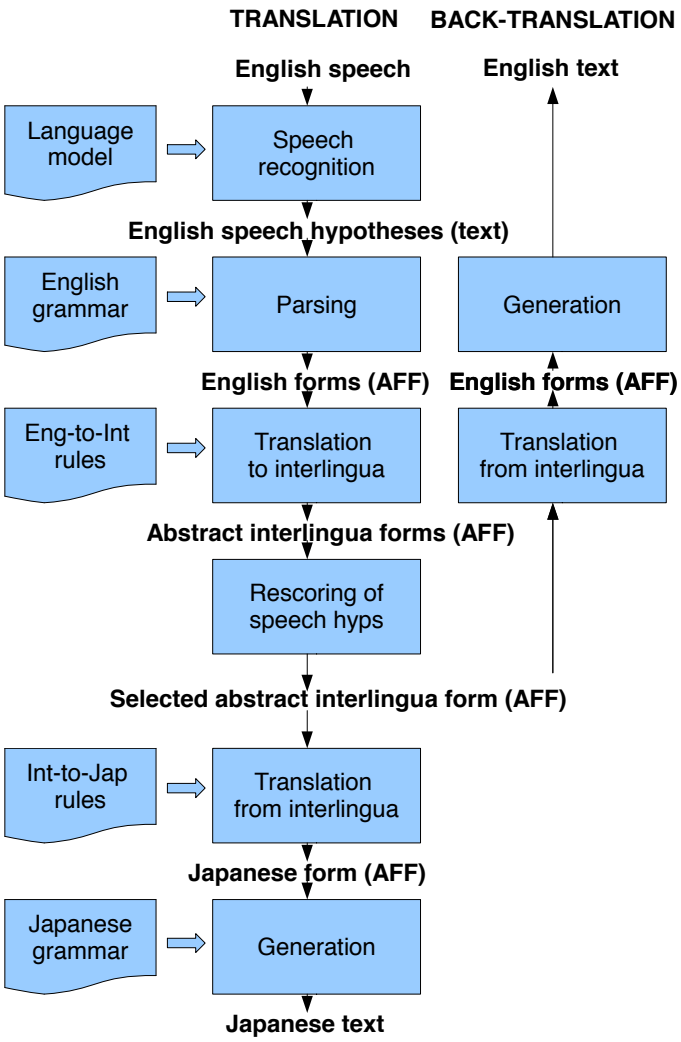

FIGURE 1 Run-time processing in the initial (fully rule-based) version of the MedSLT system. For concreteness, we consider the case where English is the source language and Japanese the target.

an incorrect result. If they do not abort, the target language string is displayed and realised as spoken output. This mode of operation is absolutely essential in a safety-critical application like medical examination. Since translation errors can have serious or even fatal consequences, doctors will only consider using systems with extremely low error rates, where they can directly satisfy themselves that the system has at least correctly understood what they have said before attempting to translate it. This also motivates use of restricted-domain, as opposed to general translation.

The interlingua grammar is built in such a way that the surface forms it defines can be used as human-readable glosses, which means that the interlingua takes on many of the characteristics of an artificial pivot language. We will make heavy use of this idea in what follows. The usual form of the "interlingua pivot language" is modelled on English. It is, however, straightforward to parametrize the grammar so that the pivot language form can also be generated with word-orders based on those occurring in other languages; in particular, we will also use one based on Japanese.

Table 2 shows examples of English domain sentences together with translations into French and Japanese and pivot language representations in English-based and Japanese-based format. Note the very simple structure of the pivot language, 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 the following structure:[4]

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

The English-oriented pivot form, "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-oriented form, "more-than one day duration pain usually last PRESENT YN-QUESTION" makes concessions to standard Japanese word-order, in which the sen-

[4] AFF representations and pivot language forms have been slightly simplified for presentational reasons.

