LANGUAGE MODELLING FOR TASK-ORIENTED DOMAINS

Cosmin Popovici

ICI - Institutul de Cercetari in Informatica
Bd. M. Averescu, 8-10
Bucuresti (Romania)

Paolo Baggia

CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli, 274
I-10148 Torino (Italy)
baggia@cselt.it

ABSTRACT

This paper is focused on the language modelling for task-oriented domains and presents an accurate analysis of the utterances acquired by the Dialogos spoken dialogue system. Dialogos allows access to the Italian Railways timetable by using the telephone over the public network.

The language model aspects of specificity and behaviour to rare events is studied.

A technique for getting a language model more robust, based on new sentences generated by grammars, is presented. Experimental results already show its benefit. The relative increment between usual language models and language models created using grammars, is higher for small amount of training material. Therefore this technique can advantage especially the development of the LM for a new domain, in the first phases.

1. INTRODUCTION

Statistical language modelling (LM) is currently used for two different classes of applications: dictation systems and task-oriented spoken dialogue systems (SDS).

The first kind of systems are tested with a very large vocabulary (60-20,000 words) and they need the availability of a huge amount of training data, for instance WSJ-NAB has a 45 million word text corpora [8].

SDSs are used in specific task-oriented domains, and they need special training material, which can be obtained either by expensive simulations [6] or by using the SDS itself. The use of a general task-independent corpus for LM of a SDS could increase, in comparison to LM that use a task-dependent one, the perplexity by an order of magnitude [9]. This is due to the mismatch between the general corpus and the specific application domain. In any case the acquired material is very limited, for instance the LM in the Air Travel Information System (ATIS) is based on a training-set of only 250,000 words [10].

This paper is focused on the language modelling for task-oriented domains. The tests made uses the utterances acquired by the Dialogos, the SDS which allows access to the Italian Railways timetable by using the telephone over the public network [1]. Other similar systems are described in [2,5,7].

The vocabulary of Dialogos contains 3,471 words, clustered in 358 classes. The semantically important words are grouped into classes, such as city names (2,983 words), numbers (76 words), and so on. During the recognition, a class-based bigram LM is used, and the 25-best sequences are rescored using a trigram LM.

Section 2 shows how well a LM captures the specificity of the domain, while Section 3 studies the behaviour of the LM to rare events. Finally Section 4 illustrates a technique for generalising a LM by adding n-grams generated by a grammar.

2. SPECIFICITY OF A LANGUAGE MODEL

A relevant characteristic of a task-oriented domain is the distribution of the user utterances in a corpus. Using the Dialogos SDS, a corpus of 1,363 spoken dialogues has been acquired, from 493 unexperienced subjects, that called the system from all over Italy [1,3].

For the present study, the collected material was divided into two parts: a training-set of 20,511 utterances and a test-set of 2,040 utterances. Each utterance was transformed in a normalised form (NU), by changing each city name, month name and number into a class tag. For instance the user utterance:

"I want to leave from Naples to Rome Monday at five (o'clock)"

becomes the following NU:

"I want to leave from CITY-NAME to CITY-NAME WEEK-DAY at HOUR-NUMBER".

For the sake of the language modelling, the NU is equivalent to the original utterance[1].

It is worth noticing that even a small number of very frequent NUs cover a great part of the acquired data (see Figure 1). The 7-th most frequent NUs cover 58% of the training-set, and 54% of test-set, and the first 191-st cover nearly 80% of test-set and over 85% of training-set. On the other hand the NUs with just one occurrence are 2,060, and more then 56% of them contain some spontaneous speech phenomena. This result shows that a few frequent NUs can already give a quite sensible picture of the user utterance distribution.

