A Methodology for Comparing Grammar-Based and Robust Approaches to Speech Understanding

Manny Rayner[1,2], *Pierrette Bouillon*[1], *Nikos Chatzichrisafis*[1], *Beth Ann Hockey*[2],
Marianne Santaholma[1], *Marianne Starlander*[1], *Hitoshi Isahara*[3], *Kyoko Kanzaki*[3], *Yukie Nakao*[3]

(1) University of Geneva, TIM/ISSCO, 40, bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
{pierrette.bouillon, nikolaos.chatzichrisafis}@issco.unige.ch,
{marianne.santaholma, marianne.starlander}@eti.unige.ch
(2) ICSI/UCSC/NASA Ames Research Center, Mail Stop T-27A, Moffett Field, CA 94035-1000
mrayner@riacs.edu, bahockey@email.arc.nasa.gov
(3) NICT, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0289
{isahara, kanzaki}@nict.go.jp, yukie-n@khn.nict.go.jp

Abstract

We present a series of experiments designed to compare grammar-based and robust approaches to speech understanding, performed in the context of an Open Source medical speech translation system. We used two versions of the system, one grammar-based and one robust, trained off the same training data, and evaluated them on test data collected using both versions of the system. The experiments were constructed so as to avoid several methodological problems which occurred in earlier work reported in the literature. We found that the grammar-based version gave significantly better results than the robust version, with the difference increasing as subjects became more familiar with the system's coverage. The rate of improvement in subject performance was positively affected by providing them with an intelligent online help system.

1. Introduction

Most spoken dialogue systems reported in the research literature during the last five to ten years use some variant of the same basic speech understanding architecture. Speech recognition is carried out using a statistical language model (SLM), usually a class bigram or trigram model. Recognition produces a word-string, which is converted into a semantic representation using some kind of phrase-spotting or robust parsing method. When an adequate quantity of training data is available, this methodology has repeatedly been shown to produce good results.

The general success of statistical/robust methods may give the impression that rule-based approaches are no longer relevant. A closer look, however, shows that the situation is not so clear. When it is feasible to train people to use the system, grammar-based language models (GLMs) can be very effective. A recent high-profile example is Clarissa [1] an experimental speech-enabled procedure reader that has been deployed on the International Space Station. Other prominent examples include CommandTalk [2], a speech interface to a military command and control simulation system, and WITAS [3], a spoken dialogue system for a simulated robotic helicopter domain. One thing that all these systems have in common is a user-base that is familiar with controlled language, and willing to acclimatise to it in order to get good performance. For example, the astronauts who use Clarissa are mostly ex-military pilots, and come from a background where controlled language is the norm.

The existence of impressive rule-based speech understanding systems, of course, proves nothing; after all, it might have been possible to get equally good or better performance using statistical/robust methods. A convincing argument must be based on direct comparison, where the same task is attempted using both methodologies. Examination of previous attempts shows though that a clean experiment is difficult to construct.

The first study of this kind known to us is [4]. Here, 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 GLM, compiled using the standard Nuance Toolkit into a recognition package. The second system instead used an SLM, trained on a corpus of about 4000 domain utterances. A robust parser post-processed the word-strings produced by the SLM-based recogniser; 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 with different constructions, the SLM comfortably outperformed the GLM. Both systems, however, performed very poorly on the naive subjects.

[4] highlights three immediate methodological problems. First, there 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. It is obviously desirable to be able to say in some well-defined sense that both systems have been constructed from the same data. A second problem is finding a suitable evaluation metric. As shown for example in [5], WER is usually a misleading measure of speech understanding performance, and should be replaced with a semantic measure. "Semantic error rate" is unfortunately not straightforward to define either. In most systems, the data structure built up at the level of semantic representation contains redundancy: it is often possible to get the same processing result from different semantic representations. [4], which was carried out in a home automation domain, suffered badly from this problem. Since determiners and articles were sometimes significant ("Switch off the light" versus "Switch off all the lights"), they were part of the

semantic representation, and an interpretation which got the article wrong was scored as semantically incorrect. In most cases, though, the article actually made no difference: “Switch off kitchen light” and “Switch off the kitchen light” would produce the same system response. Yet another problem is the lack of definition for “typical input”: it is not in any way clear whether the experienced or the naive users should be regarded as typical.