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

TABLE 2 English MedSLT examples: English source sentence, English-oriented interlingua pivot form, rule-based translation into French, Japanese-oriented interlingua pivot form and rule-based translation into Japanese.

tence 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-oriented pivot form puts "sc-when" ("subordinating-conjunction when") before the representation of the subordinate clause; the Japanese-oriented pivot form 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-oriented pivot form suppresses the personal pronoun "you", again following normal Japanese usage. In Section 4, we will demonstrate how useful the different forms of the interlingua turn out to be. The basic point is to be able to split up statistical translation into pieces where source and target always have similar word-order.

The next two sections presents the results of earlier experiments, in

which statistical components were bootstrapped by using the rule-based ones to create training data.

3 Bootstrapping statistical language models

As described in Section 2, the Regulus platform constructs grammar-based language models in a corpus-driven way. This, in principle, enables a fair comparison between grammar-based language modelling (GLM) and statistical language modelling (SLM), since the "seed corpus" used to extract the specialised grammar can also be used to train an SLM. There are, however, several ways to implement the idea. The simplest method is to use the seed corpus directly as a training corpus for the SLM. A more subtle approach is described in (Jurafsky et al., 1995, Jonson, 2005); one can randomly sample the grammar-based language model to generate arbitrarily large amounts of corpus data, which are then used as input to the SLM training process. In (Hockey et al., 2008), we showed that a statistical recogniser trained from a suitable version of a randomly generated corpus outperformed the one generated from the 948-utterance seed corpus.

Table 3 summarises the main results, contrasting different methods for building the SLM training corpora; the first line, for the GLM built using the "seed" corpus, is intended to provide a reference point. Line 2 shows the SLM built from the "seed" corpus. The other recognisers were all built from GLM-generated training corpora of the same size. The versions vary in three ways. The training corpora are of different sizes; we generate them using either plain CFG generation or PCFG generation, where the probabilistic weights attached to the CFG grammar are obtained by using the seed corpus a second time; finally, we may or may not use the interlingua to filter the results of generation.

We look first at lines 3 to 6. The corpus used in line 4 was created by starting with an initial CFG-generated set of 500K utterances, and applying interlingua-based filtering; the fact that more than 99% of the data is discarded shows that plain CFG generation produces very low-grade data. For purposes of comparison, line 3 contains results for a corpus created from an equally large sample of unfiltered CFG-generated data. Evidently, although filtering helps, CFG generation does not deliver interesting performance; results are much worse than with the seed corpus. In contrast, lines 5 and 6 show results for PCFG-generated corpora, which, at least in terms of WER, are better than the seed corpus. All the remaining experiments were consequently performed with PCFG-generated data.

When SLMs are trained on human-generated data, performance usu-

	Version	size	WER	SER
1	seed corpus GLM	948	21.96%	50.62%
2	seed corpus SLM	948	27.74%	58.40%
3	CFG/unfiltered	4 281	49.0%	88.4%
4	CFG/filtered	4 281	44.68%	85.68%
5	PCFG/unfiltered	4 281	25.98%	65.31%
6	PCFG/filtered	4 281	25.81%	63.70%
7	PCFG/unfiltered	16 619	24.84%	62.47%
8	PCFG/filtered	16 619	23.80%	59.51%
9	PCFG/unfiltered	497 798	24.38%	59.88%
10	PCFG/filtered	497 798	23.76%	57.16%

TABLE 3 Recognition performance for SLMs trained on different types of generated data. "Size" = number of utterances in training set; "WER" = Word Error Rate on test set of in-coverage and out of coverage material; "SER" = sentence error rate on test set of in-coverage and out of coverage material. GLM results included for comparison

	Version	size	WER	SER
1	seed corpus GLM	948	7.00%	22.37%
2	seed corpus SLM	948	14.40%	42.02%
3	PCFG/unfiltered	16 619	14.13%	46.11%
4	PCFG/filtered	16 619	12.76%	40.86%
5	1500K PCFG/unfiltered	497 798	12.35%	40.66%
6	1500K PCFG/filtered	497 798	11.25%	36.19%