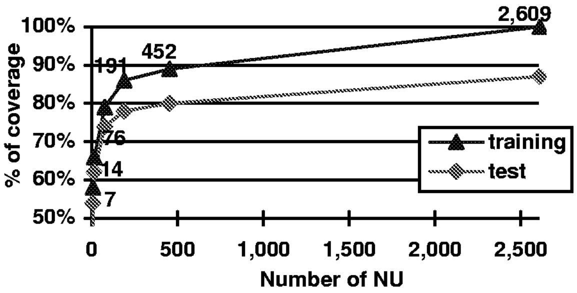

Figure 1. Coverage of training and test sets by the NUs

[1] This is because these classes are being used by the class-based LM and each word in a class has been considered with equal probability.

Moreover some partial training-sets were selected, which include the first ***n*** utterances in the whole training-set, for ***n*** ranging from 100 to 20,511 utterances. For each partial training-set a LM was created and the recognition (WA) and understanding (SU) rates are given in Figure 2. The performances of the LMs created on a partial training-set were compared with an experiment without any LM, which is even reported in Figure 2 as 0-utterance training-set. A LM trained on only 100 utterances achieves a remarkable error rate reduction of 30% of SU and 23% of WA, especially if it is compared with the error reduction when the whole 20,511 training-set is used, that is of 43% of SU and 39% of WA.

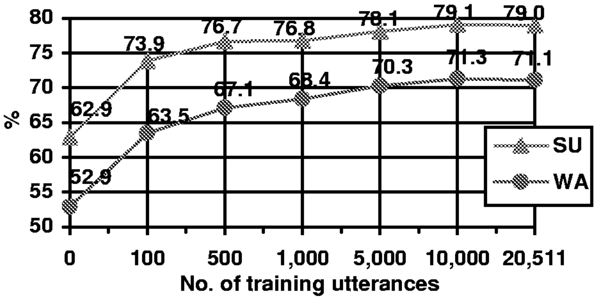

Figure 2. Variation of performance with size of training data

A coherent behaviour is also confirmed by perplexity values (PP) depicted in Figure 3, where the utterances were classified according to the kind of prompt generated by the system. Three representative points have been selected, which are the request of: departure and arrival city (**City**), time of departure (**Time**), and date of departure (**Date**). For these categories the PP of a 100-utterance LM is two times higher than a 1,000-utterance one and three times the LM trained on the whole training-set. The fact that, the PP values for the **City** requests are the highest, can be explained by the large number of city-names in the vocabulary (2,983, near 85% of the whole vocabulary).

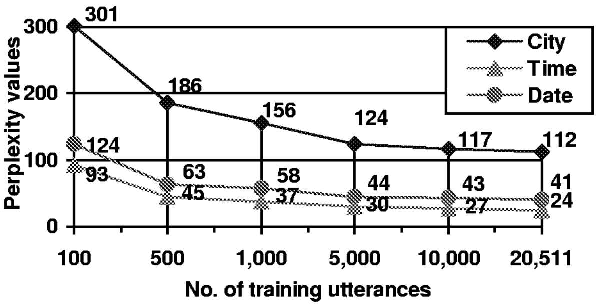

Figure 3. Variation of perplexity with size of training data

3. ROBUSTNESS TO RARE EVENTS

In this Section the behaviour of the LM with respect to rare events is studied. The test-set of 2,040 utterances was split into two parts: The first part contains 362 utterances, whose 351 NUs do not appear in any of the partial training-sets. This is referred below as the unseen part of the test-set. The second part includes the rest of the test-set (1,678 utterances, but only 257 NUs). The NUs in the partial training-sets cover progressively the utterances of the second part. For instance, the 100-utterance training-set contains only 29 NUs, which cover 1,317 of these 1,678 utterances.

Both recognition, and overall understanding results show quite similar values for the 1,678 utterances (82-85% of SU), but they are very different for the unseen part (33-46% of SU), see Figure 4. The performance on the unseen part is an indicator of the robustness of the model. In the following the reason for the low performance on the unseen part is further analysed.

