Statistical and Linguistic Clustering for Language Modeling in ASR*

R. Justo and I. Torres

Departamento de Electricidad y Electrónica,
Facultad de Ciencia y Tecnología,
Universidad del País Vasco
webjublr@lg.ehu.es, manes@we.lc.ehu.es

Abstract. In this work several sets of categories obtained by a statistical clustering algorithm, as well as a linguistic set, were used to design category-based language models. The language models proposed were evaluated, as usual, in terms of perplexity of the text corpus. Then they were integrated into an ASR system and also evaluated in terms of system performance. It can be seen that category-based language models can perform better, also in terms of WER, when categories are obtained through statistical models instead of using linguistic techniques. They also show that better system performance are obtained when the language model interpolates category based and word based models.

1 Introduction

Automatic Speech Recognition and Understanding (ASRU) Systems are currently based on Statistical Language Modeling. Thus, large amounts of training data are required to get a robust estimation of the parameters of the model. However, the availability of large amounts of training material is not always assured when designing many of usual ASRU applications. As an example, the text corpus needed to train a dialogue system consists of transcriptions of dialogue turns uttered by potential users of the system to be developed. These speakers reproduce the natural behavior of further users including spontaneous, unconstrained and most times noisy speech. This procedure only allows to obtain a limited corpus to train the language model of the ASRU system, smaller than the usual text databases.

One of the ways to deal with sparseness of data is to cluster the vocabulary of the application tasks into a reduced number of categories. Replacing words by the category they belong to entails significant reductions in the number of parameters to be estimated. Thus, smaller training corpora can be used. On the other hand, new words belonging to previously defined categories can be directly added to the vocabulary of the task without changing the training corpus.

* This work has been partially supported by the CICYT proyect TIC2002-04103-C03-02 and by the Universidad del País Vasco under grant 9/UPV 00224.310-13566/2001.

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 556–565, 2005.

The first issue when generating a category-based language model is the appropriate selection of word classes. Morphosyntactic and/or semantic knowledge is usually applied to manual clustering of linguistic categories (e.g. part of speech POS) [1]. This procedure leads to some perplexity reduction when applied to limited domain tasks. However in less constrained domains these models do not usually improve on word-based language models. Alternatively, iterative clustering algorithms using theoretic information criteria have also been proposed to reduce perplexity in large corpora [2].

In this work the categories obtained through a statistical clustering algorithm are compared with a classical linguistic set of POS. Several category-based language models are evaluated and compared in terms of ASR system performance. Thus, not only the perplexity of a text test set is evaluated but also the WER obtained through the ASR system when different category-based language models are compared. On the other hand, a category-based language model will prove coarser than a word-based model and could lose accuracy in predictions of the next word. In such a case, the language model is only based on relations between word classes and on the probability distribution of words into classes. Alternatively a second approach proposes a language model that interpolates the information associated with the relations between categories and the information associated to the relations between words [3].

These proposals were evaluated through a set of recognition experiments carried out on a Spanish human-machine dialogue system. The experiments carried out shows that category-based language models can also perform better in terms of WER, when categories are obtained through statistical models than for linguistic categories, even for limited domains. They also show that better system performance is obtained when the language model interpolates category-based and word-based models.

Section 2 deals with the method for classifying words into categories. The statistical clustering algorithm is fully explained and the POS categories are presented. In Section 3 the two category-based language model are described and their integration into the decoder is presented. Section 4 deals with experimental evaluation of the proposals and Section 5 presents the main conclusions and suggestions for future work.

2 Classification of Words into Categories

A classical clustering algorithm has been used in this work to automatically obtain a set of categories from a text corpus. The goal of a clustering algorithm is to group samples with high internal similarity. For this purpose, an objective function to be optimized should be defined [4]. This function will also measures the quality of any partition of the data. Thus, the clustering algorithm has to find the partition of the initial set of samples that optimizes the objective function. Section 2.1 fully explains the objective function to be maximized in the clustering algorithm presented in Section 2.2. Finally Section 2.3 presents the linguistic set of categories used for comparison purposes.