TABLE 4 Recognition performance as training set size increases, on in-coverage material only. "Size" = number of utterances in training set; "WER" = Word Error Rate; "SER" = sentence error rate

ally improves for some time as more data is added. A common rule of thumb when building commercial SLM-based systems is that one should aim to collect about 20 000 utterances. Lines 7 to 10 presents results for SLMs trained off PCFG-generated corpora of increasing size. In order for the filtered and unfiltered SLMs to be trained from similar amounts of data, the unfiltered data file was truncated to match the number of items left in the filtered file. So, for example, when 50K of data was generated, 16 619 items were left after filtering, and the first 16 619 items of the unfiltered file were used for training the unfiltered SLM. The amount of training data was incremented until addition of training data no longer resulted in an improvement in the error rates.

The recognisers trained on filtered data continued to improve as we increased the size of the training set. The best recogniser trained on unfiltered data had lower WER than the "seed corpus" SLM recogniser. SER, however, was almost the same between these two versions, and the difference was not significant[5]. The best recogniser trained on filtered data did better, and outscored the "seed corpus" SLM on both WER and SER. The difference on SER, however, was again not significant.

One methodological problem with the above figure is that comparisons between GLM and SLM models depend heavily on the mix of in-coverage and out of coverage data encountered in the test data. Performance of both models is generally dismal on out-of-coverage data, and consequently not very interesting; performance on in-coverage data is often a more useful metric. Table 4 summarises performance on the in-coverage subset of the data. Two points are worth noting. First, as expected, restriction to in-coverage data increases the difference between the GLM recogniser and the others in terms of both WER and SER; for both metrics, we see a relative decrease of over 35% between results for the GLM and the best of the other versions. The second point, rather more interestingly, is that the best SLM version is now the one created from filtered PCFG-generated data (line 6). This version is significantly better than the "seed corpus" SLM.

To sum up the argument for this section, the methodology of bootstrapping an SLM recogniser by using the GLM to generate more corpus data does succeed in significantly improving recognition performance. Even the best SLM version is, however, still inferior to the GLM, and substantially inferior to it on in-coverage data.

[5]All significance results in this section are in terms of the McNemar sign test. The details are presented in (Hockey et al., 2008).

4 Bootstrapping statistical translation models

In (Rayner et al., 2009, 2010), we adapted the methods from Section 3 to bootstrap Statistical Machine Translation (SMT) models from the original rule-based ones; a similar experiment, with a large-vocabulary system, is reported in (Dugast et al., 2008). As above, we started by using the source-language grammar to randomly generate a large corpus of data. We then passed the result through the interlingua-based translation components to create bicorpora for these language pairs; in the experiments reported here, we used English → French and English → Japanese versions of the system, representing a close and a distant translation pair respectively. We then used the resulting bicorpora to train SMT models with the common combination of Giza++, Moses and SRILM (Och and Ney, 2000, Koehn et al., 2007, Stolcke, 2002), training the models with MERT (Och, 2003) on a held-out portion of data

The obvious way to create the SMT models, explored in (Rayner et al., 2009), is simply to use the aligned source/target corpus as training data. As the paper shows, this gives surprisingly poor performance. We experimented with various ways to evaluate translation quality for the bootstrapped SMT, and found that a simple and effective one was to determine how often the SMT's translation differed from the original RBMT's, on examples which were not in the training data; human judges confirmed that the differences were hardly ever in the SMT's favour, and frequently showed up errors. (Differences pointed to SMT errors in about 40 to 60% of the examples for English → French, and about 60 to 80% of the examples for English → Japanese, depending on the strictness of the judge). On the "agreement" metric, performance topped out at around 67% agreement for English → French, and a dismal 27% for English → Japanese; addition of more generated training data failed to improve the results.