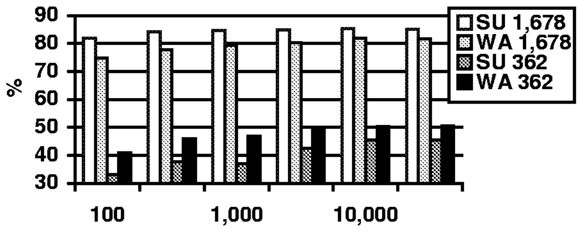

Figure 4. Evaluation of trained and untrained part of the test-DB

The NUs with more then three occurrences in the global training-set, and different one to each other, were selected. Table 1 shows the number of this NUs, that exists in each one of the partial training-sets. They were divided into groups according to the different kind system request. The growth of NUs for **City** and **Date** is fast until 5,000 utterances are reached, then it becomes very slow. This indicates that there is a kind of saturation. While **Time** NUs increase nearly proportionally.

	training utterances					
	100	500	1,000	5,000	10,000	20,511
City	11	17	26	43	46	47
Date	8	17	25	43	48	51
Time	6	12	15	36	42	49

Table 1. Number of frequent NUs in partial training-sets

Moreover, the NUs, whose frequency in the training-set is greater than 0.1%, were compared with the ones in the test-set. We observed that the selected NUs of the training-set covers more then 90% of the test-set NUs, in case of **City** and **Date**, but only 55% in case of **Time**. Therefore, the **City** and the **Date** groups are considered much more robust than the **Time** group, because the frequent NUs do not indicate a saturation, and because there is a lack of the training-set NUs in the test-set. This is due to the high variability of the time expressions.

4. INCREASING ROBUSTNESS BY ADDING N-GRAMS GENERATED BY GRAMMARS

Another coverage test was made using grammars. A grammar was created (explained in Section 4.1) on the basis of the NUs in the 500-utterance training-set. The sentences generated by the grammar showed a coverage of 85% of the NUs in the 20,511 training-set. This suggests that, the robustness of a LM may be increased by the use of a simple grammar derived from the common NUs in the training material.

At first the sentences generated by grammar were added to the training material. The obtained LMs, did not improve results, because the addition of the grammar generated sentences, greatly changes the frequency distribution of the n-grams, and reduces the specificity of the training-set.

The adopted solution was to create the LM starting from a data-base that contains n-grams, and not from a data-base of generated sentences. This made possible to add only the not-existing n-grams which do not highly affect the specificity. Therefore the tool used for training the LMs was changed, in order to be able to process both sentences and n-grams. Commonly when the n-grams are extracted from a sentence, they get automatically all their contexts (the (n-1)-gram that precedes the n-th word of the n-gram). On the other hand, if an n-gram is artificially added, it is necessary to incorporate even the missing contexts for this n-gram.

4.1. Grammar creation

The grammars used in the following tests were manually created, and they started from a set of correct NUs selected from a training-set. For each NU, semantic concepts were identified, then for each of these concepts a non-terminal was introduced, and, finally, each non-terminal was generalised. For instance, in the case of a **Time** NU:

"in the morning after seven o'clock",

the following non-terminal sequence could be identified:

"*Part_of_Day Time_Specifier Time_Identifier*".

Part_of_Day can become also "*in the afternoon*", "*in the evening*" or "*at lunch time*", *Time_Specifier* can be expressed as: "*before*", "*not earlier than*", while for *Time_Identifier* other forms are: "*a quarter to seven*", "*twenty minutes past seven*".

At this point both the 1,000-utterance training-set (SPTS-1,000) and the global one (STS) were split according to the system request. Concentrating the analysis on the **City**, **Date**, and **Time** requests, for the syntactically and semantically correct NUs in SPTS-1,000 a grammar was created. For instance, there are 107 NUs in the SPTS-1,000 **Date** requests, and 2,483 NUs in STS.

For **Date** and **Time** requests group one grammar was created (Gr_D, and Gr_T respectively), whereas two for the **City** requests: Gr_C which generalises only NUs about departure and arrival location, and Gr_Cdt which also generalises data and time, because the answers to the **City** requests could also contain that information.

