Improving out-of-coverage language modelling in a multimodal dialogue system using small training sets

Louis ten Bosch

CSLT, Radboud University Nijmegen, The Netherlands
L.tenBosch at let.ru.nl

10.21437/Interspeech.2005-404

Abstract

For automatic speech recognition, the construction of an adequate language model may be difficult when only a limited amount of training text is available. Previous work has shown that in the case of small training sets statistical language models may outperform grammars on out-of-coverage utterances, while showing comparable performance on in-coverage input. In this paper, we compare the performance of an automatic speech recognition system using a grammar and a statistical language model including garbage models in the case of very limited in-domain training data. The results show that a bigram language model and a grammar show similar performance, and that the inclusion of garbage models in statistical language models enhances their performance both on in-coverage and out-of-coverage utterances.

1. Introduction

An important issue for automatic speech recognition (ASR) for a particular domain is the availability of appropriate in-domain text for the purpose of language modelling. Since the language model (LM) spans the set of utterances that can be recognized by the speech recogniser, it must be carefully tuned to cover the set of utterances that users may apply when talking to the system. This tuning is difficult when there is only a small amount of in-domain texts available. In dialogue systems for particular domains, the problem of the construction of a language model is actually twofold: a) language coverage: the availability of appropriate in-domain texts, and b) fluency of input: users who are less familiar with the application may hesitate and therefore speak less fluent to the system.

It is well known that the performance of any ASR module in a dialogue system critically depends on the user's level of expertise and familiarity with the system. The knowledge that experienced speakers have about the limitations and capabilities of the system, such as the expected successful interpretation of particular words and phrases, can be used to guide the dialogue to a successful completion. The poor robustness of ASR systems against out-of-grammar utterances explains a great part of the difference in ASR performance between experts and naïve users – a difference which is even shown by commercial, highly developed systems (Glass, 2004).

Unfortunately, there are still no good recipes for the construction of a language model when the amount of training data is limited. Language models are sensitive to changes in style and genre of the texts on which they are trained, and this holds in particular for small training corpora (Rosenfeld, 2000).

Given the potentially poor robustness of the ASR in dialogue systems, several research groups have addressed the question of how to improve the robustness by using knowledge sources other than the LM. For example, language understanding modules and dialogue management may compensate for errors made by the ASR (e.g., Gorrell, 2003). The use of different parsing techniques has shown to yield improvements in specific cases (e.g., Wang et al, 2002). Also the wording in the system prompts can be used to gently guide the user towards expressions that are known by the ASR (e.g., Witt et al., 2003).

The ASR performance itself is sometimes improved by using linguistically motivated language models, such as long span agreements, probabilistic context free grammars and the use of latent syntactic analysis (Jurafsky and Martin, 2000). Also the use of dynamic LMs (i.e. LMs dependent on the state of the dialogue) is usually advantageous to increase ASR performance.

When there is a reasonable amount of training text, an integration of statistical language models (SLMs) and grammars (by N-gram boosting) may improve ASR performance in dialogue systems (Akiba et al., 2004; Goel, 2004). For very small amounts of training texts, these mixing techniques seem less feasible, but SLMs may show an advantage compared to grammars (e.g., Rayner et al., 2004). Grammars are usually straightforward to build and show a low word error rate on in-coverage (i.e. grammatical) utterances, but may show high error rates on out-of-coverage (i.e. ungrammatical) utterances. Compared to grammars, SLMs show comparable or slightly worse performance on in-coverage data, but they do better on out-of-coverage data in terms of word correctness (e.g. Gorrell, 2003; Knight et al., 2001). Moreover, it is often easier to relate an explicit training corpus to an SLM than to relate an 'example corpus' to a (handcrafted) grammar (Rayner er al., 2004).

In this paper, we will specifically compare the performance of an ASR system that uses a grammar or statistical language models created by different techniques. We extend this comparison by taking into account SLMs that include a 'garbage word' (cf. ten Bosch & Boves, 2004). The ASR is tested on a speech corpus with utterances of users interacting with a multimodal dialogue system. Furthermore, we will focus on the ASR performance in terms of word correctness, rather than on NLP-related measures such as concept error rates.

In the following sections, we will first discuss the context of the dialogue system that was used to log the acoustic data. The description of the dialogue system gives more insight in the background of the used speech data. Next, the language

modelling approaches, and results using two different scoring methods will be presented.

2. The COMIC dialogue system

2.1. Introduction

The European FP5 project Conversational Interaction with Computers (COMIC, http://www.hcrc.ed.ac.uk/comic/, 2002-2005) aimed at a study of fundamental aspects of multimodal interaction between users and a multimodal dialogue system. The research focus of the project was on the relation between the behaviour of subjects as function of the functionality of the system (Rossignol et al., 2003; Vuurpijl et al., 2004; Boves & den Os, 2003). The research was partly steered by experiments that were carried out with demonstrators of increasing complexity. The COMIC dialogue system was designed to interactively assist the user in designing a new bathroom, in a way that mimics human-human interaction and dialogue.