A recent attempt is reported in [6], which addresses many of the problems in [4]. The experiments were based on the Open Source MedSLT system [7], a medical speech-to-speech translator for a doctor-patient examination domain. As before, GLM and SLM versions of the recogniser were constructed, and compared on the speech translation task. The training corpus issue was resolved by constructing the GLM grammar using an example-based method, so that both recognisers could be derived from the same corpus. Evaluation was in terms of performance on the end-to-end speech translation task, giving a more appropriate measure of semantic accuracy. Given the safety-critical nature of the medical speech translation task, evaluation was modelled on the normal mode of system operation: each recognition string was echoed back to the user, who had to decide whether to accept or reject based on whether or not they believed that it was close enough to what they had said.

Although the GLM and SLM scored about equally well in terms of the number of sentences translated, [6] concluded that the GLM was preferable in terms of presenting a more predictable interface. Since the GLM-based system used the same grammar for both recognition and semantic analysis, nearly all utterances which were recognised could also be translated. This was not true for the SLM-based system: there was a much higher proportion of utterances which were judged by users as acceptably recognised, but which failed to result in a translation. For similar reasons, the SLM also produced many more incorrect translations.

Although [6] was a more convincing experiment than [4], it still contained methodological flaws. The most critical one is the presentation of recognition results to users. Simply echoing back the recognition string is a naive approach; producing a “back-translation” from the interlingua to the source language is usually preferable [8]. A second problem is the data collection methodology. [6] started each data collection session by asking subjects to read out fifty in-coverage utterances, to train them on grammar coverage. In retrospect, we were unhappy with this experimental procedure; it seems more appropriate to let subjects gain familiarity with the system by actually using it, possibly adding an intelligent help component to provide feedback on coverage [9]. Yet another issue is that data was only collected using the GLM version of the system.

In this paper, we describe a series of experiments which build on [6] and rectify all the above problems. With a newer release of the same speech translation system, we carried out a new data collection, using both GLM and SLM versions. Subjects received minimal training on the system before starting the session. Feedback on recognition results was provided by back-translating the interlingua representation into the source language, when possible. Half of the subjects used a version of the system enhanced with an intelligent help module, and no other assistance was given during the session. The utterances collected were recorded, transcribed, and then processed offline through both versions of the system.

Interestingly, we arrive at different conclusions from those in the earlier study. When back-translation is added to the system, the SLM version appears to be just as predictable as the GLM one. The key point, rather, is that GLM version is much better than the SLM one on in-coverage data. As users gain familiarity with the system, they gravitate towards grammar coverage, and the GLM version outperforms the SLM one by an increasingly large margin. Availability of an intelligent help system accelerates this trend, confirming the results of [9].

2. The MedSLT system

MedSLT [7, 6] is an Open Source limited-domain speech translation system for doctor-patient examination dialogues. Translation is one-way, in the doctor to patient direction; coverage is primarily organised around yes/no questions, with the expectation that the patient will answer non-verbally by nodding, shaking their head, pointing, or similar. Coverage is organised into sub-domains by symptom class. The coverage of each subdomain is based on standard examination questions provided by a medical professional, and supports a vocabulary of between 300 and 500 words. The versions of the system used for these experiments translate from English into French and Japanese.

The translation architecture is interlingua-based, and includes multiple processing engines, back-translation, context-dependent translation and an intelligent help component. The flow of processing is as follows. Input speech is recognised using two different recognisers, both built on top of the Nuance recognition platform. The first recogniser uses a PCFG language model, which directly produces a semantic representation; the second uses a class N-gram model, which produces a recognition string from which a representation is derived using a set of phrase-spotting rules. The two recognisers are described in Sections 2.1 and 2.2.

