ID Page

Comparing Rule-Based and Statistical Approaches to Speech Understanding in a Limited Domain Speech Translation System

Paper ID: 45

Manny Rayner[1,2], Pierrette Bouillon[1], Beth Ann Hockey[2],
Nikos Chatzichrisafis[1], Marianne Starlander[1]

(1) University of Geneva
TIM/ISSCO
40, bvd du Pont-d'Arve
CH-1211 Geneva 4, Switzerland

(2) UCSC/NASA Ames Research Center
Moffet Field, CA 94035
USA

mrayner@riacs.edu, pierrette.bouillon@issco.unige.ch, bahockey@riacs.edu,
nikolaos.chatzichrisafis@issco.unige.ch, marianne.starlander@eti.unige.ch

Comparing Rule-Based and Statistical Approaches to Speech Understanding in a Limited Domain Speech Translation System

Manny Rayner[1,2], Pierrette Bouillon[1], Beth Ann Hockey[2],
Nikos Chatzichrisafis[1], Marianne Starlander[1]

(1) University of Geneva, TIM/ISSCO
40, bvd du Pont-d'Arve
CH-1211 Geneva 4, Switzerland

(2) UCSC/NASA Ames Research Center
Moffet Field, CA 94035
USA

mrayner@riacs.edu, pierrette.bouillon@issco.unige.ch, bahockey@riacs.edu,
nikolaos.chatzichrisafis@issco.unige.ch, marianne.starlander@eti.unige.ch

Abstract

The paper directly compares two versions of a medical speech translation system, one with a grammar based language model (GLM) recognizer and the other with a statistical language model (SLM) recognizer. We construct the GLM using a corpus-based method, so that both the GLM and the SLM can be derived from the same corpus; evaluation is carried out with respect to performance on the speech translation task. Despite using a very small training set for both the GLM and the SLM, the SLM delivers much better word error rates on unseen test material. Nonetheless, evaluating both systems on translation performance rather than word error rates, the GLM-based version of the system outperforms the SLM on the actual translation task.

1. Introduction

At present, the field of speech understanding finds itself in a curious position. As far as academic research is concerned, the dominant paradigm is the statistical one. Speech recognizers are built using some kind of corpus-based process to induce a statistical language model (SLM). These recognizers are typically combined with some kind of robust parser to form a language understanding system. There is less agreement about how the parser should be constructed, and it is still reasonably popular to structure it as a rule-based system. However, the general feeling seems to be that it is in some sense "better" to try to make the language processing component trainable and corpus-based as well: the justification for this claim is typically motivated in terms of robustness, lower authoring costs, and (sometimes) better performance. A good summary of current trends is presented in (Young 2002).

In contrast, commercial speech understanding systems are mainly rule-based. For example, Nuance (2004) until fairly recently only supported grammar-based language models and it is still the case that the vast majority of Nuance applications use grammar-based methods. Rule-based systems are the norm in industry for several reasons, but the most critical one is certainly lack of training data. At the start of a project, there will typically be little or no data available to train a language model. Creating corpus data using Wizard of Oz techniques is both expensive and time-consuming; if the goal is to produce a spoken language interface for a new application in a timely fashion, rule-based methods are often the only practicable alternative.

Over the last few years, however, the above picture has become more blurred. Commercial platform vendors have begun to introduce statistical modeling tools, like Nuance's SayAnything™ module, with some success. In the other direction, academics have become more interested in grammar-based methods. The literature now contains descriptions of several research systems built using rule-based methods, which successfully use mixed-initiative strategies and complex grammars (Stent et al. 1999, Rayner et. al. 2000, Lemon et al. 2001, Rayner et. al. 2001a). Much of this work has involved the idea of compiling grammar-based language models out of descriptions written in higher-level formalisms, in particular unification grammars (Moore 1998, Kiefer & Krieger 2000, Dowding et al. 2001, Rayner et al. 2001b, Bos 2002).