The multimodal COMIC system comprised input decoders for speech (ASR) and pen input (Pen Input Interpretation), an NLP/FUSION module that first performs natural language processing and next merges the information from the speech and pen modality, a dialogue and action manager (DAM) that controls the dialogue including error handling, a FISSION module that prepares the output of the system, and the actual output modules (speech synthesis, graphics, and an avatar). The user could interact with the system via a head mounted close talk microphone and a Wacom Cintiq 15X LCD tablet. The public domain version of the HTK speech recognition system (version 3.1, later version 3.2.1) was modified in order to use it in an ASR module that could interact with the other modules.

A full user-system dialogue consists of four 'phases'. In the first phase of the dialogue, the user specifies the lay-out of the ground floor and the bathroom dimensions (by specifying the sizes of walls) as well as the location and size of sanitary ware. This phase also allows the user to move doors and windows by using spatial references in utterances ('move the window 1 meter to the door') or by pen gestures. In the second phase, the system proposes a limited number of designs that match the spatial specifications as provided by the user in the first phase. In the third phase, the user is supported in shaping the design, by specifying tiles, colours, styles, borders and decorations. The final phase shows a three-dimensional tour through the newly designed bathroom (see Rossignol et al., 2003; Vuurpijl et al., 2004).

All utterances by the users during these experiments have been logged and annotated afterwards. In parallel, the system was subject to tests during its construction and modification on a continuous basis. In this way, we collected a growing number of real-life recorded audio files.

The length of each logged utterance was fully determined by the moments on which the microphone was opened and closed. The moment of opening the microphone was determined by the FISSION module (on the basis of the moment when the prompt was finished); the moment of closing the microphone was determined by the built-in end-of-speech detection in HTK which was set to trigger after 1.2 seconds of pause.

2.2. Speech data

Three databases have been collected (see Table I). Databases A and B contain utterances from Dutch subjects interacting with the system during the actual experiments; database C contains recordings from an expert user, recorded during the various intermediate tests. The language used in these experiments is English; all speakers had a good command of the English language. Prior to the actual experiment, subjects were informed about the purpose of the interaction. However, they were not instructed about the type of utterances that they could use in order to specify the requested information about the bathroom and sanitary ware.

Table I. Overview of the databases used in this study.

Database	Subjects (non-native)	Nr of uttterances	Nr of non-silent utterances
A	18 naïve	571	309
B	22 naïve	1491	785
C	1 expert	400	400

The fourth column in Table I shows the actual number of utterances that were non-silent. Due to the character of the interaction, subjects often do not say anything during their turn, for example when they are asked to draw the shape of the bathroom. The non-silent utterances primarily contain filled pauses, length specifications, move commands, spatial references, specifications about tiles, colours, borders and decorations. Furthermore, 25 (10%), 78 (9%) and 0 utterances in A, B and C, respectively, contain speech that is not addressed to the system at all (most of these utterances contain off-topic observations). These off-topic utterances were kept in the database to specifically evaluate the out-of-coverage performance of the language models.

All recordings have been stored as little endian mono sdf files with sample frequency 16 kHz, 16 bits/sample.

2.3. Grammar-based and Statistical Language Models

At the beginning of the COMIC project, no training data was available from the domain of bathroom design for building an SLM. Therefore, first a finite state grammar was developed with out-of-the-blue example utterances and one specifically elaborated example dialogue as starting points. The example dialogue was part of the formal specification of the COMIC demonstrator. The grammar was constructed such that the most likely answers from users were covered, including meta-commands such as 'I want to quit', 'go back', repair commands ('erase this'), and more polite versions such as 'please erase this'. For example, probable user replies to the system prompt 'what is the length of the fourth wall' were covered by putting in parallel the following options, after specifying the variable `$length` by

```
$number meters [ and $number centimetres ]

$length
the length [ of this wall ] is $length
this wall is $length [ long ]
```

The grammar contained optional sub-phrases (between '[' and ']') and various loops – these loops were introduced to allow the user to provide multiple utterances within one turn. In this way, each dialogue state was associated with a limited

number of options in the grammar. All options had equal weight (i.e. in the resulting lattice, each progression is chosen via a uniform random selection). In later phases of the project, the grammar was updated and refined on the basis of utterances recorded during various intermediate tests.

