Bootstrapping A Statistical Speech Translator From A Rule-Based One

Manny Rayner
University of Geneva, TIM/ISSCO,
40 bvd du Pont-d'Arve
CH-1211 Genève, Switzerland
Emmanuel.Rayner@unige.ch

Paula Estrella
FaMAF, U. Nacional de Córdoba
5000 - Córdoba, Argentine
pestrella@famaf.unc.edu.ar

Pierrette Bouillon
University of Geneva, TIM/ISSCO,
40 bvd du Pont-d'Arve
CH-1211 Genève, Switzerland
Pierrette.Bouillon@unige.ch

Abstract

We describe a series of experiments in which we start with English → French and English → Japanese versions of an Open Source rule-based speech translation system for a medical domain, and bootstrap corresponding statistical systems. Comparative evaluation reveals that the rule-based systems are still significantly better than the statistical ones, despite the fact that considerable effort has been invested in tuning both the recognition and translation components; also, a hybrid system only marginally improved recall at the cost of a loss in precision. The result suggests that rule-based architectures may still be preferable to statistical ones for safety-critical speech translation tasks.

Index Terms: Speech translation, rule-based processing, statistical processing, bootstrapping, interlingua, evaluation

1 Introduction

This paper describes a continuation of a series of experiments centered around MedSLT (Bouillon et al., 2008a), an Open Source 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 comments we received that the community 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 pro-

cessing 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 this 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 statistical processing. Interestingly, although (Rayner et al., 2010) demonstrated that a bootstrapped statistical machine translation system is able to add substantial robustness to the original rule-based one when both are run on text data, this robustness does not carry over to speech translation.

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 machine translation; Section 4 describes the new experiments; and Section 5 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 $\rightarrow$ French and English $\rightarrow$ Japanese, which we take, respectively, as representative of a close and distant language-pair.

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[2] 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).

At run-time, the recogniser produces a source-language semantic representation in AFF (Almost Flat Functional Semantics; (Bouillon et al., 2008a)). 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.

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 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 space of well-formed interlingua representations in MedSLT is defined by yet another Regulus grammar (Bouillon et al., 2008a); this grammar is designed to have minimal structure, so checking for well-formedness can be performed very quickly. During speech recognition, 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-wellformed interlingua can safely be discarded. Use of

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

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

English	does the pain usually last for more than one day
Eng interlingua gloss	YN-QUESTION pain last PRESENT usually duration more-than one day
French	la douleur dure-t-elle habituellement plus d'un jour
Jap interlingua gloss	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 gloss	YN-QUESTION you have PRESENT ever pain sc-when you eat PRESENT
French	avez-vous déjà eu mal quand vous mangez
Jap interlingua gloss	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 gloss	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 gloss	head one side-part in-loc pain have PRESENT YN-QUESTION
Japanese	atama no katagawa wa itami masu ka

Table 1: English MedSLT examples: English source sentence, English-format interlingua gloss, rule-based translation into French, Japanese-format interlingua gloss and rule-based translation into Japanese.

this "highest-in-coverage" rescoring algorithm is found to reduce semantic error rate during speech understanding by about 10% relative (Bouillon et al., 2008b).

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; in particular, we will also use one based on Japanese.

Figure 1 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 the following structure:[3]

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

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

noun "you", again following normal Japanese usage. In Section 3.2, 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.

All the experiments described in the rest of the paper 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[4], with much of the remaining portion being out-of-domain or dysfluent.

The next section presents the results of earlier experiments, in which statistical components were bootstrapped by using the rule-based ones to create training data.

3 Previous experiments

3.1 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 and statistical language modelling, since the "seed corpus" used to extract the specialised grammar can also be used to train a statistical language model (SLM). There are, however, several ways to implement this 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 sufficiently large randomly generated corpus outperforms the one generated from the seed corpus[5]. A further refinement is to filter the randomly generated corpus by keeping only examples which, when translated into interlingua gloss form, result in well-formed representations. These improvements yielded a cumulative reduction in Word Error Rate, measured over the whole 870-utterance data set, from 27.7% to 23.6%. The best bootstrapped statistical recogniser was, however, still inferior to the grammar-based one, which scored 22.0%.

3.2 Bootstrapping statistical translation models

In (Rayner et al., 2010), we adapted the methods from Section 3.1 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 English $\rightarrow$ French and English $\rightarrow$ Japanese versions of the interlingua-based translation components, saving the source, target and interlingua gloss representations. A straightforward way to create the SMT models would be to use the aligned source/target corpora as training data. Here, however, we again showed that it was possible to get much better performance by exploiting the structure of the interlingua.

The interlingua gloss was saved both in the English-based and the Japanese-based formats (cf. Table 1). We then used the common combination of Giza++, Moses and SRILM (Och and Ney, 2000; Koehn et al., 2007; Stolcke, 2002) to train separate SMT models for the pairs English $\rightarrow$ English-format interlingua, English-format interlingua $\rightarrow$ French, and Japanese-format interlingua $\rightarrow$ Japanese; for comparison purposes, we also trained models for English $\rightarrow$ French and English $\rightarrow$ Japanese. All the models were tuned using MERT (Och, 2003) on a held-out portion of data. We experimented with several differ-

[4] 61% of the corpus is within the coverage of the current English grammar.

[5] The seed corpus used here contains 948 examples.