The source language semantic representation is passed to a discourse processing module, which interprets it in the context of the previous dialogue. The resolved representation is then transformed into its interlingual counterpart. The basic principle behind the interlingual representation is to treat each clause as a flat list of key-value pairs.

The interlingual form is first translated back into the source language and shown to the user, who has the option to abort further processing if they consider that the system has failed to understand what they said. If they approve the back-translation, the interlingual form is transferred into a target language representation. This is then transformed into a target language surface string using a generation grammar, and finally passed to a speech synthesis unit.

The system optionally invokes a simple context-sensitive help module. This uses the result of robust SLM-based recognition to display a list of in-coverage example sentences. Examples are selected from a predefined list, using a heuristic that prioritises sentences maximizing the number of bigrams and unigrams shared with those extracted from the SLM recognition result.

2.1. Grammar based recognition

The grammar-based recogniser is built using the Open Source REGULUS 2 platform [10, 11]. The recognition grammar for each of the three subdomains (headache, chest pain or abdominal pain) is derived from a single general unification grammar, using an Explanation Based Learning method driven by small corpora. The same corpus that is used to derive the domain-specific grammar is then used a second time, to perform probabilistic tuning of the CFG language model. Probabilistic tuning is done using the Nuance `compute-grammar-probs` utility.

In the context of the present paper, the attraction of using a corpus-based grammar construction method is that it permits a fair comparison of the grammar-based and N-gram based recognisers, since both can be trained using the same corpus.

2.2. N-Gram based recognition

The N-gram based recogniser employs a conventional class N-gram model constructed using the Nuance SayAnything^{TM} tool. The model uses 32 class definitions; these mostly consist of groups of semantically similar words and phrases. For example, the verbs "radiate", "spread" and "extend" form a class, and the body-part nouns "head", "eye", "face", "neck", "jaw" and form a second class.

The source language semantic representation is produced from the recognition string using a set of robust surface processing rules. Processing consists of three phases [6]. First, a set of rules is applied that attempts to detect start- and end-boundaries for subordinate clauses. Once the recognition string has been segmented into clauses, a second set of rules is applied, to guess key/value pairs. Finally, a set of post-processing rules is applied, which fills in default values for unset features in the representation of each clause. Performance of the robust and rule-based versions of the system agree to 99% on the development corpus.

3. Experiments

We used the October 2004 version of the Open Source MedSLT system to perform our experiments; we used only the headache subdomain. Both versions of the recogniser were trained from the same corpus of 575 text utterances, available from the MedSLT website. We collected data from 12 native speakers of English, using a data collection protocol based on the one described in [6]. Each subject was first given a short acclimatisation session, where they used a prepared list of ten sentences that were in-coverage for the GLM to learn how to use the microphone and the push-to-talk interface. They were then encouraged to play the part of a doctor, and conduct an examination interview, through the system, on a team member who simulated a patient suffering from a specific type of headache. The subject's task was to identify the type correctly out of a list of eight possibilities.

We modified the protocol from [6] in a few respects. Half of the subjects used the grammar-based version of the system, and half the N-gram based version ([6] only used the grammar-based version). We also shortened the initial acclimatisation session from 50 sentences to 10, in the interests of reducing the extent to which subject language was biased by the introductory material. It seems likely to us that the higher word error rates we obtained, compared to [6], are mainly due to this difference. We collected a total of 870 recorded utterances; the breakdown by subject is shown in Table 1, which also shows the proportion of utterances misrecognised by each version of the system.

The recorded data was first transcribed, and then processed through offline versions of both the GLM and SLM processing paths in the system. This was done as follows. We first set the system to translate from English into English (via the interlingua), and then had an English-speaking judge evaluate the "back-translation" of each utterance to determine whether or not it was an acceptable paraphrase of what had been said. Utterances for which the back-translation was judged acceptable were regarded as correctly recognised, and were then translated further into the two target languages French and Japanese.