2.1 Objective Function: Class Bigram Models

We first describe a class bigram model which is the basis of the objective function to be selected [5]. Suppose a distribution of the W words of the vocabulary into N_C classes using a function $C(\cdot)$, which maps a word w into a class C_w, $C(\cdot) : w \rightarrow C_w$. If each word is assigned to a single class, a word bigram (w_i, w_j) will correspond to the class bigram (C_{w_i}, C_{w_j}).

According to a class bigram model:

$$p(w_j|w_i) = p(w_j|C_{w_j})p(C_{w_j}|C_{w_i}) \tag{1}$$

Given a training corpus and the map function C, $p(w_j|C_{w_j})$ and $p(C_{w_j}|C_{w_i})$ can be estimated taking into account the number of times that particular events have been seen in the training corpus, $N(\ )$.

$$p(w|C_w) = \frac{N(w)}{N(C_w)} \tag{2}$$

$$p(C_{w_j}|C_{w_i}) = \frac{N(C_{w_i}, C_{w_j})}{N(C_{w_i})} \tag{3}$$

The clustering algorithm consist of finding the function C that maximizes the log-likelihood function of the class bigram model described in 1, on the training corpus.

The likelihood is defined as the joint probability of the training samples and using the bigram model is expressed as follows:

$$P(w_1 \dots w_N) = P(w_1) \prod_{n=2}^{N} P(w_n|w_{n-1}) \tag{4}$$

From equation 4, the function log-likelihood is developed for a bigram class model:

$$F_{bi}(\mathcal{C}) = \sum_{n=1}^{T} \log P(\omega_n|\omega_{n-1}) = \sum_{v,w} N(v,w) \log p(w|v) =$$

$$= \sum_{w} N(w) \log \tfrac{N(w)}{N(C_w)} + \sum_{C_v, Cw} N(C_v, C_w) \log \tfrac{N(C_v, C_w)}{N(C_v)} =$$

$$= \sum_{C_v, Cw} N(C_v, C_w) \log N(C_v, C_w) - 2 \sum_{C} N(C) \log N(C) + \sum_{w} N(w) \log N(w) \tag{5}$$

where each term is defined in 1.

2.2 Clustering Algorithm

The goal of this algorithm is to find the function C , thus, the way the words can be grouped, which maximizes the log-likelihood function of the bigram class

Table 1. Notation for the expression 5

$F_{bi}(\mathcal{C})$	Log-likelihood function for a bigram class model.
T	Training corpus size.
C	Word class.
C_v, C_w	Classes containing the words v and w respectively.
$N(C)$	Number of occurrences of the C class in the training corpus.
$N(C_v, C_w)$	Number of occurrences of the C_v class after C_w class have been seen in the training corpus.
$N(w)$	Number of times w word has appeared in the training.

model, $F_{bi}(\mathcal{C})$ on the training corpus. An iterative clustering algorithm based on sample exchange is used for this purpose [5].

Iterative algorithm.
Start with some initial mapping: N_C classes and $\mathcal{C} : w \rightarrow C_w$.
do
for each w of the vocabulary **do**
for each class k **do**
- tentatively exchange word w from class C_w to class k.
- compute likelihood for this tentative exchange.

exchange word w from class C_w to class k which maximizes the likelihood.
until the stopping criterion is met.

The result of such algorithms strongly depends on the initialization, thus different classes are generated depending on the initial distribution of words into categories.

The initial distribution in [5] is based on placing each of he most frequent words in a single class and the rest in the last class. This technique lets the most frequent words determine the way the words are grouped because they are evaluated first in the process. However, the algorithm used does not permit the use of this technique with the same result because putting each of the most frequent words in a single class means they are unable to leave that class until they are not only word in it. Therefore, a different distribution has been selected [6], consisting of placing each of the most frequent words each in a class, except for the less frequent $N_C - 1$, which are placed in the $N_C - 1$ remaining classes.

2.3 Linguistic Categories

The categories obtained from a text corpus, using classical clustering algorithms, have been compared to the linguistic categories obtained from a morphosyntactic analyzer: "Freeling", in terms of ASR system performance.

"Freeling" is a free software developed in Barcelona's Polytechnic University by the *talp* group. The FreeLing package consists of a library providing language analysis services (such as morphosyntactic analysis, date recognition, PoS tagging, etc.) In this case, only morphosyntactic analysis is used to obtain classes

comparable with those obtained automatically. The classes given by the Freeling correspond to the following "eagle" labels: Adjectives, adverbs, determinants, names, verbs, pronouns, conjunctions, interjections, prepositions and another class was defined for the word P, corresponding to silences.

