Combining word- and class-based language models: A comparative study in several languages using automatic and manual word-clustering techniques

G. Maltese, P. Bravetti, H. Crépy, B. J. Grainger, M. Herzog, F. Palou

IBM Voice Systems, European Speech Research
gmaltese@tivoli.com

Abstract

This paper compares various class-based language models when used in conjunction with a word-based trigram language model by means of linear interpolation. For class-based language models where classes are automatically derived we present a comparative analysis in five languages (French, British English, German, Italian, and Spanish). With regard to classes corresponding to parts-of-speech, we present results for three languages (British English, French, and Italian). For each language, we present results for varying training corpus size and test script complexity. We achieved significant perplexity and word error rate reduction for all five languages and for several language models and recognition tasks. This work extends previous research by covering more languages and showing positive impact of these techniques with very large corpora, whereas prior work mostly focused on addressing data sparseness issues caused by small corpora.

1. Introduction

Data sparseness is a well-known problem when building language models (LM) for large-vocabulary speech recognition, specifically when estimating the a-priori probability $P(W)$ that the speaker will utter the sequence $W=w_1 \ldots w_n$. The well-known trigram LM provides an estimate of $P(W)$ ignoring past history beyond last two words, i.e.:

$$(1)\quad P(W) \equiv P_W(W) = \prod_{i=1}^{n} p(w_i \mid w_1 \ldots w_{i-1}) \cong \prod_{i=1}^{n} p(w_i \mid w_{i-2} w_{i-1})$$

One disadvantage of the trigram LM is its large number of parameters. If V is the vocabulary size, there may be $O(V^3)$ trigrams, a number significantly larger (even considering grammatical constraints) than the number of distinct trigrams which may be found in the available corpora with current values of V (typically $O(10^5)$). Another disadvantage of the trigram LM is its high dependence on the discourse domain as can be seen by measuring perplexity on a set of different texts belonging to (or outside of) the linguistic domain of the training corpus.

A partial solution to the data sparseness problem has been offered by smoothing techniques, which have led to the backing-off [1] and to the combined LM [2], the latter being that currently employed in the IBM ViaVoice large-vocabulary speech recognizers. In these models the probability of the current word w_i is estimated by taking into account the estimates supplied by trigram ($w_{i-2}\ w_{i-1}\ w_i$), bigram ($w_{i-1}\ w_i$), and unigram (w_i) distributions, considered one at a time (backing-off LM), or linearly combined.

However, particularly as the vocabulary size increases, it may well be the case that neither sequence $w_{i-2}\ w_{i-1}\ w_i$ nor $w_{i-1}\ w_i$ occur in the training corpus, or even the unigram w_i when the vocabulary is not derived by merely including the most frequent words from the corpus [3]. This is a significant drawback.

Another way to overcome the data sparseness problem, and, possibly, to reduce the dependence on the discourse domain, consists of grouping words together into equivalence classes and using the probabilities of word classes, instead of those of individual words. We therefore have a *triclass* LM:

$$(2)\quad P(W) \equiv P_C(W) \cong \prod_{i=1}^{n} k(w_i \mid c_i) h(c_i \mid c_{i-2} c_{i-1})$$

where we made the tacit assumption that each word w_i can belong to just one class. The probability h of class c_i given the context $c_{i-2}\ c_{i-1}$ is evaluated in the same way as for probability p in Eq. (1), i.e. according to the linearly interpolated LM scheme. The probability of word w_i given its class c_i can be conveniently estimated with the following formula:

$$(3)\quad k(w_i \mid c_i) = \frac{C(w_i) + 1}{S(c_i) + C(c_i)}$$

where $C(w_i)$ and $C(c_i)$ are the number of occurrences of word w_i and class c_i respectively, and $S(c_i)$ is the size of class c_i.

Class-based (hereafter C) LMs are more compact and generalize better with regard to unseen word sequences than purely word-based (hereafter W) LMs [4]. C LMs have proved effective for training on small datasets, or for fast language model adaptation. They can also be useful for speech recognition applications requiring small footprint LMs, as in the case of small devices supporting speech recognition applications.

However, for large-vocabulary dictation tasks, W LMs are still considered superior in capturing collocational relations between words, especially for texts belonging to the same discourse domain of the corpus used to train the LM. Thus, an effective way to take the best from both W and C LMs is to combine them. The combination scheme may rely upon backing-off [5] or linear interpolation [4], the latter using the formula:

$$(4)\quad P(W) = \gamma_W P_W(W) + \gamma_C P_C(W)$$

where $\gamma_W + \gamma_C = 1$.