Subject	#Utts	Version		Misrecognised	
				GLM	SLM
AG	31	SLM	H–	19.4%	25.8%
NA	48	SLM	H–	33.3%	43.8%
AL	53	SLM	H–	45.3%	47.2%
SU	104	SLM	H+	47.1%	45.2%
AN	77	GLM	H+	51.9%	51.9%
GA	70	GLM	H–	51.4%	61.4%
RA	59	SLM	H+	57.6%	71.2%
DA	80	SLM	H+	63.8%	57.5%
AD	65	GLM	H+	66.2%	70.8%
MA	33	GLM	H+	66.7%	69.7%
IA	80	GLM	H–	67.5%	80.0%
EM	170	GLM	H–	70.0%	69.4%

Table 1: Results of data collection. For each subject, we specify the number of utterances collected (#Utts), whether the GLM or the SLM version of the system was used to collect data for that subject, whether the help system was switched on (H+) or off (H–), and the proportion of utterances misrecognised when the recorded utterances are processed using the GLM and SLM versions respectively.

Table 2 presents overall Word Error Rate (WER), Sentence Error Rate (SER) and Semantic Error Rate (SemER) for both versions of the system, where SemER is measured as the proportion of utterances not receiving an acceptable back-translation. Since performance of the recognisers, particularly the GLM version, differs greatly depending on whether or not it was within the coverage of the GLM grammar, we present separate figures for in-coverage data (417 utterances) and out-of-coverage data (453 utterances).

	In coverage		Out of coverage	
	GLM	SLM	GLM	SLM
WER	5.7%	12.7%	57.5%	47.8%
SER	19.4%	36.7%	99.8%	91.4%
SemER	18.5%	28.1%	87.9%	89.0%

Table 2: WER, SER and SemER for SLM and GLM versions of the recogniser, on in-coverage and out-of-coverage data.

Translations to French and Japanese were judged for acceptability by native speaker judges for each language: there were six judges for French, and three for Japanese. Judges were asked to categorise translations as "good", "acceptable" or "bad". For each target language, and each processing method (GLM or SLM), we consolidated the results using a majority voting scheme. If two-thirds of the judges (i.e. four for French, or two for Japanese) agreed that the translation was clearly "good" or "bad", we counted the translation as belonging to the appropriate category. Otherwise, we counted it as "acceptable". The results of this judging are shown in Table 3.

For each target language, we next extracted the sets of utterances where one type of processing yielded a clearly better result than the other. By "clearly better", we mean that either one result was judged "good" and the other not, or that one was "acceptable" and the other was some kind of error. The balance in favour of GLM processing was 83–46 for French, and 77–41 for Japanese. According to the McNemar sign test, both these results are significant at well under the 1% level.

	French		Japanese	
	GLM	SLM	GLM	SLM
Bad recognition	54.6%	59.8%	54.6%	59.8%
Good translation	34.4%	30.8%	36.4%	32.8%
Acceptable trans	8.7%	7.7%	3.6%	3.3%
Bad translation	0.3%	0.2%	0.5%	0.5%
No translation	2.0%	1.5%	4.9%	3.7%

Table 3: Breakdown of English → French and English → Japanese translation using SLM and GLM versions. Translation judgements have been consolidated using majority voting.

4. Discussion

We will now pull the threads together, and draw some general conclusions about the relative performance of GLM and SLM processing in these experiments. Looking again at Table 2, the striking difference is on the in-coverage data; the GLM scores much better than the SLM, especially on WER (5.7% versus 12.7%) and SER (19.4% versus 36.7%). Robust processing lets the SLM recover somewhat on SemER (18.5% versus 28.1%), but the GLM still comes out a clear winner. Interestingly enough, although the SLM scores better than the GLM on the out-of-coverage data in terms of WER (47.8% versus 57.5%) and SER (91.4% versus 99.8%), the two systems score about equally even here in terms of semantic error rate (87.9% versus 89.0%).