4.2. Creation of generalised LMs

The merge between the n-grams extracted from a training-set and from sentences generated by the grammar was done using the following technique. At first, both the training-set and the sentences generated by a grammar were transformed in n-grams (n=3), then three type of events were considered: n-grams which are present both in the training-set and in the generated sentences (called *usual events*), n-grams which exist only in the training-set (called *rare events*), and n-grams which exist only in the generated sentences (called *unknown events*).

Into the new LM, the unknown events were added only once, while the rare events maintained their frequencies (which is quite low). In many cases the number of unknown events is much more higher than the number of usual events. For instance in the case of time there are 276 usual events obtained from SPTS-1,000, 36 rare events and 1,748 unknown events. Therefore the quantities of usual and unknown events are weighted, by multiplying them with a balance-factor. At this point, a language model is created, then the best value for the balance-factor (BaFa) is empirically determined by the minimisation of the PP on the test-set.

request	grammar	usual		rare		unknown		BaFa	
groups	used	part	all	part	all	part	all	part	all
City	Gr_C	534	9861	159	859	166	43	1	1
City	Gr_Cdt	568	10088	125	632	1326	1156	4	2
Date	Gr_D	316	6921	96	849	150	82	1	1
Time	Gr_T	276	6431	36	616	1748	1488	10	1

Table 2. Event composition of the training-sets

Using Table 2, the event composition of each one of the studied LMs can be computed. For each request group many LMs were created by the generalisation of SPTS-1,000 and STS, respectively *part* and *all* in the Table. It is worth noticing that in a baseline LM only the usual and rare events are considered.

4.3. Experimental Results

In this Section, the performances of the LMs that include n-grams generated by a grammar were compared with baseline LMs which does not make use of grammar n-grams. These baseline LMs are reported in the Tables 3-6, with the tag *unused* in the grammar column.

request	grammar	SPTS-1000		STS	
groups	used	WA	SU	WA	SU
City	*unused*	77.5	68.5	82.3	71.4
City	Gr_C	78.8	69.3	82.5	71.4
City	Gr_Cdt	80.0	70.1	82.3	72.6
Date	*unused*	82.0	80.9	82.8	80.9
Date	Gr_D	82.7	81.3	82.9	80.9
Time	*unused*	79.7	85.5	83.6	86.7
Time	Gr_T	82.5	86.1	83.6	86.7

Table 3. Recognition and understanding results

Table 3 shows that the LMs created using the grammars, obtain better results for the SPTS-1,000 LMs, while for the STS LMs the increment is rather limited. In particular, for **Time** and **City** the improvement of WA is significant. The reasons are: the high variability of time expressions and the fact that sometimes the **City** requests even include information about **date** and **time**, especially in the first utterance to the system. This fact is evident from the improvement obtained by the use of the Gr_Cdt grammar, which even increases the performance of the STS LM.

Moreover the merge of with SPTS-1,000 with grammars improve the results, but they could not reach the performances of the baseline STS LMs. An explanation is that the used grammars do not model the highly frequent extra-linguistic phenomena.

In addition the perplexity of these LMs has been studied. For each group the analyses of the PP has been performed on the test-set and even on the sentences generated by the grammar. Table 4 shows PP results for all the LMs tested on the specific part of the test-set. The generalisation of the LMs by using grammar n-grams does not significantly affect the PP.

request	grammar	SPTS-1000	STS
groups	used	PP	PP
City	*unused*	117	79
City	Gr_C	118	78
City	Gr_Cdt	122	96
Date	*unused*	33	24
Date	Gr_D	32	24
Time	*unused*	20	14
Time	Gr_T	19	15

Table 4. Perplexity results on the test-set

The use of a test-set of sentences generated by the grammars, even if it does not give a correct insight of the behaviour of the system on a test-set acquired from real users, because the sentence distribution is artificial, it can show the degree of generalisation. These PP results have been reported in Table 5 and Table 6 according to the number of unknown events reported in Table 2. In the former are shown the results for small grammars (G_C, and G_D), while in the latter the results for large ones (G_Cdt, and G_T).

