Training Language Models without Appropriate Language Resources: Experiments with an AAC System for Disabled People

Tonio Wandmacher, Jean-Yves Antoine

Université François-Rabelais de Tours
Laboratoire d'Informatique – Equipe BdTln
3 place Jean Jaurès, 41000 Blois, France
E-mail: tonio.wandmacher@univ-tours.fr, jean-yves.antoine@univ-tours.fr

Abstract

Statistical Language Models (LM) are highly dependent on their training resources. This makes it not only difficult to interpret evaluation results, it also has a deteriorating effect on the use of an LM-based application. This question has already been studied by others (e.g. Bellegarda, 2004). Considering a specific domain (text prediction in a communication aid for handicapped people) we want to address the problem from a different point of view: the influence of the language register. Considering corpora from five different registers, we want to discuss three methods to adapt a language model to its actual language resource ultimately reducing the effect of training dependency: (a) A simple cache model augmenting the probability of the *n* last inserted words; (b) a user dictionary, keeping every unseen word; and (c) a combined LM interpolating a base model with a dynamically updated user model. Our evaluation is based on the results obtained from a text prediction system working on a trigram LM.

1 Introduction

Language models depend strongly on the similarity of the training data with the task. In particular, they have reduced capacities of generalization apart from the style of language on which they are trained. In the domain of speech recognition, evaluation campaigns showed that journalistic corpora are well adapted to the automatic transcription of broadcast news. On the contrary, such corpora are completely inoperative on spontaneous speech.

A simple solution to this problem consists in building specific resources for every task. This solution is however expensive. For this reason it is generally preferred to interpolate a background language model with a second one trained on a specific corpus (Woodland et al, 1998; Bellegarda, 2004).

This paper investigates the relevance of some adaptation techniques for difficult tasks, where the language to be modelled strongly varies according to the context of use. Such situations raise the question of adaptation when very limited data are available. This is particularly the case in the domain of *Augmentative and Alternative Communication* (AAC) for disabled people, where every association of a specific patient and a communication goal (loose conversation, official or private correspondence, literary writing etc.) will constitute a very specific situation of communication.

At first, we will present the problem of word prediction for AAC systems. Then, we will detail experiments which show the influence of register, as defined by Biber (1993), on language models. We will then compare well-known techniques of adaptation such as the cache model, a user dictionary and a dynamic interpolated user model. In conclusion, we will explain our ideas on using information provided by *Latent Semantic Analysis* (LSA) for an adaptation to the topic of the current context of communication.

2 AAC: the Problem of Missing Data

AAC aims at restoring communicative abilities for persons with severe speech and motion impairments (cerebrally and physically handicapped persons). Whatever the disease considered, oral communication is impossible for these persons who have also serious difficulties to control physically their environment. In particular, they are not able to use input devices of a computer. Communication with an AAC system means communicating by the help of a table of symbols (words, letters or even icons), where the handicapped person selects successively item after item. The selection is achieved by pointing on a virtual keyboard displayed on the screen of the computer.

Basically, an AAC system consists of four components: At first, a physical input interface connected to a computer. This interface is adapted to the control capacities of the user. Often, these only amount to a binary reply (e.g. an eye glimpse): the control of the environment is therefore restricted to a *Yes/No*-command.

Secondly, a virtual keyboard allowing the user to select successively symbols to compose messages. In our *SIBYLLE* AAC system (Schadle et al, 2004), key selection is achieved by a linear scan: a cursor highlights successively each key, which can then be selected.

The last two components are a text editor (to write e-mails or other documents) and a speech synthesis which is activated in case of oral communication.

The main weakness of AAC systems results from the slowness of message composition (on the average 1 to 5 words per minute). Moreover, this task is extremely tiring for the patients.