Over the whole dataset, GLM processing is slightly worse on WER (34.3% versus 31.6%), but slightly better on SER (62.0% versus 66.3%), and SemER (54.6% versus 59.8%). This reflects the balance between in-coverage and out-of-coverage utterances in the data. SemER is about 10% better for the GLM on in-coverage data, and the two versions are about equal on out-of-coverage data; since we have about half in-coverage and half out-of-coverage, the GLM does about 5% better overall.

The ratio of in-coverage to out-of-coverage in the dataset is mainly a function of how familiar the subjects are with the system's coverage. An experienced user will produce mostly in-coverage data; a novice user will produce mostly out-of-coverage data. Performance improves with increased familiarity, and this favours the GLM system: as people become more expert, they gravitate towards the intended coverage. One way to quantify this effect is to contrast performance for the two architectures, averaged over the first and last quarters of each session. Table 4 presents the results, divided up according to whether the help system was enabled or not. Even without the help system, performance improved on average by about 7% over the course of the session. With GLM processing and the help system turned on, the improvement was much larger, averaging over 17%.

Part session	Help OFF		Help ON	
	GLM	SLM	GLM	SLM
First quarter	58.6%	65.8%	63.1%	64.1%
Last quarter	52.1%	58.1%	45.9%	56.0%
Improvement	6.5%	7.7%	17.2%	8.1%

Table 4: Performance of GLM and SLM processing, measured by proportion of unacceptably recognised utterances, for versions with and without access to help system.

5. Summary and conclusions

We have presented a carefully constructed experiment, intended to produce a fair comparison of grammar-based and robust processing. For the kind of "all or nothing" task considered here, where partial results are not relevant, grammar-based processing clearly outperforms robust processing. In particular, the semantic error rate for grammar-based processing on in-coverage material is reduced by about 35% (relative) compared to robust processing, while the semantic error rate on out-of-coverage material is about the same. As users become more familiar with the system's coverage, the imbalance in favour of grammar-based processing becomes more marked. It is possible to accelerate the rate at which users acclimatise to the system by providing an intelligent help system, which provides help feedback based on the results of robust processing.

6. References

[1] Clarissa, http://www.ic.arc.nasa.gov/projects/clarissa/, 2005, as of 4 April 2005.

[2] A. Stent, J. Dowding, J. Gawron, E. Bratt, and R. Moore, "The CommandTalk spoken dialogue system," in *Proceedings of the 37th ACL*, 1999, pp. 183–190.

[3] O. Lemon, A. Bracy, A. Gruenstein, and S. Peters, "Multimodal dialogues with intelligent agents in dynamic environments: the WITAS conversational interface," in *Proceedings of 2nd NAACL*, Pittsburgh, PA, 2001.

[4] 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 *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 1779–1782.

[5] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Proceedings of Eurospeech 2003,* Geneva, Switzerland, 2003, pp. 609–612.

[6] M. Rayner, P. Bouillon, B. Hockey, N. Chatzichrisafis, and M. Starlander, "Comparing rule-based and statistical approaches to speech understanding in a limited domain speech translation system," in *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, 2004.

[7] MedSLT, http://sourceforge.net/projects/medslt/, 2005, as of 8 January 2005.

[8] R. Frederking, A. Rudnicky, and C. Hogan, "Interactive speech translation in the diplomat project," in *Proceedings of the Spoken Language Translation Workshop at the 35th ACL*, Madrid, Spain, 1997.

[9] B. Hockey, O. Lemon, E. Campana, L. Hiatt, G. Aist, J. Hieronymus, A. Gruenstein, and J. Dowding, "Targeted help for spoken dialogue systems: Intelligent feedback improves naive user's performance," in *Proceedings of the 10th EACL*, Budapest, Hungary, 2003.

[10] Regulus, http://sourceforge.net/projects/regulus/, 2005, as of 8 January 2005.

[11] M. Rayner, B. Hockey, and J. Dowding, "An open source environment for compiling typed unification grammars into speech recognisers," in *Proceedings of the 10th EACL (demo track)*, Budapest, Hungary, 2003.