In order to compare the performance of the ASR using a grammar and using an SLM, a bigram LM has been constructed on the basis of a text corpus that was generated by using the grammar in 'generation mode'. The exact procedure was as follows:

a) In the first step, the grammar was updated such that it covered the example dialogue and the verbatim transcription of all user utterances collected during the intermediate tests. Only off-topic utterances were excluded. (The utterances that were collected during the actual experiments served as test set and were therefore not included in the LM.)

b) The HTK tool HSGen was used to randomly generate in-coverage sentences. In this way, a text corpus was made with 200k utterances of word length < 15. The number 200k was chosen to guarantee that all possible sentences that were contained in the grammar were also covered by the bigram LM (the number of different sentences that could be generated by the grammar with all loops removed was about 89k). The constraint of the sentence length served to avoid ultra-long sentences that resulted from the various loops in the finite-state grammar. It was verified that the restriction on sentence length did not significantly influence the eventual bigram.

c) Next, this corpus has been copied into a variant in which garbage words ('GARB') were introduced. For example, if the sentence 'A B B C' occurs in (b), the sentence 'GARB A GARB B GARB B GARB C GARB' is constructed in step (c).

d) The LM training corpus was made by joining the text constructed under (b) and the text created under (c) in a user-specified ratio (see below).

e) On the text constructed under (d), one bigram LM was constructed via the HTK tool HLStats by an absolute discount of 0.5 and a floor probability of 0.02 for bigrams. A second bigram was made by using the Good-Turing discounting.

This construction of the SLM is related to the way described in Knight et al. (2001), with the difference that in their case 200 utterances have been used to create an initial SLM and a few thousand utterances for updated SLMs. Moreover, they also trained trigrams and used LM-classes consisting of names of rooms and devices. In our case, trigrams and class-based LMs were left out of consideration because there were not directly applicable in the HTK recogniser.

3. Results

The performance results of the ASR have been evaluated for six corpora, four types of language models, and two different ways of evaluating the word correctness. Three of the corpora are the original corpora A, B and C; the other three are sub-corpora consisting of the grammatical utterances (denoted A-gram, etc.). An utterance is defined as 'grammatical' if its verbatim annotation is exactly included by the finite-state grammar.

In Table II, the ASR performance values (in terms of percentage word correct) are shown. Each row refers to a corpus. The columns refer to the finite-state grammar (indicated by Grammar), the bigram LM including garbage (SLM1+G), the same bigram but without garbage (SLM1), and the SLM based on the Good-Turing method, also without garbage (SLM2). For the language model SLM1+G, we have used a ratio of 10:1 in step (d) for the text without garbage compared to the text with garbage, such that the unigram probability of the garbage word was 0.05. To avoid tuning on the test set, this ratio was not further tuned on these test corpora. With SLM1+G, the test set bigram perplexities of the six sets are 18.1, 8.6, 15.4, 7.3, 8.8 and 8.5, respectively.

The upper part of Table II shows the default NIST scoring results. This scoring method is not entirely adequate in the case that garbage words have to be aligned, because the scoring treats the token GARB as a normal word. An example of the effect is given below:

```
REF this wall is two meters
HYP this wall GARB GARB GARB meters
```

According to the original NIST scoring, the reference-hypothesis pair (REF-HYP) would lead to 2 substitutions, 1 insertion, and 3 correct words. Thus, the insertion of each additional GARB in the sequence of garbage words is penalised. As a compromise, a second scoring algorithm was designed that *removes* a GARB if it was inserted while being part of a GARB sequence, and, reversibly, *adds* a GARB if it was considered as part of a deletion and a part of a GARB sequence. The resulting scores, presented in Table II (bottom), give a fairer account of the reference-hypothesis mismatch.

4. Discussion and conclusion

On the basis of the upper part of Table II, we can make the following observations. Firstly, when comparing A with A-gram and B with B-gram, the difference in ASR performance is largest for the case in which the language model is the finite-state grammar itself. This is in line with earlier results (cf. Knight et al., 2001). For obvious reasons this observation also holds for the bottom part of Table II. Comparing language models, we observe that the statistical language model that includes garbage (SLM1+G) performs best for A and B, but that the grammar and the SLM1 (without garbage) outperform SLM1+G on A-gram, B-gram, C and C-gram. The differences between SML1 and SLM2 (Good-Turing) are small – the added value of the Good-Turing method probably disappears because there is no large subset of LM tokens with very low counts. Knight et al. found an advantage for using the SLM compared to the grammar for out-of-coverage data, but in our data this effect was not clearly visible, which might indicate that the gain of a flexible decoding by the SLM is compensated by the loss of longer span information. A clear advantage of SLMs is their flexibility to introduce garbage words which evidently lead to a better performance on the corpora A and B for both scoring methods. For example, step (c) in section 2.3 can be refined by inserting GARB in linguistically motivated locations.

A comparison between C and A or B shows the difference between an expert and naïve subjects (cf. Glass, 2004). However, the difference is not only a matter of coverage – it is also the familiarity with the interface itself that probably plays a role, as follows from the difference between A-gram, B-gram and C-gram. Furthermore, part of the differences between A and B on the one hand and A-gram and B-gram on the other hand can be attributed to the occurrence of off-topic utterances in A and B.

