Category-based Language Models in a Spanish Spoken Dialogue System*

Raquel Justo e Inés Torres
Departamento de
Electricidad y Electrónica
Universidad del País Vasco
E-48080 Leioa, Spain
webjublr@lg.ehu.es, manes@we.lc.ehu.es

Jose Miguel Benedí
Departamento de Sistemas
Informáticos y Computación
Universidad Politécnica de Valencia
E-46071 Valencia, Spain
jbenedi@dsic.upv.es

Resumen: El objetivo principal de este trabajo es comprobar si un modelo de lenguaje basado en categorías puede mejorar el rendimiento de un sistema de diálogo, de la misma forma que lo hace para aplicaciones que utilizan bases de datos no espontáneas y de mayores dimensiones en inglés. En primer lugar, se obtienen diversos conjuntos de categorías generados en base a diferentes criterios de clasificación. Para cada grupo de categorías se generan dos modelos: Un modelo de lenguaje basado en k-gramas de categorías y un modelo híbrido que es una interpolación de un modelo de lenguaje basado en palabras y uno basado en categorías. Finalmente, se presentan los experimentos realizados sobre un corpus de diálogo espontáneo en castellano para los que se han obtenido resultados de Perplejidad y Word Error Rate
Palabras clave: modelo de lenguaje, categorización, reconocimiento automático del habla, sistema de diálogo

Abstract: The main goal of this work is to study if a language model based on categories could improve the performance of a dialogue system application as it does when not spontaneous and bigger English corpora are used. Firstly, several sets of categories, which are generated on the basis of different classification criteria, are obtained. Then, for each criterion, two language models are generated: A language model based on category k-grams and a hybrid model that is an interpolation of a word-based language model and a category-based language model. Finally, experiments on a spontaneous dialogue corpus in Spanish are reported. These experiments have been carried out in terms of Perplexity and Word Error Rate.
Keywords: language model, categorization, automatic speech recognition, dialogue system

1 Introduction

Dialogue systems are one of the most interesting applications in the field of speech technologies. The aim of these systems is to speak naturally with users to provide them with services such as information of interest or the functional control of machines. The description of these kind of systems can be found in (Zue et al., 2000; Lamel et al., 2000; Seneff and Polifroni, 2000). In a dialogue system there are usually several modules that cooperate to perform the interaction with the user. This is the case of the Automatic Speech Recognition module, the Language Understanding Module, the Dialogue Manager, the Answer Generator and the Synthesizer (Gianchin and McGlashan, 1997).

In this work, we deal with the Automatic Speech Recognition (ASR) module in a dialogue system, specifically with the Language Model (LM). Natural human language is based on a large amount of prior knowledge and this allows us to make several assumptions and to simplify the language that is used. Because of this fact, the use of an appropriate LM that is adapted to the requirements of the application and capable of capturing the structure of the sentences uttered by the speakers is very important.

Nowadays, Statistical LMs are used in the recognition process. Large amounts of training data are required to get a robust estimation of the parameters of such models. However, in the case of dialogue systems, there

* This work has been partially supported by the CICYT project TIN2005-08660-C04-03 and by the Universidad del País Vasco under grant 9/UPV 00224.310-15900/2004.

is not a great deal of training material available. One way to deal with the sparseness of data is to cluster the vocabulary of the application into a smaller number of categories (Niesler and Woodland, 1996).

In this work, we study how categorization (linguistic and statistical) improves the LM of an ASR module in a dialogue system application. The task consists of telephone queries about long-distance train schedules, destinations, and prices uttered by potential users of the system. Firstly, several groups of categories were obtained using different classification criteria. For each criterion, two language models were constructed: a language model based on category k-grams and a hybrid model. The hybrid model is a linear combination of a language model based on word k-grams and a language model based on category k-grams. The experiments that were carried out in terms of Perplexity (PP) and Word Error Rate (WER) showed improvements in the ASR system performance when the hybrid language model was used. These results, obtained using the spoken Spanish, prove the usefulness of category-based models not only with bigger English corpus as Wall Street Journal, as can be seen in previous works (Niesler, Whittaker, and Woodland, 1998), but also in dialogue system applications with fewer training data and spontaneous natural language.

Section 2 deals with the methods used int this work for classifying words into categories, Section 3 describes the two category-based language models. Section 4 details the features of the task and the corpus, and Section 5 deals with the experimental evaluation of the proposals. Finally, Section 6 presents the main conclusions and suggestions for future work.

2 Word categorization

Taking into account that the task under consideration is limited to a restricted semantic domain, two kinds of word classification are proposed: task-dependent categories and task-independent categories. Task-dependent categories seek to take advantage of the knowledge that can be extracted from the semantics of the sentences in the application task. Two different classification criteria are used to generate task-independent categories: a linguistic criterion and a statistical criterion.