Given that a great deal of practical and theoretical work is going on using both statistical and rule-based methods, it is remarkable that there is almost no reported work attempting a systematic comparison of the two approaches. The only example known to us is (Knight et al. 2001), in which one of the present authors participated. In this study, two speech understanding systems were constructed for the same domain, a medium-vocabulary command and control task. Both systems ran on the Nuance 7 platform. The first had a hand-coded grammar-based language model (GLM), compiled using the standard Nuance Toolkit into a recognition package. The second system instead used a statistical language model (SLM), trained using the SRILM package (Stolcke et al. 2002) on a corpus of about 4000 domain utterances. A robust parser post-processed the word-strings produced by the SLM-based recognizer; it built semantic representations in the format used by the grammar-based system, so that an exact comparison was possible. The results were interesting, but inconclusive. The SLM-based system had a slightly lower WER, but a considerably higher semantic error rate when evaluated on data provided by experienced users who knew the system's coverage. When evaluated on naive users who were encouraged to experiment freely with different constructions, the SLM comfortably outperformed the GLM.

(Knight et al. 2001) highlights the problems involved in carrying out a methodologically sound comparison between the statistical and rule-based approaches. An immediate problem is the role of the corpus. It is straightforward to determine exactly what corpus material has been used to train an SLM, but the identity of the data used to construct a GLM is usually much less clear.

In the current paper, we will present a comparative study, using a version of the MedSLT spoken language translation system (Rayner et. al. 2003). We construct the GLM using a corpus-based method, so that both the GLM and the SLM can be derived from the same corpus; evaluation is carried out with respect to performance on the speech translation task, to avoid the problem of defining semantic error rate.

The results were surprising. We had expected that one of the advantages of a GLM would be that it would yield better recognition performance when there was little training data available. In fact, despite using a very small training set for both the GLM and the SLM, the SLM delivers much better word error rates on unseen test material. None the less, the GLM-based version of the system performs much better on the actual translation task than the SLM-based version.

The rest of the paper is structured as follows. Section 2 describes the MedSLT spoken language translation system. Section 3 describes the corpus-based framework used to construct the GLM, and Section 4 describes how we built an SLM-based version of MedSLT. The meat of the paper is in Sections 5 and 6, which respectively describe the experiments and the results. Section 7 concludes.

2. The MedSLT System

The top-level goal of the MedSLT (2004) project is to develop an Open Source framework for rapid construction of practically useful limited-domain spoken language translation systems for the doctor-patient examination domain. The basic scenario envisaged is one-way communication. The doctor asks questions, which are translated by the system, and the patient responds non-verbally, for example by nodding or shaking their head, or pointing.

The key requirement of the project is a high level of accuracy: doctors are only interested in using systems they can trust. Since speech recognition can never be wholly reliable, the user interface is structured so that the initial recognition hypothesis is echoed back to the user, who has the option to abort further processing if necessary. Reliability thus means reliability on the utterances which the user considers to be correctly recognized.

The current prototype system (Rayner et al. 2003) translates questions in a headache domain from English into Japanese or French, using a vocabulary of about 200 words; a production version would need to cover at least another 25 to 50 similar subdomains. The recognition component for the SLM and GLM version was built with a training corpus of 450 utterances. The initial set of training utterances were supplied to us by a physician, who then interacted with us to expand it by adding enough synonyms and alternative phrasings to make the coverage reasonably habitable. Later versions may use larger development sets, but it is unreasonable to assume large corpora for such specialized domains during development.

The semantic representations used by MedSLT are simple lists of argument/value pairs: thus for example the representation of "where is the pain" is

```
[[utterance_type,whq], [loc,where], [tense,present], [verb,be],
[spec,the_sing],[symptom,pain]]
```

The pros and cons of this approach are discussed in (Rayner & Bouillon 2002). For the purposes of the present paper, the most important point is that it is easy for a shallow parser to produce this kind of flat representation, facilitating a fair comparison between the grammar-based methods used in the standard version of MedSLT and the SLM-based methods described in the next section.

Translation is carried out by first using a set of transfer rules to map the source-language semantic representations into similar ones in the target language, and then generating a target-language string using another REGULUS grammar compiled in a form suitable for generation. Output target speech is produced using the Nuance Vocalizer™ TTS engine for French, and concatenated pre-recorded audio files for Japanese.