Table II. Top: results (percentages word correct) obtained by the regular NIST scoring without correction for GARB alignments. Bottom: results obtained after correction for GARB alignments. For technical reasons, the results for SLM2 are not available for the sets B and B-gram.

Corpus	Grammar	SLM1+G	SLM1	SLM2
A	48.5	51.5	49.9	49.1
A-gram	70.4	64.7	68.4	68.8
B	55.2	61.3	54.1	-
B-gram	76.9	73.3	74.2	-
C	92.7	89.0	92.1	91.5
C-gram	93.7	89.0	91.6	91.6

Corpus	Grammar	SLM1+G	SLM1	SLM2
A	48.5	61.0	54.1	55.3
A-gram	71.0	73.6	68.4	68.2
B	55.4	66.0	56.8	-
B-gram	78.4	75.0	76.1	-
C	93.1	95.1	93.1	92.3
C-gram	93.3	95.2	93.2	93.9

We can conclude that, given all possible cases considered here, and given the limitations indicated (limited training data, no trigrams, no class-based LM), the results suggest that a bigram model, enriched with flexible and adjustable garbage penalties, is one of the best options. For a fair account of the results, the alignment must treat garbage words in a way different from words.

5. Acknowledgements

Earlier comments from colleagues improved this text. The first author participated in the FP5 project COMIC (IST-2001-32311). Johan de Veth created the Good-Turing SLM.

6. References

[1] Akiba, T., Fujii, A., Itou, K. (2004). Effects of language modelling on speech-driven question answering. Proc. Interspeech 2004, Korea (cd-rom).

[2] Boves, L.W.J., Neumann, A., Vuurpijl, L.G., Bosch, L.F.M. ten, Rossignol, S., Engel, R., Pfleger, N. (2004). Multimodal Interaction in Architectal Design Applications. In Proceedings 8th ERCIM Workshop on "User interfaces for all" (UI4All) workshop, Palais Eschenbach, Vienna, Austria. (cd-rom).

[3] Glass, J. (2004). Tutorial on Conversational interfaces. Proc. Interspeech 2004, Jeju Island, Korea (cd-rom).

[4] Goel, V. (2004). Conditional maximum likelihood estimation for improving annotation performance of N-gram models incorporating stochastic finite-state grammars. Proc. Interspeech 2004, Korea (cd-rom).

[5] Gorrell. G. (2003). Recognition error handling in spoken dialogue systems. Proc. Sec. Intern.Conf. on Mobile and Ubiquitous multimedia, Norrköping, Sweden (cd-rom).

[6] Jurafky D., Marin, J. (2000). Speech and language processing. Prentice Hall, New Jersey.

[7] Knight, S., Gorrell, G., Rayner, M., Milward, D., Koeling, R., Lewin, I. (2001). Comparing grammar-based and robust approaches to speech understanding: a case study. Proc. Eurospeech (cd-rom).

[8] den Os, E.A and Boves, L. (2003). Towards Ambient Intel-ligence: Multimodal computers that understand our intentions, Proc. eChallenges, Bologna, October 2003.

[9] Oviatt, S., (2000). Taming Speech Recognition Errors Within a Multimodal Interface. In: *Communications of the ACM,* vol. 43 (9), pp. 45-51.

[10] Rabiner, L. & Juang, B.-H. (1993). Fundamentals of speech processing. New Jersey: Prentice Hall.

[11] Rossignol, S., ten Bosch, L., Vuurpijl, L., Neumann, A., Boves, L., den Os, E., de Ruiter, J-P. (2003). Human-Factors issues in multi-modal interaction in complex design tasks. Human Computer Interaction Conference. Greece, June 2003. pp. 79-80.

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

[13] Rosenfeld, R. (2000). Two decades of statistical modelling: where do we go from here? Proceedings of the IEEE, vol. 88 (8), pp. 1270-1278.

[14] Ten Bosch, L., Boves, L. (2004). Survey of spontaneous speech phenomena in a multimodal dialogue system and some implications for ASR. Proc. Interspeech 2004, Korea (cd-rom).

[15] Vuurpijl, L.G., Bosch, L.F.M. ten, Rossignol, S., Neumann, A., Pfleger, N., Engel, R. (2004). Evaluation of multimodal dialog systems. In Proceedings of LREC 2004, Workshop on Multimodal Corpora (cd-rom).

[16] Wang, Y., Acero, A., Chelba, C., Frey, B., Wong, L. (2002). Combination of statistical and rule-based approaches for spoken language understanding. Proc. ICSLP 2002, Denver, Colorado (cd-rom).

[17] Witt, S., Williams, J. (2003). Two studies of open vs. directed dialog strategies in spoken dialog systems. Proc. Eurospeech 2003 (cd-rom).