The unimpressive figures are perhaps not as strange as they first appear. Most SMT performance results are on tasks where vocabulary is much larger, but where demands on translation quality are also much lower. A translation which would be quite acceptable in Google Translate would often be completely unacceptable in the context of the safety-critical medical speech translation task. None the less, we felt that there had to be some way to get better performance. In (Rayner et al., 2010), we showed that it was possible to do this by once again exploiting the "pivot language" forms described in Section 2, which provide a text form of the interlingua. This allows us to construct aligned corpora which pair source or target sentences with pivot lan-

guage forms, and train separate SMT models for translation from source language to pivot language, and from pivot language to target language.

Splitting up SMT translation to make it go through the interlingua pivot language 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: as described earlier, each sentence produced by the speech recogniser is first translated into the interlingua, and then "backtranslated" into the source language, so that the user can if necessary abort the translation before a target language sentence is produced. If SMT is performed in two stages, using the text form of the interlingua as a pivot language, it is possible to employ the same basic architecture.

The interlingua/pivot language can also be exploited in other ways. First, if the SMT decoder is set to produce N-best output, we can use the interlingua/pivot grammar as a knowledge source to reorder N-best hypotheses as we do in recognition, 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 the English-oriented pivot language, then reformulate the result into a Japanese-ordered "Japlish", and finally use RBMT to generate Japanese.

In our concrete experiments, we created, as before, a randomly generated English corpus of 500K utterances, and passed it through English → French and English → Japanese versions of the RBMT system, storing the source, target and pivot language results. The pivot language was saved both in the English-oriented and the Japanese-oriented formats (cf. Table 2). We then trained separate SMT models for the pairs English → English-oriented pivot language, English-oriented pivot language → French, and Japanese-oriented pivot language → Japanese; for comparison purposes, we also trained models for English → French and English → Japanese. We experimented with several different ways of combining these resources, of which we finally used two. The first of these, which involves only statistical processing, was the following pipeline:

1. Translation from English to English-oriented pivot language using SMT, with the decoder set to produce N-best output (N was set to 15);
2. Rescoring of the N-best output to choose the highest well-formed string, where one was available;
3. If the target is Japanese, reformulation from English-oriented pivot language to Japanese-oriented pivot language;
4. Translation from the appropriate form of pivot language to the target language using SMT

We will call this pipeline **Bootstrapped-Statistical-SMT**. As shown in (Rayner et al., 2010), **Bootstrapped-Statistical-SMT** massively decreases the error rate for the difficult pair English → Japanese, compared to the naïve method of training a single SMT model. The key advantage is that SMT translation, which is very sensitive to differences in word-order, only has to translate between languages with similar word-orders. Even in the relatively easy pair English → French, a substantial performance gain was achieved by interposing the N-best rescoring step. On in-coverage input, both bootstrapped pivot-based SMT systems were able to reproduce the translations of the original rule-based systems on about 79% of the data; recall that the corresponding figures for the naïve method were 67% for English → French and 27% for English → Japanese.

Performance can be further improved if we move to a hybrid method, which we will call **Bootstrapped-Hybrid-SMT**, where the last step in the preceding pipeline is replaced by RBMT translation from the pivot to the target language. With this further enhancement, the bootstrapped system is able to reproduce the RBMT's translations on about 87% of the data for both target languages.

Though considerably improved, the bootstrapped SMT systems are thus still not quite as good as the original RBMT ones on in-coverage data. The payoff, of course, is that the bootstrapped systems are also able to translate out-of-coverage sentences. When evaluated on the out-of-coverage portion of the test set (358 text utterances), 81 sentences (23%) produced a backtranslation judged to be correct. Of these 81 sentences, 76 (94%) were judged to produce good translations for French, and 71 (88%) for Japanese when we used the pure bootstrapped statistical translation method. With the hybrid method, 75 of the 81 sentences which gave a good backtranslation also produced a French translation, and all of these translations were judged correct. For Japanese, all 81 sentences produced a translation, and 77 of these translations (95%) were judged correct.