2.1 Task-dependent categories

Firstly, a group of task-dependent semantic categories is defined. In the corpus used, speakers ask for information about long-distance train schedules, destinations, and prices, so the more recurrent items are chosen as task-dependent categories: **cities**, **months**, **days**, and **trains**. Different category groups, that have lower appearance ratio, have been studied but they had worse performances. This categorization involves classifying only some words in the vocabulary and not all of them. In this way, a partially categorized corpus is obtained. Words that have not been classified can be viewed as categories that contain a single word.

2.2 Task-independent categories

2.2.1 Linguistic categories

In this section, categories are automatically obtained using a linguistic criterion. A free software application, FreeLing (Carreras et al., 2004), has been used for this purpose. The classes given by FreeLing correspond to the following EAGLE (EAGLES project, 1993–1996) labels and are independent of the task: **adjectives**, **adverbs**, **determinants**, **names**, **verbs**, **pronouns**, **conjunctions**, **interjections**, and **prepositions**. Once the words in the training corpus are labelled, several of them present ambiguity due to the kind of categorization. Thus, some words in Spanish, for example, "deseo", could be a noun or a verb depending on the context, so they cannot be assigned to a single class. Ambiguity in words is solved by the following procedure: the category assigned to the word is decided in accordance with the word's predominant function in the task.

2.2.2 Statistical clustering

In this section, categories are automatically generated using a classical clustering algorithm. 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 (Duda, Hart, and Stork, 2000). The objective function selected in this work is the log-likelihood function in a class bigram model (Martin, Liermann, and Ney, 1998), and the clustering has been done using an iterative algorithm. In this algorithm the number of classes must be set at the beginning. In this work, several sets of 5, 10, 15, 20, 25, 50, 75 and 100

statistical classes were obtained for different purposes.

3 Category-based language models

In this study, a category-based LM is used. Alternatively, an interpolated model has been proposed in order to enrich the category-based LM.

3.1 A language model based on category k-grams

In a first approach, the LM only captures the relations between groups of words and "forgets" about the relations between specific words (Niesler and Woodland, 1996).

Equation 1 shows that the probability of a word sequence $(w_1, \ldots, w_N)$ can be represented as a product of conditional probabilities.

$$\begin{aligned} &P(w_1, \ldots, w_N) = \\ &= P(w_1) \prod_{n=2}^{N} P(w_n | w_1, \ldots, w_{n-1}) \end{aligned} \quad (1)$$

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

Using a language model based on word k-grams (M_w), the conditional probability of the previous expression is approximated as follows:

$$\begin{aligned} &P(w_n|w_1, \ldots, w_{n-1}) \cong \\ &\cong P_{M_w}(w_n|w_{n-k+1}, \ldots, w_{n-1}) \end{aligned} \quad (2)$$

where $P_{M_w}(w_n|w_{n-k+1} \ldots w_{n-1})$ represents the probability of w_n when the sequence of the previous $k-1$ words $(w_{n-k+1}, \ldots, w_{n-1})$ is observed.

Assuming a model based on category k-grams (M_c), the probability of w_n conditioned to its $k-1$ predecessors can be written as follows (Niesler and Woodland, 1996; Nevado, Sánchez, and Benedí, 2001):

$$\begin{aligned} &P_{M_c}(w_n|w_{n-k+1} \ldots w_{n-1}) = \\ &= \sum_{j=1}^{N_C} P(w_n|C_j) P(C_j|C_{j-k+1} \ldots C_{j-1}) \end{aligned} \quad (3)$$

where N_C is the number of different word categories.

Assuming now the restriction that each word belongs to a single class the previous equation is rewritten using the equation 4:

$$\begin{aligned} &P_{M_c}(w_n|w_{n-k+1} \ldots w_{n-1}) = \\ &= P(w_n|C_{w_n}) P(C_{w_n}|C_{w_{n-k+1}} \ldots C_{w_{n-1}}) \end{aligned} \quad (4)$$