3 Category-Based Language Models into a Speech Recognition System

In this section two category-based language models are defined and then integrated into a speech recognition system.

3.1 A Language Model Based on Category N-Grams

In a first approach, the language model only collects the relations between word groups, "forgetting" the relations between particular words [7].

The probability of a sentence $(w_1, \ldots, w_N)$ can be represented as a product of conditional probabilities:

$$P(w_1, \ldots, w_N) = P(w_1) \prod_{n=2}^{N} P(w_n | w_1 \ldots w_{n-1}) \tag{6}$$

where $P(w_n|w_1 \ldots w_{n-1})$ represents the probability of w_n when the sequence of words $(w_1, \ldots, w_{n-1})$ has been observed.

Assuming that when the categorization process is finished, the set of words in the lexicon belongs to a smaller group of "a priori" defined categories, the probability of w_N conditioned to its $N-1$ predecessors can be defined as follows [7] [8]:

$$P(w_N|w_1 \ldots w_{N-1}) = \sum_{j=1}^{N_c} P(w_N|C_j) P(C_j|C_1 \ldots C_{j-1}) \tag{7}$$

where N_C is the number of different word categories.

The classification algorithm restricts the membership of words to a single class, so a single label corresponding to a category is assigned to a word and the above equation assumes the follows form:

$$P(w_N|w_1 \ldots w_{N-1}) = P(w_N|C_{w_N}) P(C_{w_N}|C_{w_1} \ldots C_{w_{N-1}}) \tag{8}$$

The parameters of the distributions of words into categories are calculated as follows:

$$P(w|C) = \frac{N(w, C)}{\sum_{w'} N(w', C)} \tag{9}$$

where $N(w, C)$ is the number of times a word w is labeled by C in the training corpus.

On the other hand, $P(C_{w_N}|C_{w_1} \dots C_{w_{N-1}})$ represents the probability of C_{w_N} being the next class if up to now $C_{w_1} \dots C_{w_{N-1}}$ category sequence has been observed and C_{w_i} represents the class w_i belongs to.

It is important to notice that probabilities are calculated using category n-gram based models, analogous to word n-grams, so the history of an event is reduced to the n-1 previous events, thus:

$$P(w_N|w_1, \dots, w_{N-1}) \cong P(w_N|w_{N-n+1}, \dots, w_{N-1}) \tag{10}$$

and expression 8 is rewritten as:

$$P(w_N|w_1 \dots w_{N-1}) = P(w_N|C_{w_N})P(C_{w_N}|C_{w_{N-n+1}} \dots C_{w_{N-1}}) \tag{11}$$

An automatic speech recognition system based on the Viterbi algorithm looks for the sequence of states that has the maximum probability given the sequence of acoustic observations, and thus estimates the sequence of words the speaker pronounced

The transition probability between each pair of words is calculated in accordance with expression 11. This model only considers the probability distribution of words into categories and the category n-gram model.

3.2 Interpolating Category and Word N-Gram Models

The category based language model described in the equation 11 does not need so many parameters as the one based on word n-grams. Thus, it may be better estimated, with a higher confidence level. But it fails to capture the relationships between particular words so it is less accurate in predicting the next word.

The hybrid model to be described try to integrate both information sources, i.e. the one relative to relationships between particular words and the one associated with the relationships between groups of words.

This hybrid model is an interpolation of a model based on category n-grams and a model based on word n-grams. It is defined as a linear combination of both models. The probability of the word w_N conditioned to the N-1 previous words, would be represented as follows [3]:

$$\begin{aligned} &P(w_N|w_1 \dots w_{N-1}) = \lambda P(w_N|w_1 \dots w_{N-1}) + \\ &+(1-\lambda)P(w_N|w_1 \dots w_{N-1}, M_c) \end{aligned} \tag{12}$$

If n-grams based models are used as in the previous sections:

$$\begin{aligned} &P(w_N|w_1 \dots w_{N-1}) = \lambda P(w_N|w_{N-n+1} \dots w_{N-1}) + \\ &+(1-\lambda) \sum_{j=1}^{C} P(w_N|C_j)P(C_j|C_{j-n+1} \dots C_{j-1}) \end{aligned} \tag{13}$$

and assuming that each word belongs to a single class

$$P(w_N|w_1 \dots w_{N-1}) = \lambda P(w_N|w_{N-n+1} \dots w_{N-1}) + \\ +(1-\lambda)P(w_N|C_{w_N})P(C_{w_N}|C_{w_{N-n+1}} \dots C_{w_{N-1}}) \quad (14)$$

In this case the speech recognizer calculates the transition probability between each pair of words taking into account three probability distributions: distribution of words into categories, category n-grams and word n-grams.