Two complementary approaches are possible to speed up communication. The first aims at optimizing the selection on the virtual keyboard : most probable symbols are dynamically presented at first on screen. The second improvement consists in minimizing the number of

keystrokes: the system tries to predict the words which are likely to occur just after those already typed. Several approaches can be used to carry out this prediction, among which language models that provide a list of word suggestions, depending on the *n* (typically 1-4) last inserted words. Other, more complex models (structural model, e. g. Schadle et al, 2004) can be used, but we will limit ourselves here to a tri-gram model giving already satisfactory results.

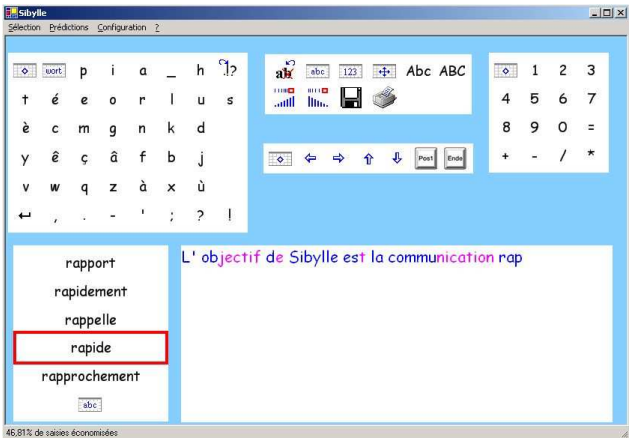

Figure 1: The interface of the SIBYLLE AAC system

As can be seen in figure 1, after each keystroke, a list of (usually 3-7) word suggestions is presented on the screen. If the user selects one of these proposals, the text is automatically supplemented, which avoids the selection of the last letters of the word. Classically, word predictors are evaluated by an objective metric called *Keystroke Saving Rate* (*ksr*) :

$$ksr = (1 - \frac{k_p}{k_a}) \cdot 100$$

with k_p, , k_a being the number of keystrokes needed on the input device when typing a message with (k_p) and without prediction (k_a = number of characters in the corpus). We do not consider perplexity, being often used in the SLM domain, since it does not well reflect the actual gain in prediction.

3 AAC and the Problem of Language Model Adaptation

Experiments on a newspaper corpus have shown that SIBYLLE is able to arrive at a *ksr* varying between 50% and 60% (Schadle et al, 2004). However, the performance of this system decreases strongly with handicapped users, especially when the patient is agrammatic. This loss of performance is due to the differences in each situation of use. Since the users respond to very varied clinical patterns and will use AAC systems for varied purposes, we face multi-factorial requests for adaptation. Some works (e.g. Trost et al, 2005) already emphasized the importance of adaptation for AAC systems. However, these works did not consider this multiplicity of influences in terms of language registers.

3.1 Influence of Language Registers on Training

Our experiments have been conducted with a text prediction system based on a tri-gram model using backoff absolute discounting for smoothing. It was trained on a French newspaper corpus (*Le Monde*, 5,6M words); the vocabulary size amounted to 141.022 words. This model was assessed on several test corpora corresponding to various styles :

a. *Newspaper* (control situation) : part from *Le Monde*, not incl. in the training data; 20.009 words.
b. *Scientific*: a scientific article (unpublished) from the domain of NLP; 8.766 words.
c. *Literary*: first chapter from *Germinal* by Emile Zola; 20.928 words
d. *Speech*: transcription of spontaneous dialog between tourist agents and customers (*OTG* corpus; Antoine et al, 2002); 15.435 words.
e. *E-mail*: personal e-mails; headers, replies and hyperlinks were removed; 8.874 words.

Using a simulation device, we computed the *ksr* of the system on the five corpora. The ratio of out-of-vocabulary words (OOV) was determined as well (see table 1).
In the control situation (same register as training corpus), the prediction system showed a *ksr* of 50,5%. The percentage of OOV amounts to 3,4%. These results correspond to state-of-the-art performances in text prediction.
On the contrary, as can be seen in figure 2, a considerable degradation is observed for the other corpora : the *ksr* decreases by 8 to 16% (scientific register). OOV are obviously more frequent (up to 16,3%, see table 1). These results show a very important influence of the language register on training.