$P(C_{w_n}|C_{w_{n-k+1}} \ldots C_{w_{n-1}})$ represents the probability of C_{w_n} when $C_{w_{n-k+1}} \ldots C_{w_{n-1}}$ category sequence has been observed; and C_{w_i} represents the class that w_i belongs to. The parameters of the distributions of words into categories are calculated using expression 5

$$P(w|C) = \frac{N(w|C)}{\sum_{w'} N(w'|C)} \quad (5)$$

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

3.2 Interpolation of word-based and category-based k-gram models

In this section we describe a hybrid model (M_h) that seeks to take advantage of two information sources, i.e., the relations between specific words and between groups of words. It is an interpolation of a word-based LM and a category-based LM and is defined as a linear combination of the two. The probability of the word w_n conditioned to the $k-1$ previous events is given by equation (6), again assuming k-gram based LMs and words belonging to a single class (Benedí and Sánchez, 2005):

$$\begin{aligned} &P_{M_h}(w_n|w_{n-k+1} \ldots w_{n-1}) = \\ &= \lambda P(w_n|w_{n-k+1} \ldots w_{n-1}) + (1-\lambda) \\ &P(w_n|C_{w_n}) P(C_{w_n}|C_{w_{n-k+1}} \ldots C_{w_{n-1}}) \end{aligned} \quad (6)$$

4 Task and corpus

Within the framework of the DIHANA project (DIHANA project, 2005) a human-machine dialogue corpus in Spanish was acquired. Here, speakers ask for information about long-distance train schedules, destinations, and prices by telephone. The features of this corpus are detailed in Table 1.

CORPUS	
dialogues	900
speakers	225
no. sentences	9985
no. training sentences	8606
no. test sentences	1379
vocabulary	938
no. total words	89841

Table 1: Features of the corpus

5 Experimental results

The experiments were carried out using the corpus mentioned above. In this work, several category-based LMs were generated using the different groups of categories described in Section 2. The LMs were evaluated in terms of **PP**, which is the term usually used to measure the quality of LMs. Then, they were integrated into an ASR system and evaluated in terms of **WER**.

5.1 Perplexity results

Firstly, different language models were evaluated in terms of PP. In these preliminary experiments, the PP was measured over the categorized test corpus. Thus, once the test set was labelled with the classes corresponding to each word, the PP was measured over the categories as if they were words. Therefore, these values are very dependent on the number of classes.

Different category-based language models were generated using the following groups of categories:

- task-dependent categories
- linguistic categories
- statistical categories: 10 (in order to compare them to the 10 linguistic categories) and 50.

Different sets of statistical categories (with 15, 20 and 25 classes) were used in order to generate LMs. However, the most representative results were obtained for 10 and 50 classes. These values were compared to the results of PP obtained for a word-based LM.

As Table 2 shows, as the number of categories increased, the values of PP also increased because the size of the vocabulary increased as well. Nevertheless, the values of PP increased slightly compared to the growth in the size of the vocabulary. Table 2 also shows that the values of PP were better with greater values of K, up to a threshold. In spite of this, note that as the value of k increased the size of the model (measured in terms of the number of transitions) also increased substantially. As can be seen in Table 2 the values of PP for 10 statistical and 10 linguistic categories are very similar, however, the values of PP were slightly better for linguistic classes. Word-based LMs, and the category-based LMs that uses task-dependent categories, cannot be compared in terms of PP because of the differences in the size of the vocabulary. The same happens with 50 statistical classes.

K	**PP**				
	word based	**category based**			
		task dep.	**ling.**	**statis.**	
	938	**817**	**10**	**10**	**50**
2	18.17	14.31	5.19	5.02	8.23
3	14.59	11.22	4.42	4.70	6.87
4	14.60	11.10	4.14	4.52	6.61
5	14.85	11.32	4.06	4.47	6.64
6	15.03	11.51	4.07	4.50	6.75

Table 2: PP results for a word-based LM (938 classes) and several LMs based on different sets of categories (task-dependent categories, 10 linguistic categories and statistical categories, 10 and 50). There are 817 task-dependent classes (4 classes + 813 words).

5.2 WER

Finally, the category-based LMs described in Section 3 were generated using the sets of categories mentioned above: task-dependent categories, 10 linguistic categories, and statistical categories (10 and 50). Due to the improvement in the results given by 50 classes compared to the results given by 10 statistical classes, a new experimentation phase was carried out. In this second phase, sets of 75 and 100 statistical classes were obtained, and the corresponding category-based LMs were generated. Then, the models were integrated into the ASR system, using k-gram LMs with a value of $K = 3$ to obtain a reasonable value of PP and a model of limited size. Finally, the models were evaluated in terms of WER.

Table 3 shows the values of WER when a model based on category k-grams was used. A significant reduction in the value of WER was observed when task-dependent categories were used compared to the other groups of categories. Furthermore, a slight improvement was shown compared to the word-based model performance. Note that task-dependent categories are ad-hoc categories, and they take into account the semantics of the sentences, using the categorization of words only in very specific cases.