	Comparison			Judged by		
	Version 1	Version 2	Dataset	Judge1	Judge2	Unanimous
English → French						
1	Rule-based	Bootstrapped statistical	All data	**261–43**	**259–43**	**247–33**
2	Rule-based	Bootstrapped statistical	Only good backtrans.	**69–25**	**71–27**	**62–20**
3	Hybrid	Rule-based	All data	**29–180**	**30–181**	**25–177**
4	Hybrid	Rule-based	Only good backtrans.	18–12	19–15	15–12
English → Japanese						
5	Rule-based	Bootstrapped statistical	All data	125–98	**147–96**	**101–67**
6	Rule-based	Bootstrapped statistical	Only good backtrans.	**61–25**	**66–41**	**49–21**
7	Hybrid	Rule-based	All data	49–62	**30–81**	**23–55**
8	Hybrid	Rule-based	Only good backtrans.	17–8	19–9	14–8

Table 2: 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**.

ent ways of combining these resources. The best method turned out to be the following pipeline:

1. Translation from English to English-format interlingua 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-format interlingua to Japanese-format interlingua;
4. Translation from the appropriate format of interlingua to the target language using SMT

As shown in the paper, this combination 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 interlingua-based SMT systems were able to reproduce the translations of the original rule-based systems on about 79% of the data; the corresponding figures when the naïve method was used were 67% for English → French and 27% for English → Japanese. In cases where the bootstrapped SMT output differed from the RBMT one, hand-examination showed that the SMT version was hardly ever better, and was often worse (Rayner et al., 2009).

The bootstrapped SMT systems are thus not quite as good as the original RBMT ones on in-coverage data. The payoff, of course, is that the bootstrapped system 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.

4 Combining recognition and translation

The preceding sections have shown how we were able to use Open Source resources 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 describe how we combined these modules to compare a full bootstrapped statistical speech translation system against the orginal rule-based one; we also compare the rule-based system with a hybrid version which com-

bines rule-based and statistical processing.

We took the best versions of the bootstrapped statistical recogniser from Section 3.1 and the bootstrapped statistical translation models from Section 3.2, ran the 870-utterance speech corpus from (Rayner et al., 2005) through them, and compared the results with those obtained from the original speech translation system (grammar-based recognition and rule-based translation). In both 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 in the following way. The English $\rightarrow$ English backtranslations were evaluated by a native English judge; 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 $\rightarrow$ French and English $\rightarrow$ 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 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 the rule-based systems, the bootstrapped statistical systems, and a hypothetical hybrid system which produces the result from the bootstrapped system if the rule-based system produces no translation or no backtranslation, and otherwise produces the result from the rule-based system. The results are summarised in Table 2; 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.

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 2 and 6 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 62 cases where rule-based processing gave a better result than statistical, mostly due to more accurate recognition or translation. There were 20 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 results from (Rayner et al., 2010) led us to hope that the hybrid system would add robustness to the rule-based system without compromising accuracy; (Seneff et al., 2006) reports a similar result when the text component of a speech translation system is evaluated in isolation. Combination with the speech recognition front-end, with its concomitant noisy input, unfortunately appears to change the picture. Without backtranslation (lines 3 and 6), the hybrid system is inferior to the rule-based one for the reasons we have already seen.

When backtranslation is included (lines 4 and 8), we do indeed see a very small gain in re-

call, but this comes at the price of a substantial loss of precision. Examination of the cases where the rule-based system diverges from the hybrid one shows disturbing examples where the rule-based system produces no output, and the hybrid one an output which is meaningful but incorrect. For instance, "Do you take medicine for your headaches?" produced no translation in the rule-based English → French system, but *Avez-vous vos maux de tête quand vous prenez des médicaments?* ("Do you have headaches when you take medicine?") in the hybrid one; a mistake which would certainly worry any doctor who used the system!

5 Summary and conclusions

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. The rule-based system is still better than the statistical one, despite the fact that considerable ingenuity has been invested in tuning both the recognition and translation components.

The naïve hybrid system gave a small improvement in recall, but at an unacceptable cost in precision. It is conceivable that a more subtle way of creating the hybrid system may still succeed in adding useful robustness. At the moment, though, the evidence at our disposal suggests that rule-based systems are more appropriate for the kind of task, and that any gain from adding robust methods is at best likely to be small.

We are well aware that our result is at odds with the currently prevailing wisdom, namely that statistical methods are preferable to rule-based ones, and the obvious question is why this should be. We think there are two main reasons. First, most academic papers are written about systems that have been created to address a shared task. These tasks typically use large training sets that represent a substantial investment in time and effort. When building real world applications, it is unusual to be given a large training set at the start of the project; it is much more common to have no training set at all.

The second reason is that medical speech translation applications are safety-critical. Mistranslations can have serious consequences, and this needs to be reflected in the evaluation metric. A metric which maximizes BLEU score or recall, typical of most current evaluations, is inappropriate. No doctor we have talked to would consider BLEU a useful metric.

In both respects, the application we describe is closer to real world ones than is common in the literature, and we therefore think it reasonable to claim that our results should not be dismissed as irrelevant; we suspect that similar problems will emerge in many other real world applications. The Open Source framework we have used make it easy for sceptical researchers to check the details of our methods and data.

References

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

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

Hockey, B., Rayner, M., and Christian, G. (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., Wooters, C., Segal, J., Stolcke, A., Fosler, E., Tajchman, G., and Morgan, N. (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., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 2.

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

Rayner, M., Bouillon, P., Chatzichrisafis, N., Hockey, B., Santaholma, M., Starlander, M., Isahara, H., Kanzaki, K., and Nakao, Y. (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., Estrella, P., and Bouillon, P. (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., Estrella, P., Bouillon, P., Hockey, B., and Nakao, Y. (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.

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

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

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