5 Combining recognition and translation

The preceding sections have described how we were able to bootstrap good robust versions of the original speech recognition and machine translation components, using only the original, very small training set of 948 sentences. We now combine these modules to simulate a full bootstrapped statistical speech translation system and a hybrid version which combines rule-based and statistical processing.

Specifically, we took the best version of the bootstrapped statistical recogniser from Section 3 and the best versions of the pure bootstrapped statistical translation model (**Bootstrapped-Statistical-SMT**) and the hybrid bootstrapped statistical translation (**Bootstrapped-Hybrid-SMT**) from Section 4. We first ran the 870-utterance speech corpus from (Rayner et al., 2005) through the bootstrapped statistical recogniser, and then passed the results through both the bootstrapped statistical translation models. We define the three notional speech translation systems we will compare as follows. **Rule-Based** (Figure 1) is the original rule-based system, i.e. rule-based recognition followed by rule-based translation. The **Full-Bootstrapped-Statistical** system (Figure 2) is the composition of the bootstrapped statistical recogniser and **Bootstrapped-Statistical-SMT**. **Full-Bootstrapped-Hybrid** (Figure 3) is a system which uses **Rule-Based** if this produces a non-null translation, and otherwise backs off to the composition of the bootstrapped statistical recogniser and **Bootstrapped-Hybrid-SMT**.

For all three configurations, we also produced rule-based backtranslations (cf. Section 2), in order to be able to simulate normal use of the system. The material was annotated by human judges as follows. The English → English backtranslations were evaluated by two native English judges; they were asked to mark the backtranslation as good if they were sufficiently sure of its correctness that they would have considered, in a real medical examination dialogue, that the system had understood and should be allowed to pass its translation on to the patient. The English → French and English → Japanese translations were evaluated by two native speakers of French and two native speakers of Japanese respectively, who were all fluent in English. They were presented with a spreadsheet containing three columns, in which the first column was the source English sentence, and the other two were the output of the orginal rule-based system and the output of the bootstrapped system. If one of the systems produced no output, for whatever reason, this was marked as "NO TRANSLATION". The order of presentation of the two systems was randomised, so that the judge did not know, for any given line, which version was shown in the