With regard to word classification schemes, the approaches employed so far belong to two main paradigms. The first of these are manually based classification schemes employing POS

information [4], possibly improved through some automatic technique [6]. The other consists of automatically derived word classes, where words are clustered with respect to a statistical criterion (see, for example, [7]). In the following paper, we will refer to automatic and grammatical classing as AC and GC, respectively. Recent papers have compared the performance of both kinds of LMs [5, 8, 9].

Whilst there is evidence of the advantages of combining W and C LMs for large-vocabulary speech recognition tasks, we still feel that there are issues requiring clarification. For example, the merits of combining C and W LMs have not yet been fully explored in the area of large or very large text corpora or vocabularies through an extensive set of recognition results and perplexity figures. Moreover, most of the available results cover English, other languages having rarely been examined [9, 10].

In this paper, we report a comprehensive study in several languages concerning combinations according to Eq. (4), of W and C LMs. In each language, we employed training corpora of varying size, and test scripts of varying perplexity and recognition complexity. Five languages, namely British English (En), French (Fr), German (Gr), Italian (It), and Spanish (Sp), will be involved in the experiments with AC LMs. For three languages (En, Fr, It) we will also provide results with GC LMs. We will give both perplexity measures (PP) and word error rates (WER). For the latter, we employed the standard IBM ViaVoice large-vocabulary continuous speech recognizer.

Applying the same basic technology and the same techniques (to the extent which intrinsic differences between languages allow us to do) to different languages, and running a multi-lingual set of experiments provides us with a deeper insight into the effectiveness of combining W and C LMs in order to obtain robust and general purpose language models in a large-vocabulary speech recognition task.

2. Automatic determination of lexical classes by clustering

For four languages we employed an algorithm based on the simulated annealing principle [9, 11, 12] in order to automatically cluster words in each vocabulary. For British English we employed hierarchical tabu algorithm [13]. Both algorithms aim at minimizing the PP of a class bigram LM on the training corpus. In Table 1 we give the relevant parameters of the word-clustering step.

	Voc. size (kw)	Corpus size (Mw)	No. of classes	Initial PP	Final PP
En	147.4	187	2050	1221	301
Fr	120.2	65.3	3343	767	126
Gr	83.5	48.8	4624	463	148
It	100.1	218	946	1337	363
Sp	102.5	116.3	2048	721	159

Table 1. Relevant quantities for the word-clustering step in various languages.

The differences in the final PP values were probably due to differences in the number of classes for each language. Table 2 contains two examples per language of classes that we found interesting.

En	1) afraid, glad, grateful, hoping, oblivious, recommitted, thankful; 2) adapted, designed, opposed, suited
Fr	1) nombrer, prendre, provisionner, republier, typer; 2) position, possession, présence, quête, volute
Gr	1) dreimal, fünfmal, sechsmal, siebenmal, viermal, zehnmal, zweimal; 2) dicht, eng, freundschaftlich, großzügig, lose, nah, prominent, reich, unauflösbar
It	1) abbandona, benedice, confuta, esplora, esporta, percepisce; 2) cedevole, colmabile, controproducente, dannoso, entusiasmante, negligente, pretestuoso
Sp	1) Fernando, Imanol, Inocencio, Isidoro, Rodolfo, Valeriano, Victorino; 2) empresarial, estudiantil, gremial, gubernamental, sindical, vecinal

Table 2. Examples of automatic word groupings.

3. Manually derived grammatical classes

For three languages (Fr, It, En) we built a word-class scheme based on part-of-speech (POS) information. This would typically include information about the word's main syntactic category, such as noun, verb, adjective and, depending on the granularity of the class scheme, additional information such as number, tense, gender, etc. The first step towards this goal might be to have just one word class per each POS tag. However, in each language, many words can be assigned different POS tags. This means that one either has to cope with overlapping word classes, or that we have to be able to guess for each occurrence of a word what POS tag to assign to it given its context.

To circumvent intrinsic language ambiguity with respect to POS tags, we adopted an approach allowing ambiguity of word classing [3, 4]. This meant that we grouped together words that could be assigned to the same set of possible POS tags. Instead of trying to disambiguate, we have chosen to acknowledge the intrinsic ambiguity of classes by allocating classes for words that belong to multiple pure POS classes.