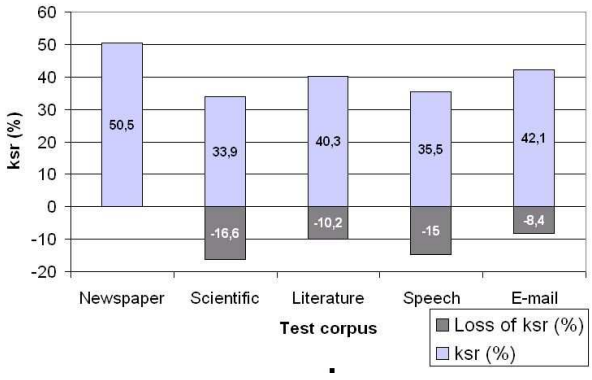

Figure 2: Results (*ksr*) for the five test corpora

4 Adapting a Language Model

As we have shown in the previous section, the performance of a language model is highly dependent on its actual usage. This problem has already been mentioned by others (e.g. Bellegarda, 2004), who also discuss several techniques to diminish this dependency on the training corpus. We investigate here three techniques of adaptation: a cache model, a user dictionary and a dynamically adapted user LM.

4.1 The Cache Model

Assuming that a word recently used has an elevated probability in the current discourse, the cache model keeps track of the *n* last inserted words (Kuhn & De Mori, 1990; Rosenfeld, 1996). The probability of these *n* words (*n* usually being between 50 and 200) is then augmented by a constant or an exponentially decaying factor *c* (see also Clarkson & Robinson, 1997). For our experiments we implemented a rather simple cache model: the maximum size *n* was set to 100 words; *c* = 0,0005. The mass of probability reserved for the cache did therefore not exceed 0,05. Stopwords were not added to the cache.

4.2 The User Dictionary (UD)

Like nearly all applications in NLP, a text prediction system has to deal with unknown words (out-of-vocabulary words, OOV). As an OOV is not predictible, its *ksr* remains at 0. The amount of OOV (i.e. their deteriorating effect) depends strongly on the similarity between the training and the test (or usage) data; however, as we could show, even when training and test data belong to the same register, it remains a non-negligible problem (OOV for the newspaper corpus : 4,83%).

We integrated to our prediction system a dynamically adapted user dictionary (UD) keeping track of every unknown word. As the frequency of these words is kept as well, we can properly calculate their probability just as if they had been part of the base vocabulary. The use of this dictionary reduces the percentage of OOV by up to 7%.

	Newsp.	Scient.	Lit.	Speech	E-mail
%OOV w/o UD	3,40	16,33	4,44	2,18	8,83
%OOV w UD	1,84	9,34	2,93	0,96	5,19

Table 1: % of OOV for the five test corpora

Considering the case of the UD, it becomes obvious why perplexity is a misleading measure of evaluation for our purposes: previously unknown words being included in the model will be assigned a very low probability. If they are encountered in the test corpus, model perplexity (reflecting average word probability) will necessarily rise. However, it is still beneficial to include unknown words, since they can further reduce the amount of typing events.

4.3 The Dynamic User Model (DUM)

This model integrates two LM's : a base trigram model and a user specific model being trained on every text being inserted. It follows the framework of Bellegarda (2004) with the difference that every text being inserted is instantly used within the language model. As unknown words are considered as well, this model comprises the former one (UD), and to a lesser extent as well a cache model, as it is sensitive to the current word usage. To arrive at a common probability estimate, the two models are linearly interpolated by the (well-known) formula:

$$P'(w_i) = \lambda_1 \cdot P_{Base}(w_i | w_{i-2} w_{i-1}) + \lambda_2 \cdot P_{DUM}(w_i | w_{i-2} w_{i-1})$$

where λ_1, λ_2 ($1 = \lambda_1 + \lambda_2$) are weighting factors. We estimated them also dynamically by applying an EM-like algorithm (Jelinek & Mercer, 1980) summing up the previous probabilities assigned by each of the two models. Empirical testing showed that this approach finds indeed an optimal weighting factor. Moreover, our factors were rather constant (0,40 - 0,49 for λ_1 and 0,60 - 0,51 for λ_2.)