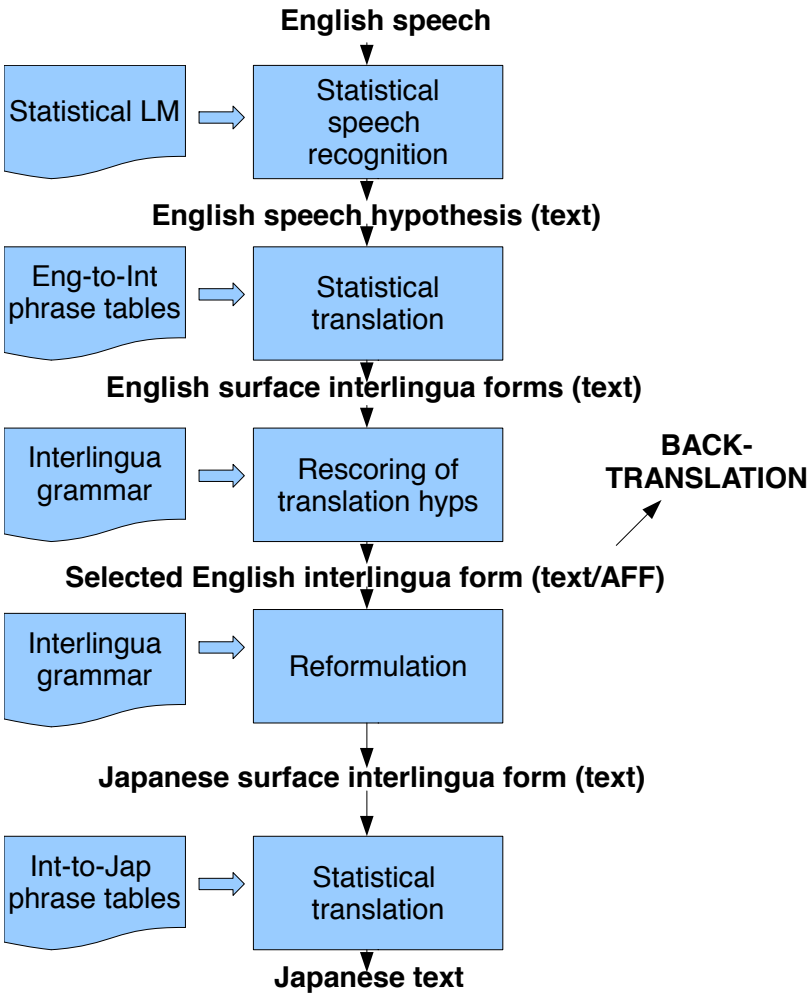

FIGURE 2 Run-time processing in the **Full-Bootstrapped-Statistical** version of the English-to-Japanese system. The statistical language model used for speech recognition and the two sets of phrase tables used for statistical translation are all bootstrapped from the correspoding rule-based components. Note that the selected English interlingua form is considered both as text (for statistical translation to Japanese) and as AFF (for rule-based back-translation to English).

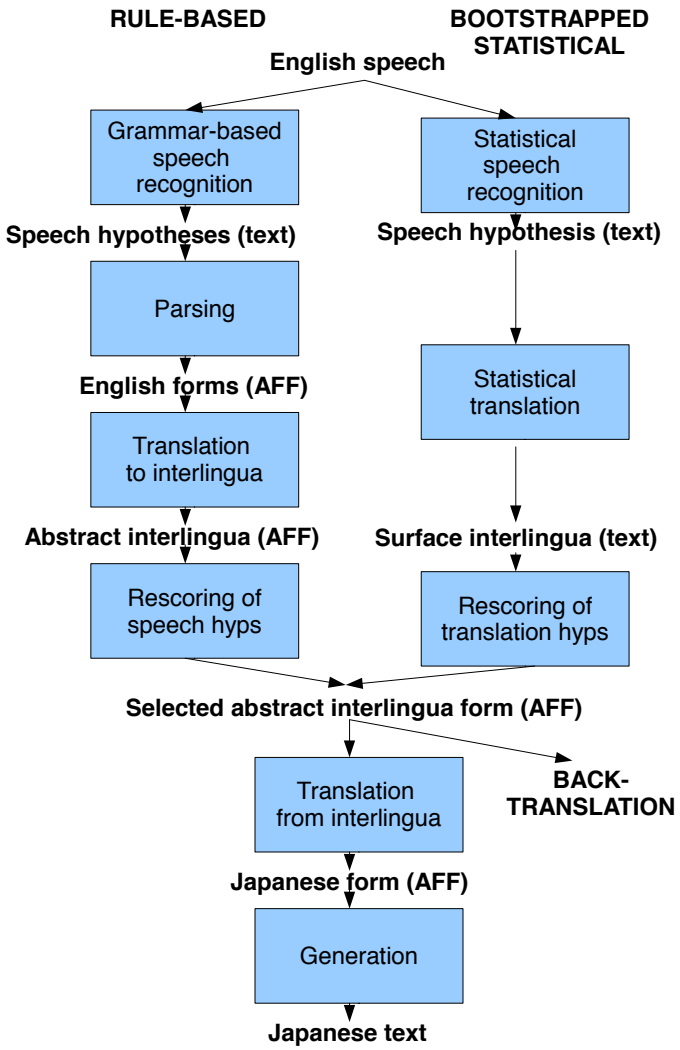

FIGURE 3 The **Full-Bootstrapped-Hybrid** version of the English-to-Japanese system combines modules from the **Rule-Based** version (Figure 1) and the **Full-Bootstrapped-Statistical** version (Figure 2). Run-time processing follows the Rule-Based path if that produces a well-formed interlingua form, otherwise backs off to the Full-Bootstrapped-Statistical path. In both cases, translation from interlingua to target is performed using Rule-Based modules.