For French, starting with a POS tag inventory of 287 purely grammatical classes, this scheme produces 870 classes. Of those, 389 are actually singletons, including many frequent function words. Furthermore, we added 209 expert-made categories, such as "days of the week" or "male first names".

For Italian, we started with a POS scheme of 261 purely grammatical classes, and enlarged it to 338 classes including 77 ambiguous classes. We put in singleton classes all frequent function words and frequent phrases having a well-defined grammatical role, such as adverbial and prepositional phrases. As in Fr, we included some expert-made classes, like family names or measurement units; we ended up with a 947-class scheme.

For British English, we started from a 20 POS set and enlarged it to 211 classes including ambiguity and expert-made classes. We also included 251 singleton classes for frequent function words, giving a total of 462 classes.

4. Experimental results

4.1 Test scripts and training corpora

In Table 3 we list the vocabulary sizes used in the experiments for the various LMs.

	W	AC	GC
En	150.6	147.5	150.6
Fr	236.3	120.2	236.3
Gr	94.5	83.5	-
It	100.1	100.1	100.1
Sp	102.8	102.6	-

Table 3. Vocabulary sizes (kw) for W and C LMs.

For each language, we trained two W and two AC LMs based on two corpora C1 and C2, where C2 is a large corpus, superset of the smaller C1. We measured PP and WER on two test scripts, one of medium-low PP (T1) and the other of medium-high PP (T2). In Table 4 we list the size of training corpora and test scripts, together with the number of speakers reading each test script.

	C1 (Mw)	C2 (Mw)	T1 w*spks	T2 w*spks
En	206.8	1736	1825*15	1464*15
Fr	65.3	326.7	1057*14	1271*6
Gr	98	466.8	288*20	448*20
It	96.4	782.5	5952*4	350*5
Sp	31.3	116.3	684*20	265*20

Table 4. Size of training corpora and test scripts and number of speakers reading each test script.

Journalism and law form a major part of training corpora. For Fr, It and Gr we also included some reports and office correspondence (OC). For En, corpus C2 includes the British National Corpus. Test scripts are everyday sentences for En; OC (T1) and legal texts (T2) for German; OC+journalism+a thesis abstract (T1) and an historical travel report (T2) for Fr. For It, we have OC+journalism+homework+school reports in T1 script, and recipes in T2. Finally, for Sp, we have journalism for both training corpora and test scripts. There are no out-of-vocabulary words in any of the test scripts.

4.2 Combining word-based and class-based language models

In Table 5 we list PP and WER for each LM, measured both on T1 and T2 scripts. We can see that, for the T1 scripts, in the majority of cases, PP increases from W to AC and GC LMs. This is not always true for WER, as sometimes AC LMs outperform W ones (En_C1, Fr_C1, Gr_C1/C2, Sp_C1). This never happens for GC LMs. For T2 scripts in several cases AC LMs show lower WER values (En_C1/C2, It_C1/C2, Sp_C1/C2) and lower PP values (En_C1, Fr_C1, It_C1/C2, Sp_C1/C2) than W ones. This also happens for It GC LMs, for both PP and WER values.

	PP/**WER**(T1)			PP/**WER**(T2)		
LM	W	AC	GC	W	AC	GC
En_C1	353 **9.8**	391 **9.14**	825 **15.0**	4358 **15.0**	3752 **13.9**	6500 **18.1**
En_C2	268 **7.87**	363 **8.43**	868 **14.7**	3162 **13.8**	3367 **12.9**	6949 **17.7**
Fr_C1	217 **11.0**	225 **10.5**	373 **11.7**	395 **19.9**	380 **20.1**	553 **22.7**
Fr_C2	174 **8.65**	205 **9.16**	370 **11.5**	308 **18.6**	346 **18.9**	539 **22.1**
Gr_C1	301 **7.06**	300 **6.13**	-	407 **11.7**	493 **12.7**	-
Gr_C2	255 **6.13**	282 **5.48**	-	360 **10.7**	460 **12.9**	-
It_C1	579 **6.56**	741 **6.61**	1025 **8.06**	13118 **24.6**	7596 **21.6**	7299 **21.4**
It_C2	457 **5.48**	756 **6.23**	1093 **7.93**	8844 **21.4**	7386 **20.1**	7958 **20.7**
Sp_C1	159 **5.72**	158 **5.41**	-	808 **9.32**	660 **8.02**	-
Sp_C2	137 **5.0**	150 **5.12**	-	647 **8.11**	603 **7.02**	-

Table 5. PP and **WER** (%) values for each LM measured on T1 and T2 scripts.