4 Experimental Results

Several speech recognition experiments were done using a human-machine dialogue corpus in Spanish, BASURDE [9]. The speakers ask for information about long distances trains schedules, destinations and prices by telephone (8KHz). It was acquired by the "Wizard of Oz" technique and has 227 dialogues uttered by 75 speakers, 43 male and 32 female. The total number of phrases is 1657 and 1340 of them are used for training and the remaining 308 for testing. The starter set of the vocabulary consists of 788 words and the total number of words is 21088.

The language models proposed in this work were evaluated, as usual, in terms of perplexity (PP) of the text corpus. Then they were integrated into an ASR system and evaluated in terms of both, Word Error Rate (WER) and Category Error Rate (CER).

Continuous HMM were used to model a set of context independent units corresponding to the basic set of Spanish phones. These models were previously trained over a task-independent phonetically balanced Spanish corpus (SENGLAR) uttered by 57 speakers and then integrated into the ASR system.

K-testable grammars in the strict sense (K-TSS) were used to get the proposed language models. This formalism is considered as the grammatical approach to the well known n-gram models [10]. The ASR consists finally in a single stochastic automaton integrating acoustic, lexical, word-based and category-based language models along with the required smoothing technique. The search of most probable hypothesis over the full network is based on the Viterbi algorithm.

Category based language models based on 5, 10 and 20 categories, obtained through the clustering algorithm, as well as the language model based on the 10 linguistic classes obtained by "Freeling" were evaluated in these experiments. For comparison purposes a classical word-based language model was also considered. In such a case, the number of classes is equal to the size of the vocabulary, i.e. 788. Table 2 shows the perplexity evaluation for k=2, 3, and 4 models. This table reveals, as expected, important reductions of the PP values for the category-based language models. These PP values increased with the number of categories but similar value was achieved by linguistic and statistical methodologies for equal number of classes.

A first evaluation of the defined categories was achieved using the word sequences obtained through the ASR system. A conventional word-based language model was integrated into the ASR system. Then, once the recognition process

Table 2. Perplexity values for a classical word-based language model (788 classes) and for language models generated using a labeled corpus with 5, 10 and 20 automatically obtained classes on one hand and with 10 linguistic classes on the other hand

PP	without categories	statistical clusterings			linguistic classes
	788 words	5	10	20	10 (9+1)
k=2	29.8	3.06	4.73	7.38	5.15
k=3	27.36	3.02	4.69	7.35	4.64
k=4	27.65	3.03	4.70	7.83	4.36

Table 3. Values of CER when the word based language model is used but the recognized phrases are labeled with the labels corresponding to 5, 10 and 20 automatically obtained categories on one hand and to 10 linguistic categories on the other hand. The value of ($CER \equiv WER$) when the word based language model is used and no categories are considered, i.e. 788 categories, also appears.

	without categories	statistical clustering			linguistic classes
	788 words ($WER \equiv CER$)	5	10	20	10 (9+1)
CER(%)	38.19	27.97	31.43	33.87	39.96

was finished, both the sequences of words obtained by the recognizer and the reference sequences of words were labeled according to the different set of categories defined. Thus, a Category Error Rate (CER) can be calculated. Table 3 shows that CER is clearly lower for statistical categories than for linguistic ones. In this case, CER is similar, even greater, to CER obtained when any clustering was considered. Thus, the confusions, i.e. substitution errors, between words belonging to different cluster seems to be lower for statistical categories than for linguistic ones. Let us note that in certain sense linguistic categories also model the order of the phrase in agreement with the general syntax of the language. However this fact does not seem important in this case, perhaps due to the natural and spontaneous type of speech in the corpus.