second column and which in the third. If there were two translations, the judges were instructed to mark one of them if they considered that it was clearly superior to the other. If one of the translations was null they were instructed to mark the non-null translation as preferable if they considered that it would be useful in the context of the medical speech translation task.

We used the data and the judgments to compare **Full-Bootstrapped-Statistical** and **Full-Bootstrapped-Hybrid** against **Rule-Based**. The results are summarised in Table 5; we present figures for each comparison both on the complete dataset, and also on the subset for which backtranslation produced a result judged as good. The last three columns give the results first for each judge separately, then for the cases where the two judgements coincide.

	Comparison		Judged by		
	Version 1	Version 2	Judge1	Judge2	Agree
English → French (all data)					
1	Rules	Bootstrapped	**261–43**	**259–43**	**247–33**
2	Hybrid	Rules	**22–81**	**28–56**	**26–55**
English → French (good backtranslation only)					
3	Rules	Bootstrapped	**69–25**	**71–27**	**62–20**
4	Hybrid	Rules	**17–2**	**15–3**	**15–2**
English → Japanese (all data)					
5	Rules	Bootstrapped	125–98	**149–84**	**104–62**
6	Hybrid	Rules	**26–75**	**22–78**	**14–67**
English → Japanese (good backtranslation only)					
7	Rules	Bootstrapped	**61–25**	**64–30**	**48–18**
8	Hybrid	Rules	**12–1**	**15–2**	**10–0**

TABLE 5 Comparisons between different versions of the English → French and English → Japanese MedSLT systems. The result NN–MM indicates that the judge(s) in question considered that the first version gave a clearly better result NN times, and the second version a clearly better result MM times. Differences significant at $P < 0.05$ according to the McNemar test are marked in **bold**.

Although statistical processing, as usual, adds robustness, we can see that it suffers from two major problems. As lines 1 and 5 show, the statistical system, without backtranslation, is much worse than the rule-based one, since it frequently produces incorrect translations due to bad recognition. (The statistical system almost always produces a translation; the rule-based one fails to do so about on about 30% of

the data, since rule-based recognition most often fails altogether on out-of-coverage data, as opposed to producing a nonsensical result). With backtranslation added, lines 3 and 7 at least demonstrate that this first problem disappears, and the result is closer. However, we still have the second problem; there are long-distance dependencies which the statistical algorithms are unable to learn. For example, in French, both judges agreed that there were 55 cases where rule-based processing gave a better result than statistical, mostly due to more accurate recognition or translation. There were 26 cases which went the opposite way, with statistical processing better than rule-based: in most of these, rule-based processing gave no result, and statistical a good result. For both language pairs, the figures suggest that the lack of long-distance constraints is more important than the added robustness.

The hybrid system, however, *is* able to deliver a better result than the rule-based one when backtranslation is added; according to the McNemar test, the improvement is significant at $P < 0.01$. This positive result appears to depend heavily on the fact that the hybrid system uses rule-based translation to translate from the interlingua/pivot level, both for producing the backtranslation and the target translation. The statistical recognition and translation components add recall, but considerably reduce precision. By checking backtranslations, however, the user can catch cases where the statistical processing result has resulted in incorrect interlingua/pivot language, and be confident that the remaining examples will be correctly translated into the target language.