3. Grammar-based Recognition Using REGULUS

The grammar based system was built using REGULUS (Rayner et. al. 2003), an Open Source environment that supports efficient compilation of typed unification grammars into speech recognizers. The basic intent is to provide a set of tools to support rapid prototyping of spoken dialogue applications in situations where little or no corpus data exists.

The core functionality provided by the REGULUS environment is compilation of typed unification grammars into annotated context-free grammar language models expressed in Nuance Grammar Specification Language (GSL) notation (Nuance 2004). GSL language models can be

converted into runnable speech recognizers by invoking the Nuance Toolkit compiler utility, so the net result is the ability to compile a unification grammar into a speech recognizer.

Experience with grammar-based spoken dialogue systems shows that there is usually a substantial overlap between the structures of grammars for different domains. This is hardly surprising, since they all ultimately have to model general facts about the linguistic structure of English and other natural languages. It is consequently natural to consider strategies which attempt to exploit the overlap between domains by building a single, general grammar valid for a wide variety of applications. A grammar of this kind will probably offer more coverage (and hence lower accuracy) than is desirable for any given specific application. It is however feasible to address the problem using corpus-based techniques which extract a specialized version of the original general grammar.

REGULUS implements a version of the grammar specialization scheme which extends the Explanation Based Learning (EBL) method described in (Rayner et al. 2002). A specialized Nuance grammar is derived from the general grammar in the following processing stages:

- The training corpus is converted into a "treebank" of parsed representations. This is done using a left-corner parser representation of the grammar.
- The treebank is used to produce a specialized grammar in REGULUS format, using the EBL algorithm (van Harmelen & Bundy 1988, Rayner 1988).
- The final specialized grammar is compiled into a Nuance GSL grammar.

It follows automatically from the EBL algorithm that the semantic representations produced by the specialized grammar are identical to those produced by the general one.

In the context of this paper, the critical point about the REGULUS grammar development framework is that it makes explicit the role of the corpus. Since the final specialized grammar can only include words and constructions licensed by the training corpus, it is meaningful to make comparisons between a grammar-based recognizer built using REGULUS, and an SLM-based recognizer built from the same corpus using standard methods.

4. An SLM-based Version of MedSLT

We constructed a robust version of the prototype MedSLT system from the same 450 utterances of data used to train the original grammar-based version. We first built a class N-gram language model, using the Nuance SayAnything™ tool. The model used 17 class definitions; these were determined by hand examination of the REGULUS lexicon, and mostly consisted of groups of semantically similar words and phrases. For example, the words "radiate", "spread" and "extend" form a class.

Syntactic and semantic processing is carried out by a simple surface parser, which applies a set of about 300 hand-coded patterns to identify semantic attribute/value tags. For example, the pattern used to identify the tag [prep, in_time] (temporal use of "in") is

```
pattern([in,'…',morning|afternoon|evening],[prep,in_time]).
```

The initial set of patterns was automatically extracted from the REGULUS lexicon. The raw set of tags found by pattern-matching is post-processed by a small set of extra rules which look for global information, and in particular add a tag classifying the utterance as a Y-N question, a WH-question, or an elliptical phrase. The output of the robust parser matches that of the REGULUS parser on 98% of the 450 utterance training corpus.

The robust version of the system uses the same transfer and generation components as the grammar-based one, so translation fails if transfer is unable to create a well-formed semantic representation on the target side.

5. Experiments

Data collection was divided into three parts. In the first part, subjects were handed an introductory text and were informed about the domain of the application. The second part consisted of the subject reading a sample of 55 system sentences that were chosen to illustrate the range of system coverage.

The last part of the data collection was designed to collect live application data. The subject was instructed to perform the task of diagnosing the patient's headache type using the translation system. To accomplish the task, the subject was given a set of eight headaches along with typical symptoms divided into categories like 'where', 'pain type' and 'frequency/duration'. Many headaches had similar symptoms, which made it necessary to ask questions from different categories in order to differentiate amongst headaches. Data collection moderators acted the patient role, giving answers that would suggest a specific predetermined headache type.