Generally speaking, C LMs are better suited than W ones to high-PP tasks. As expected, for all languages WER and PP values decreased when increasing training corpus from C1 to C2 for W LMs. As far as WER is concerned, this also happened for C LMs (the only exceptions being the Gr AC LM on the T2 script), whilst PP of C LMs shows a less pronounced effect (PP always decreases when enlarging training corpus for AC LMs on script T2).

Figure 1 shows relative WER improvement combining W and AC LMs compared to W LMs vs. weight of AC LM (γ_C in Eq. 4). We find a substantial decrease in WER for a wide range of values of γ_C.

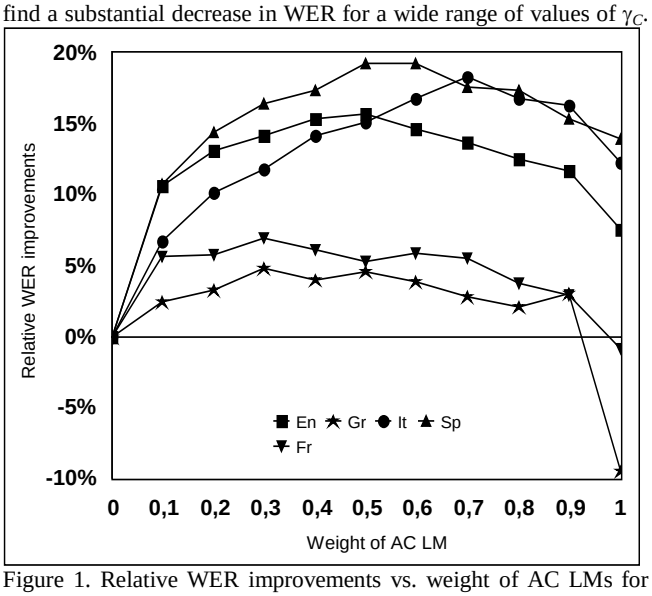

Figure 1. Relative WER improvements vs. weight of AC LMs for the C1/T2 combination.

In Table 6 we list the highest percentage decrease in WER and PP

for each LM combination compared to standalone W LM.

	Percentage decrease in PP/**WER** for T1 script with respect to W LM		Percentage decrease in PP/**WER** for T2 script with respect to W LM	
LM	W+AC	W+GC	W+AC	W+GC
En_C1	13.9 **13.1**	4.0 **4.0**	32.5 **15.7**	21.4 **6.2**
En_C2	4.48 **5.34**	0.75 **0.45**	23.3 **14.7**	13.9 **5.0**
Fr_C1	10.6 **14.9**	9.7 **12.1**	17.0 **6.97**	12.7 **6.92**
Fr_C2	6.9 **5.9**	4.6 **2.2**	9.7 **6.47**	7.8 **8.46**
Gr_C1	11.6 **19.1**	-	11.1 **4.81**	-
Gr_C2	6.7 **11.6**	-	9.4 **4.09**	-
It_C1	17.6 **15.7**	13.1 **9.45**	49.6 **18.3**	48.3 **16.0**
It_C2	11.6 **13.9**	8.53 **8.58**	39.7 **17.3**	39.4 **14.1**
Sp_C1	9.43 **11.9**	-	25.7 **19.2**	-
Sp_C2	5.84 **7.8**	-	15.0 **18.4**	-

Table 6. Performance when interpolating W and C LMs with respect to standalone W LMs.

We observe that we consistently achieve a reasonable improvement in both PP and WER when considering W+AC with respect to W LMs. The improvements are somewhat smaller when increasing the corpus from C1 to C2. This is due to the fact that W LMs benefit more than AC LMs from an increase in training corpus size (see also the discussion on WER and PP figures of data shown in Table 5). Yet W+AC LMs significantly outperform W LMs even when large or very large training corpora are taken into account. PP shows the same trend as WER.

W+GC LMs show the same behaviour with respect to W LMs as W+AC LMs. Relative improvements are in most cases somewhat smaller with respect to W+AC LMs, the only significant exception being French C2/T2 combination. In the latter figure the increased complexity of the task and the high degree of inflection of the French language probably play a significant role.

In several cases, when passing from T1 to T2, we observe an increase in effectiveness of combined models. For W+AC LMs, this happens for En_C1/C2; Fr_C2; It_C1/C2, and Sp_C1/C2. For W+GC LMs this happens for En_C1/C2, Fr_C2, and It_C1/C2 LMs. This may support the validity of employing C LMs when dealing with more complex recognition tasks, or tasks unrelated to the training corpora.