5 Results

Table 2 shows the *ksr* measured for the three methods on every test corpus. The results displayed in the first line can be seen as a baseline for the three approaches discussed before.

	Newsp.	Scient.	Lit.	Speech	E-mail
Tri only	50,51	33,97	40,29	35,50	42,11
Tri + cache	51,07	35,14	40,76	39,03	42,98
	+0,56	**+1,17**	**+0,47**	**+3,53**	**+0,87**
Tri + UD	51,31	36,42	41,26	35,65	42,46
	+0,80	**+2,45**	**+0,97**	**+0,15**	**+0,35**
Tri + DUM	61,58	43,09	46,89	50,14	51,62
	+11,07	**+9,12**	**+6,60**	**+14,64**	**+9,51**

Table 2: Results (*ksr*) and advantages for all conditions

For the three adaptation techniques we can see beneficial effects on every corpus tested. For the cache model as well as the user dictionary the advantages are however not very high. There are in turn two interesting outliers: Firstly, for the cache model we measured for the speech corpus an advantage of 3,53% over the baseline. This indicates that oral communication relies much more on the current content of discourse; words are here more likely to re-occur.

Secondly, the UD scores 2,45% better than the baseline for the scientific corpus. This corpus had the highest rate of OOV (16,6%), which is not surprising, since the scientific register relies on a rather distinct vocabulary. In this case the UD was able to reduce the rate of OOV by 7%, meaning that nearly half of the unknown words occurred twice or more.

For the dynamic user model, however, different results can be observed. Here, we get advantages of 6,6% - 14,6% for all test corpora. Even for the test corpus of the same register (newspaper) we get an improvement of more than 11%. This was not expected, since the language style is rather homogeneous in newspaper text, and the amount of OOV words is not very high. This result further underlines the efficacy of the DUM approach.

6 Conclusion

We started from the observation that the performance of language models depends to a large extent on the similarity of the training data and the actual task or register. Our experimentations, conducted on test corpora from five different registers, showed a loss of up to 16% in *ksr*; a deterioration, which has to be expected in real-life conditions as well.

We then discussed three techniques to adapt an LM to its current usage: a cache model, a user dictionary and an interpolated dynamic user model. Whereas all three approaches proved to be beneficial, the advantages of the

cache and the user dictionary are rather limited (+0,1%-+3,5% gain of *ksr*). The dynamic user model however, showed very interesting gains (up to 14,6%), even for the test corpus belonging to the same register (newspaper) as the training data. From the superiority of the latter model we conclude that local syntactic information, as provided by the DUM, is of much more importance for our purposes than simple lexical knowledge. Knowing that a word has already occurred in the context (cache model) does not seem to have a big influence on prediction. But how about the semantic domain it belongs to?

7 Perspectives: Thematic Adaptation

A track we have not yet pursued in the context of adaptation is exploiting semantics or topical information.

There is no doubt that the probability of content words depends strongly on the particular thematic context (s. a. Lesher et al, 2002). For example, a rare word like 'abstention' or 'ballot' will have an elevated probability in the context of presidential elections.

For the exploitation of topic several approaches have been presented (e.g. Gildea & Hofmann, 1999). The trigger model (Rosenfeld, 1996; Matiasek et al. 2003), uses collocations to adapt word probabilities to a given context. We want to investigate a different approach: in the LSA model (Deerwester et al, 1990) a word w_i is represented as a high-dimensional vector, derived by *Singular Value Decomposition* (SVD) from a term × document (or a term × term) matrix of a training corpus.