We collected data from 11 subjects with no prior experience in spoken language systems, where the first subject was used for testing and refining the experimental setup and the data collection procedure. Collected speech data was transcribed, fed through both the GLM and SLM versions of the MedSLT system, and given to three judges fluent in both English and French.

System performance was evaluated in two phases. In the first phase, the judge was shown only the transcription of each utterance together with the recognition output, and asked to mark it as either "acceptable" or "unacceptable". The intention was to simulate operation of the "abort" button in the live system (cf. Section 2).

In the second phase, sentences marked as "acceptable" in the first phase were further classified into 'good', 'ok' and 'bad'. Judges were instructed to label sentences as 'good' to correctly translated sentences, with colloquial language and good grammaticality. Judges were advised to label 'ok' translations with minor problems in grammar or word choice, but which preserved the meaning of the source sentence. Finally the label 'bad' was given to sentences with translations which did not preserve the original meaning.

6. Results

Table 1 through Table 4 present the main results. Looking at Table 1, we see that the SLM version has much better word error rate (WER) than the GLM version (24.10% versus 36.67%) and approximately equal SER (59.16% versus 60.69%). The first line of Table 2 also shows that users mark far fewer SLM sentences as unacceptable (207 versus 283). Despite this, it is striking to note that the GLM delivers much better overall performance. Going further down Table 2 we see that the SLM-based system fails to translate 94 of the "acceptable" utterances, while the GLM misses only 2; this means that the GLM actually ends up producing more translations than the SLM.

Comparing Tables 3 and 4, we can see that although the SLM and GLM produce roughly comparable proportions of "good" and "ok" utterances, the proportion of "bad" judgments by each judge is roughly twice as high for the SLM system. In fact, comparison between judges shows that the imbalance is even greater. For the GLM system, there were only 4 utterances

judged bad by two judges, and none judged bad by all three. For the SLM system, in contrast, there were 15 sentences judged bad by two judges, and 5 judged bad by all three. One could thus reasonably argue that the SLM version is mistranslating four or five times as many sentences as the GLM version.

It is at first sight surprising that the GLM system can achieve better task performance than the SLM one, despite its greatly inferior WER. The roughly equal SER scores for the two systems, however, give a first hint at an explanation. The GLM enforces global constraints on the recognized utterance, so it is harder for an utterance to be only partially correct. In contrast, the local constraints in the SER make it easier for part of an out-of-coverage utterance to be correctly recognized, while the rest is more or less badly scrambled. The character of the task, though, is "all-or-nothing": partial solutions are rarely useful, even though they improve the SLM version's WER score.

Further analysis suggested that the reason why the performance of the SLM seems so much less reliable than that of the GLM appears to lie not so much in the systems themselves as in their interaction with the user. When the user marks an utterance as "acceptable", they are making an intuitive judgment based on a very incomplete understanding of how the system operates. For example, with the system architecture and domain used in this evaluation, it is usually the case that a missing determiner or article makes no difference to the translation: thus if the user says "is the pain in the front of the head", it will be fine to accept a recognition hypothesis like "is a pain in front of head". During the experiments, users rapidly learned that small recognition errors like these were often harmless.

Unfortunately, other apparently minor recognition errors turned out to be fatal. In a typical example, the user utterance was "is the frequency of the headaches increasing", recognized as "does the frequency of the headaches increasing". The user plausibly accepted this as a correct recognition, but the system was not able to translate at all. In a similar example, "does change in temperature make it worse" was recognized as "does changes in temperature make worse" and accepted; once again, the system was unable to translate.

	SLM	GLM
WER	24.10%	36.67%
SER	59.16%	60.69%

Table 1 SER and WER in SLM and GLM versions

	SLM	GLM
Unacceptable Recognition	207	283
No result	94	2
Translated	223	239
Total	524	524

Table 2 Breakdown of examples translated by SLM and GLM.