For the final evaluation the category based models described in the section above were integrated into the ASR system.

Table 4 shows the WER and CER obtained when the category based language model, defined in section 3.1 was used. Statistical and linguistic categories were compared using corresponding K-TSS language models. Reductions in CER and WER can be seen in the mentioned table when statistical categories were used. However, the integration of the category based language model does not improve the system performance measured in terms of both WER and CER (see table 3 to compare).

Finally the hybrid language model interpolating a category based model and a word based model (see section 3.2) was integrated into the ASR system. Table 5 shows WER and CER obtained when statistical and linguistic categories were considered in the hybrid model. In this table an important reduction in CER and

WER can be seen when compared to Table 4. The interpolation of word-based models and category-based models improves the WER and CER obtained by a simple category category-based language model. Nevertheless, the final performance of the ASR system is similar, maybe a little better, than the reference ASR system which did not consider any category model.

Finally let us note that the objective function used in the statistical clustering algorithm seems to work quite well since the values of CER are quite low for these categories.

Table 4. Values of WER and CER using a category based language model with 5, 10 and 20 statistical clusters on the one hand and 10 linguistic classes on the other one

	statistical clustering			**linguistic classes**
number of classes	**5**	**10**	**20**	**10 (9+1)**
WER (%)	51.06	47.12	46.57	52.42
CER (%)	33.07	36.20	39.63	41.33

Table 5. Values of WER and CER using a hybrid language model where the category based language model has been generated with 5, 10 and 20 statisitcal clusters on the one hand and with 10 linguistic classes on the other one

	statistical clustering			**linguistic classes**
number of classes	**5**	**10**	**20**	**10 (9+1)**
WER (%)	38.34	37.92	38.16	37.67
CER (%)	27.39	30.2	33.02	40.35

5 Conclusions and Future Work

In this work several sets of categories obtained by a statistical clustering algorithm, as well as a linguistic set, were used to design category-based language models. The language models proposed were evaluated, as usual, in terms of perplexity of the text corpus. Then they were integrated into an ASR system and also evaluated in terms of system performance.

The experiments carried out shows that category-based language models can perform better, also in terms of WER, when categories are obtained through statistical models instead of using linguistic techniques, even for limited domains. They also show that better system performance are obtained when the language model interpolates category based and word based models.

These preliminary experiments have shown the power of statistical clustering of words for language modeling, even for limited domain application tasks. However, an in-depth experimentation is required to explore new objective functions and initializations in cluster algorithm. Alternative formalism s to interpolate and integrate models into the ASR system should also be explored.

References

1. Niesler, T.: Category-based statistical language models. PhD thesis, Department of Engineering, University of Cambridge, U.K. (1997)
2. Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Comput. Linguist. **18** (1992) 467–479
3. Linares, D., Benedí, J., Sánchez, J.: A hybrid language model based on a combination of n-grams and stochastic context-free grammars. ACM Trans. on Asian Language Information Processing **3** (2004) 113–127
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2nd edn. Wiley-Interscience (2000)
5. Martin, S., Liermann, J., Ney, H.: Algorithms for bigram and trigram word clustering. Speech Communication **24** (1998) 19–37
6. Barrachina, S.: Técnicas de agrupamiento bilingue aplicada a la inferencia de traductores. PhD thesis, Universidad Jaume I, Departamento de Ingeniería y Ciencia de los Computadores. (2003)
7. Niesler, T.R., Woodland, P.C.: A variable-length category-based n-gram language model. In: IEEE ICASSP-96. Volume I., Atlanta, GA, IEEE (1996) 164–167
8. Nevado, F., Sánchez, J., Benedí, J.: Lexical decoding based on the combination of category-based stochastic models and word-category distribution models. In: IX Spanish Symposium on Pattern Recognition and Image Analysis. Volume 1., Castellón (Spain), Publicacions de la Universitat Jaume I (2001) 183–188
9. Proyecto BASURDE: Spontaneus-Speech Dialogue System in Limited Domains. Comisin Interministerial de Ciencia y Tecnologa TIC98-423-C06 (1998-2001) http://gps-tsc.upc.es/ veu/basurde/Home.htm.
10. Torres, I., Varona, A.: k-TSS language models in speech recognition systems. Computer Speech and Language **15** (2001) 127–149