In this framework, a context can be represented by the vector sum of the vectors corresponding to the words it contains (Landauer et al. 1997); these vectors can be compared by well-known similarity measures (scalar product, cosine). The vector of the actual history reflects the meaning of the preceding, already typed section, and can be compared with the term vectors of the vocabulary. The terms of the closest vectors should be semantically related to the history. We now can exploit this semantic similarity to make our model sensitive to the current topic by interpolating the similarity scores with the previously calculated probabilities. Promising approaches in this direction have been done by Coccaro & Jurafsky (1998).

As the words being contained in the history tend to be very close to the history vector, this approach also works as an improved cache model. We cannot present any results yet, but we are confident that this approach will enhance the *ksr* more than a simple cache. In the near future, this model will be subject to a thorough evaluation.

Acknowledgements

Part of this work was financed by the *German Academic Exchange Service*, DAAD.

References

ANTOINE, J.-Y., LETELLIER-ZARSHENAS, S., NICOLAS, P., SCHADLE I. (2002). Corpus OTG et ECOLE_MASSY : vers la constitution d'un collection de corpus francophones de dialogue oral diffusés librement. Actes *TALN'2002*. Nancy, France. Juin 2002

BELLEGARDA, J. (2004.: "Statistical language model adaptation: review and perspectives", *Speech Communication*, 42, pp. 93-108.

BIBER, D. (1993). Using Register-Diversified Corpora for General Language Studies, *Computational linguistics* 19(2), pp. 219-241.

CLARKSON, P. R. and ROBINSON, A.J. (1997). "Language Model Adaptation using Mixtures and an Exponentially Decaying Cache", in *Proc. IEEE ICASSP-97*, Munich.

COCCARO, N. and JURAFSKY, D. (1998). "Towards better integration of semantic predictors in statistical language modeling", *Proc. of the ICSLP-98*, Sydney.

DEERWESTER, S. C., DUMAIS, S., LANDAUER, T., FURNAS, G. and HARSHMAN, R. (1990). „Indexing by Latent Semantic Analysis", *JASIS* 41(6), pp. 391-407.

GILDEA, D, HOFMANN, T. (1999). "Topic-based Language Models using EM". *Proc. of Eurospeech-99*, Budapest.

JELINEK, F. and MERCER, R. (1980). "Interpolated estimation of Markov source parameters from sparse data". In *Pattern Recognition in Practice*, pp. 381- 397.

KUHN, R. and DE MORI, R. (1990). "A Cache-Based Natural Language Model for Speech Reproduction", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12 (6), pp. 570-583.

LANDAUER, T. K., LAHAM, D., REHDER, B. and SCHREINER, M. E. (1997). „How well can passage meaning be derived without using word order? A comparison of LSA and humans", *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pp. 412-417, Erlbaum Mawhwah, NJ.

LESHER, G. W., MOULTON, B. J, HIGGINBOTHAM, D.J. and ALSOFROM, B. (2002). "Limits of human word prediction performance", *Proceedings of the CSUN 2002*, California State University, Northridge.

MATIASEK, H. and BARONI, M. (2003). "Exploiting long distance collocational relations in predictive typing", *Proceedings of the EACL-03 Workshop on Language Modeling for Text Entry Methods*, Budapest.

ROSENFELD, R. (1996). "A maximum entropy approach to adaptive statistical language modelling", *Computer Speech and Langua*ge, 10 (1), pp. 187-228.

SCHADLE, I., ANTOINE, J.-Y., LE PÉVÉDIC, B. and POIRIER, F. (2004). "SibyMot: Modélisation stochastique du langage intégrant la notion de chunks", *Proceedings of the TALN-2004*, Fès.

TROST, H., MATIASEK, J. and BARONI, M. (2005). "The Language Component of the FASTY Text Prediction System", *Appl. Artificial Intelligence*, 19(8), 743-781.

WOODLAND, P.C., ODELL, J.J., HAIN, T., MOORE, G.L., NIELSER, T.R., TUERK, A. & WHITTAKER, E.W.D. (1998). "Improvements in Accuracy and Speed in the HTK Broadcast News Transcription System" In: *Proc. of the Eurospeech'98* , Budapest, Hungary.