	Good	OK	Bad
Judge 1	106	127	6
Judge 2	169	55	15
Judge 3	174	57	8
Average	150	80	10

Table 3 Quality of translation with GLM version

	Good	OK	Bad
Judge 1	91	113	19
Judge 2	158	34	31
Judge 3	151	57	15
Average	133	68	22

Table 4 Quality of Translation with SLM version

The worst problem with the SLM version, as shown by the second line of Table 2, is that the differences between harmless and dangerous recognition errors are often incomprehensible to the user. Correspondingly, the key advantage of the GLM version in this respect is that all recognized utterances are guaranteed to be within the system's grammar coverage, which makes it easy for the user to differentiate between acceptable and unacceptable recognition results. The subjective experience with the GLM version is that the user is in control. When the system produces something that looks like a plausible hypothesis, she can be confident that pressing the "accept" button will almost always result in the production of a translation, which will almost always be correct or at least acceptable. With the SLM version, the subjective experience is rather that the system is in control. Plausible looking recognition results frequently result in no translation, and the quality of the translations is much less reliable. It was immediately apparent during the tests that most of the GLM subjects found the experience enjoyable, while the SLM subjects experienced it as frustrating.

Since the problems in the SLM version are largely caused by the fact that we are interfacing a statistical/surface oriented speech understanding module to a rule-based translation module, one obvious solution would be to add a robust backup translation component that took over if the current rule-based translation component failed. This would certainly make the SLM system feel more user-friendly, but would also have the effect of rendering translation quality even less reliable, as surface translation methods typically trade precision for recall. For example, the experiments in (Carter et al. 2000) on a multi-strategy speech translation system showed that the proportion of clearly incorrectly translated utterances increased from 5% using rule-based translation, through 11% for phrasal translation, to 22% for completely surface-based translation. Given that the proportion of bad translations in the SLM-based system is already two to five times higher than in the GLM-based version, this seemed like a dangerous step.

Another possible argument is that the rule-based translation in the SLM-based system could be improved simply by doing more work; poor performance is maybe just due to lack of implementation effort. In order to estimate the potential for straightforward improvement to the system, we began by manually going through the 94 utterances tagged "acceptable" where the SLM failed to produce a result, investigating in each case why processing had failed. We classified the example as "bug" if the problem appeared to be a simple mistake or lack of a rule; "missing heuristic", if a plausible heuristic could have been added which would at least be correct in most cases; and "confusion" if the problem was most likely that the user was confused about the system's capabilities, and no obvious fix would resolve it.

The examples divided fairly evenly across the three categories, with 30 "bugs", 29 "heuristics", and 35 "confuseds". Subsequent system development broadly justified our intuitive classification. We were able to fix a large proportion of the problems in the first and second groups, but it was essentially impossible to make any progress on the "confused" utterances; when we re-ran the same data on a later version of the system, about half of the 94 problem utterances gave acceptable results, but the SLM version was still failing to produce a translation more than 10 times as often as the GLM one. It seems reasonable to us to conclude that safety-critical nature of the task implies that the cautious, high precision GLM architecture offers real advantages compared to the adventurous, high-recall SLM alternative, especially when this agrees well with the intuitive user experience.

7. Conclusions

We have presented a series of experiments directly comparing two versions of a medical speech translation system, designed to investigate the tradeoffs between grammar-based and statistical speech understanding strategies. Evaluated on the medical speech translation task, the GLM system produced better results than the SLM system, despite the SLM's lower WER and comparable SER. Analysis of the results suggests that the success of the GLM version is largely due to its ability to present a more predictable interface to the user, which in practice appears to outweigh the loss of recall inherent in a rule-based strategy. We hypothesize that this tradeoff may well apply to other safety-critical speech processing tasks.

8. References

Johan Bos. Compilation of Unification Grammars with Compositional Semantics to Speech Recognition Packages. In Proc. of the 19th International Conference on Computational Linguistics, 2002.

D. Carter, M. Rayner, R. Eklund, C. MacDermid, and M. Wiren. Evaluation. In M. Rayner, D. Carter, P. Bouillon, V. Digalakis, and M. Wiren, editors, The Spoken Language Translator. Cambridge University Press, 2000.