For Fr and It, further improvements were achieved by simultaneously combining W, AC, and GC LMs. For Fr, in the C1/T1 and the C2/T1 tasks, we got a WER reduction compared to standalone W LM of 17.1% and 8.63%, respectively. For It, WER reduction went up to 17.2%, 20.4%, 14.2%, 18.4% in tasks C1/T1, C1/T2, C2/T1 & C2/T2, respectively. Comparing these values with corresponding ones listed in Table 6, we see that by using three LMs for the aforementioned tasks results in additional reduction in WER ranging from 0.3 to 2.2%.

5. Conclusions and future work

For several languages we have demonstrated that combining class- and word-based LMs led to significant reductions both in PP and WER and even for large or very large training corpora. This result was more evident for complex than for simple tasks, thereby confirming the effectiveness of class-based LMs in making word-based LMs less dependent on the domain of the training corpus when combined. Class-based LMs can be used in a variety of circumstances as an effective way of integrating the predictive power of word-based LMs. The interest of these results is that they demonstrate the effectiveness of class-based LMs in many languages, and extend the proof of their worthiness into the realm of very-large corpora LMs, a domain that previous investigations left relatively untouched. Future work will be devoted to improving existing grammatical class schemes, creating them in Gr, Sp, and testing the approach outlined in this paper with more scripts/corpora. Dynamic weighting of combined LMs will also receive due attention.

Acknowledgements. We thank our colleagues in the IBM European Speech Research (Cairo, Heidelberg, Hursley, Paris, Rome, and Seville) and in the IBM Human Language Technologies Group (T. J. Watson Research Center, Yorktown Heights) for many valuable suggestions and the continuous exchange of ideas. Special thanks to Burn Lewis (ideas, implementations), Jean-Christophe Marcadet (grammatical classing), and Marc Richarme (automatic classing).

6. References

[1] Katz, S. “Estimation of probabilities from sparse data for the language model component of a speech recognizer”; IEEE Trans. ASSP, vol. 35, no. 3, March 1987, pp. 400-401.

[2] Jelinek, F.; Mercer, R.L. “Interpolated estimation of Markov source parameters from sparse data”; Proc. of the Workshop on Pattern Recognition in Practice, North-Holland, Amsterdam, The Netherlands, pp. 381-397, May 1980.

[3] Crépy, H.; Marcadet, J.-C.; Waast, C.; “Dictée à grand Vocabulaire en Français: IBM Voicetype 3.0, un Produit de la Recherche”, 1° JST, Francil, 1997, pp. 19-23.

[4] Samuelsson, C.; Reichl, W.; “A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics”, Proc. of ICASSP 1999, paper no. 1781.

[5] Niesler, T.R.; Woodland, P.C.; “Combination of word-based and category-based language models”, Proc. of ICLSP 1996, pp. 220-223.

[6] Witschel, P.; “Optimized pos-based language models for large vocabulary speech recognition”, Proc. of ICLSP 1998, paper no. 471.

[7] Brown, P.F.; DeSouza, P.V; Mercer, R.L.; Della Pietra, V.J; Lai, J.C; “Class-Based n-gram Models of Natural Language”, Computational Linguistics, vol. 18, pp. 467-480, 1992.

[8] Niesler, T.R.; Whittaker, E.W.D.; Woodland, P.C.; “Comparison of part-of-speech and automatically derived category-based language models for speech recognition”, Proc. of ICASSP 1998, vol. 1, paper no. 2003, pp. 177-180.

[9] Smaïli, K.; Brun, A. Zitouni, I.; Haton, J.P.; “Automatic and manual clustering for large vocabulary speech recognition: a comparative study”, Proc. of EUROSPEECH 1999, paper no. S044.

[10] Whittaker, E.W.D.; Woodland, P.C.; “Comparison of Language Modelling Techniques for Russian and English”, Proc. of ICSLP 1998, paper no. 968.

[11] Jardino, M.; Adda, G.; “Automatic word classification using simulated annealing”, Proc. of ICASSP 1993, vol. II, pp. 41-44.

[12] Kirkpatrick, C.D.; Gelatt, Jr., C.D.; Vecchi, M.P.; “Optimization by Simulated Annealing”, Science, vol. 220, pp. 671-680, 1983.

[13] Glover, F.; Laguna, M.; “Tabu Search”, Kluwer, Boston, 1997.