6 Summary and discussion

We have described a series of experiments in which we started with a rule-based speech translation system for a medical speech translation system, and used it to bootstrap a corresponding statistical system. We see two main conclusions, one positive and one negative. On the negative side, the statistical system is still inferior to the rule-based one, despite the fact that considerable ingenuity has been invested in tuning both the recognition and translation components. It is conceivable that a more subtle way of creating the statistical system might succeed in producing a system whose accuracy was comparable to that of the rule-based version. At the moment, though, the evidence at our disposal suggests that, if we are making a straight choice between rule-based and statistical, then rule-based systems are more appropriate for this kind of task.

We are aware that our conclusions are at odds with the currently prevailing wisdom; it is clear that some of our colleagues view our re-

sults with suspicion. When we have discussed the methodology with these people, there have been two main objections. The first is that large-vocabulary recognition, trained off general corpora, may give better recognition than the domain-specific recognition used here. This is possibly true, though our sentence error rate on in-domain utterances is in fact quite competitive. But the argument, even if correct, is basically irrelevant. Even when we filter to keep only examples with good backtranslations, thus in effect simulating near-perfect recognition, we find that the statistical system is at a severe disadvantage. The critical issues, in other words, lie in translation rather than recognition.

The second criticism we have received is that the SMT is being trained on RBMT output, and hence can only be worse; a common argument is that a system trained on human-produced translations should yield better results. It is entirely plausible that an SMT trained on this kind of data would perform better on material which is outside the coverage of the RBMT system. In our 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). Human-generated translations would sometimes, 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 dubious 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, this is not true; when the two translations differ, although the SMT translation can occasionally be better, it is usually worse. 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 accuracy 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.

As already noted, it is of course possible that some better way of doing SMT would resolve the technical issues that have emerged here, and allow us to train a more accurate model. Until this has been concretely demonstrated, however, the theoretical question of whether statistical systems may eventually be capable of outperforming rule-based ones on this kind of task seems to us rather uninteresting. We would prefer to focus on the positive result: the architecture we have developed here already provides a straightforward recipe for constructing hybrid limited-domain speech translation systems which are better than either type of pure system.

In the experiments described above, the statistical components are bootstrapped from the rule-based ones, and have only slightly more coverage; even so, they are able to make the hybrid system significantly more robust. It would be possible to use the same methods to combine a rule-based speech translation system with broad-coverage recognition and translation modules, as long as they are able to map source language into the interlingua, and this could result in a *substantial* increase in robustness. Since the interlingua has a text realisation as the pivot language, the construction is in principle quite straightforward; just as we have done in the current system, use of rule-based methods for backtranslation and translation to the target language means that the hybrid system can retain all the rule-based system's precision. There are some non-trivial problems — in particular, it is not immediately clear how to create the data needed to train a broad-coverage SMT module for the source language/pivot language pair — but if they can be solved it should be possible to create a very interesting type of application. We hope to investigate these questions further.

References

Bouillon, P., G. Flores, M. Georgescul, S. Halimi, B.A. Hockey, H. Isahara, K. Kanzaki, Y. Nakao, M. Rayner, M. Santaholma, M. Starlander, and N. Tsourakis. 2008a. 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.

Bouillon, P., S. Halimi, Y. Nakao, K. Kanzaki, H. Isahara, N. Tsourakis, M. Starlander, B.A. Hockey, and M. Rayner. 2008b. Developing non-European translation pairs in a medium-vocabulary medical speech translation system. In *Proceedings of LREC 2008*. Marrakesh, Morocco.

Dugast, L., 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.

Hockey, B.A., 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.

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

Jurafsky, A., 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.

Koehn, P., 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*, vol. 45, page 2.

Och, F.J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

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

Rayner, M., 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.

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

Rayner, M., P. Estrella, and P. Bouillon. 2010. A bootstrapped interlingua-based SMT architecture. In *Proceedings of the 14th Conference of the European Association for Machine Translation (EAMT)*. St Raphael, France.

Rayner, M., 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. Association for Computational Linguistics, Singapore.

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

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

Xu, Y. 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.