J. Dowding, B.A. Hockey, J.M. Gawron, and C. Culy. Practical issues in compiling typed unification grammars for speech recognition. In Proc. of the 39th Annual Meeting of the ACL, Toulouse, France, 2001.

B. Kiefer and H. Krieger. A context-free approximation of head-driven phrase structure grammar. In Proc. of the 6th International Workshop on Parsing Technologies, pages 135-146, 2000.

S. Knight, G. Gorrell, M. Rayner, D. Milward, R. Koeling, and I. Lewin. Comparing grammar-based and robust approaches to speech understanding: a case study. In Proc. of Eurospeech 2001, pages 1779-1782, Aalborg, Denmark, 2001.

O. Lemon, A. Bracy, A. Gruenstein, and S. Peters. Multimodal dialogues with intelligent agents in dynamic environments: the WITAS conversational interface. In Proc. of 2nd Meeting of the North American ACL, Pittsburgh, PA, 2001.

MedSLT. http://sourceforge.net/projects/medslt/, 2004. As of 26 March 2004.

R. Moore. Using natural language knowledge sources in speech recognition. In Proc. of the NATO Advanced Studies Institute, 1998.

Nuance. http://www.nuance.com, 2004. As of 26 March 2004.

S.G. Pulman. Syntactic and semantic processing. In H. Alshawi, editor, The Core Language Engine, pages 129-148. MIT Press, Cambridge, Massachusetts, 1992.

M. Rayner. Applying explanation-based generalization to natural language processing. In Proc.of the International Conference on Fifth Generation Computer Systems, pages 1267-1274, Tokyo, Japan, 1988.

M. Rayner and P. Bouillon. A phrasebook style medical speech translator. In Proc. of the 40th Annual Meeting of the ACL (demo track), Philadelphia, PA, 2002.

M. Rayner, P. Bouillon, V. Van Dalsem III, B.A. Hockey, H. Isahara, and K. Kanzaki. A limited-domain English to Japanese medical speech translator built using REGULUS 2. In Proc. of the 41st Annual Meeting of the ACL (demo track), Sapporo, Japan, 2003.

M. Rayner, J. Dowding, and B.A. Hockey. A baseline method for compiling typed unification grammars into context free language models. In Proc. of Eurospeech 2001, pages 729-732, Aalborg, Denmark, 2001b.

M. Rayner, B.A. Hockey, and J. Dowding. Grammar specialisation meets language modelling. In Proc. of the 7th International Conference on Spoken Language Processing (ICSLP), Denver, CO, 2002.

M. Rayner, B.A. Hockey, and J. Dowding. An open source environment for compiling typed unification grammars into speech recognisers. In Proc. of the 10th EACL (demo track), Budapest, Hungary, 2003.

M. Rayner, B.A. Hockey, and F. James. A compact architecture for dialogue management based on scripts and meta-outputs. In Proc. of the 6th Applied Natural Language Processing Conference, Seattle, WA, 2000.

M. Rayner, I. Lewin, G. Gorrell, and J. Boye. Plug and play spoken language understanding. In Proc. of the 2nd ACL SIGDIAL Workshop on Discourse and Dialogue, Aalborg, Denmark, 2001a.

REGULUS. http://sourceforge.net/projects/regulus/, 2003. As of 26 March 2004.

A. Stent, J. Dowding, J. Gawron, E. Bratt, and R. Moore. The CommandTalk spoken dialogue system. In Proc. of the Thirty-Seventh Annual Meeting of the ACL, pages 183-190, 1999.

A. Stolcke. SRILM - an extensible language modeling toolkit. In Proc. of the International Conference on Spoken Language Processing (ICSLP), 2002.

J. van Eijck and R. Moore. Semantic rules for English. In H. Alshawi, editor, The Core Language Engine, pages 83-116. MIT Press, 1992.

T. van Harmelen and A. Bundy. Explanation-based generalization partial evaluation (research note). Artificial Intelligence, 36:401-412, 1988.

S. Young. Talking to machines (statistically speaking). In Proc. of the 7th International Conference on Spoken Language Proc. (ICSLP), pages 9-16, Denver, CO, 2002.