request	grammar	SPTS-1000	STS
groups	used	PP	PP
City	*unused*	72	36
City	Gr_C	24	24
Date	*unused*	52	25
Date	Gr_D	17	16

Table 5. Perplexity results on the grammar sentences.

request	grammar	SPTS-1000	STS
groups	used	PP	PP
City	*unused*	177	207
City	Gr_Cdt	36	42
Time	*unused*	231	53
Time	Gr_T	12	11

Table 6. Perplexity values for Gr_C and Gr_T

In Table 5, a clear reduction of the PP could be observed for the LMs which includes grammar n-grams. This reduction is higher for the LMs trained over SPTS-1,000 (66%), but it is relevant even for the LMs trained on STS (33%).

Making a similar comparison of the PP results, presented in Table 6, for the large sets of unknown evens, as expected, a more significant reduction was obtained, that goes from a minimum of 77% to a maximum of 94%.

5. CONCLUSIONS

This papers shows that, in a task-oriented domain, a LM trained out with a small amount of material (about 1,000 utterances) acquired form naive users allows to obtain rather good results, in the case of the more common NUs. Because this common NUs are a few but very frequent, also the result obtained over all NUs, decrease a bit.

Secondly, in a task-oriented domain with a very limited training-set (about 1,000 utterances), the robustness of a LM was increased by the use of a simple grammar derived from the common NUs in the training material.

A technique for the generalisation of a LM adding n-grams generated by a grammar is described. The advantage of the use of this technique is reflected by the experimental results. The improvements obtained using this technique, are explicitly good for LMs trained with small training material, and therefore the first phases of the development of the LM, for a new domain, can be advantaged. Even if the generalised LMs does not increase recognition and understanding results, the obtained perplexity values indicates a better behaviour in case of the events.

References

[1] Albesano D., P. Baggia, M. Danieli, R. Gemello, E. Gerbino, C. Rullent, "Dialogos: A Robust System for Human-Machine Spoken Dialogue on the Telephone", in *Proc. of ICASSP'97*, München, 1997, to appear.

[2] Aust H., M. Oerder, F. Seide, V. Steinbiss, "The Philips Automatic Train Timetable Information System", in *Speech Communications*, 1995, vol. 17, pp. 249-262.

[3] Baggia P., E. Gerbino, E. Giachin, C. Rullent, "Experiences of Spontaneous Speech Interaction with a Dialogue System", in *Proc. of CRIM/FORWISS Workshop*, München, 1994, pp. 241-248.

[4] Besling S., H.-G. Meier, "Language Model Speaker Adaptation", in *Proc. of EUROSPEECH'95*, Madrid, 1995, pp. 1755-1758.

[5] Eckert W., T. Kuhn, H. Niemann, S. Rieck, A. Scheuer, E.G. Schukat-Talamazzini, "A Spoken Dialogue System for German Intercity Train Timetable Inquiries", in *Proc. of EUROSPEECH'93*, Berlin, 1993, vol. 3, pp. 1871-1874.

[6] Fraser N., G.N. Gilbert, "Simulating Speech Systems", in *Computer Speech and Language*, 1991, vol. 5, pp. 81-99.

[7] Goddeau D., E. Brill, J. Glass, C. Pao, M. Phillips, J. Polifroni, S. Seneff, V. Zue, "GALAXY: A Human Language Interface to On-line Travel Information", in *Proc. of ICSLP'94*, Yokoama, 1994.

[8] Gauvain J.L., L. Lamel, G. Adda, D. Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB New Task", in *Proc. of ICASSP'96*, Atlanta, 1996, vol.1, pp. 73-76.

[9] Placeway P., R. Schwartz, P. Fung, L. Nguyen, "The Estimation of Powerful Language Models from Small and Large Corpora", in *Proc. of ICASSP'93*, Minneapolis, 1993, vol. 2, pp. 33-36.

[10] Ward W., S. Issar, "Recent Improvements in the CMU Spoken Language Understanding System", in *Proc. of ARPA HLT Workshop*, March 1994, pp. 213-216.