Alternatively, models based on the remaining classes showed worse results than the word-based language model. This could be

WER (%)		
word-based		19.92
task-dependent		19.57
linguistic (10)		31.91
statistic.	10	30.27
	50	24.20
	75	23.05
	100	22.21

Table 3: Comparison of WER results using word-based and category-based LMs. Different sets of categories are used, task-dependent categories, 10 linguistic categories and statistical categories (10, 50, 75 and 100).

due to the reduced size of the vocabulary in the task. If a larger corpus was used, the reduction in the size of the vocabulary due to categorization would be more noticeable, and the results might be better.

Statistical clusters work better than linguistic classes even with the same number of classes (10). Note that the process used for ambiguous words in linguistic classes is a procedure that may be wrong in some cases, while the statistical clustering algorithm assumes no ambiguity for words as an initialization. Using statistical classes, better results of WER were obtained when a higher number of classes (100) was used; but the WER diminishes asymptotically as the number of classes grows.

Table 4 presents the values of WER for the hybrid model. This table shows a major reduction in WER compared to the values obtained with the k-gram based LMs. Note that the value of λ can be selected in order to minimize the value of WER. In this work we varied the value of λ manually to obtain different results of WER. The results given in the table 4 were obtained for a value of $\lambda = 0.8$, although for high number of classes better results were obtained with lower values of λ. It can be inferred from these results that the interpolation of word-based and category-based LMs improves the values of WER obtained by a simple category-based LM, especially when statistical clustering is used. Furthermore, the hybrid model that used task-dependent and statistical classes achieved better values of WER than the simple word-based model. Table 4 shows that the best results are obtained using 50 statistical classes. Note that the number of classes is an only one factor of those that are involved in the performance of the hybrid model, so by changing the number of classes, we might not reach the absolute minimum, as shown in Table 4.

WER (%)		
word-based		19.92
task-dependent		19.47
linguistic (10)		19.89
statistic.	10	19.53
	50	18.92
	75	19.04
	100	18.95

Table 4: Comparison of WER results for a word-based LM and the hybrid LMs generated with different sets of classes. Task-dependent categories, 10 linguistic categories and statistical categories (10, 50, 75 and 100).

6 Concluding remarks and future work

The experiments show that the hybrid model performs better than the simple category-based model and the word-based model, when it is integrated into the ASR module of a dialogue system. This is due to the fact that the hybrid model takes into account the relations between word classes as well as the relations between specific words. These results show the usefulness of the mentioned model for both an application that uses a big corpus in English, as can be seen in previous works, and a spontaneous Spanish corpus of a reduced size acquired for a dialogue system application.

However, since the improvement is not very significant, an in depth experimentation is required to study new interpolations of models.

In the field of categorization, it could be interesting for future work to study a more general unit of categorization, i.e., segments whose components could be words, combinations of words, or combinations of categories and words.

Bibliografía

Benedí, J.M. and J.A. Sánchez. 2005. Estimation of stochastic context-free grammars and their use as language models. *Computer Speech and Language*, 19(3):249–274.

Carreras, X., I. Chao, L. Padró, and M. Padró. 2004. Freeling: An open-source suite of language analyzers. In *LREC'04*, Lisbon, Portugal. http://garraf.epsevg.upc.es/freeling.

DIHANA project. 2005. Dialogue System for Information Access Using Spontaneous Speech in Different Environments. Comisión Interministerial de Ciencia y Tecnología TIC2002-04103-C03-03. http://www.dihana.upv.es.

Duda, R. O., P. E. Hart, and D. G. Stork. 2000. *Pattern Classification*. Wiley-Interscience, 2nd edition.

EAGLES project. 1993–1996. Expert advisory group on language engineering standards. http://www.lsi.upc.es/∼nlp/tools/parole-sp.html.

Gianchin, E. and S. McGlashan. 1997. Corpus-Based Methods in Speech Processing. *Kluwer Academic, ch. Spoken Language Dyalogue Systems*, pages 67–117.

Lamel, L., S. Rosset, J.L. Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Prouts. 2000. The LIMSI ARISE System. 31(4):339–354, Aug.

Martin, S., J. Liermann, and H. Ney. 1998. Algorithms for bigram and trigram word clustering. *Speech Communication*, 24:19–37.

Nevado, F., J.A. Sánchez, and J.M. Benedí. 2001. 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, pages 183–188, Castellón (Spain).

Niesler, T. R., E. W. D. Whittaker, and P. C. Woodland. 1998. Comparison of part-of-speech and automatically derived categroy-based language models for speech recognition. In *IEEE ICASSP-98*, Seattle, Washington. IEEE.

Niesler, T. R. and P. C. Woodland. 1996. A variable-length category-based n-gram language model. In *IEEE ICASSP-96*, volume I, pages 164–167, Atlanta, GA. IEEE.

Seneff, S. and J. Polifroni. 2000. Dialogue management in the mercury flight reservation system. In *Proc. ANLP-NAACL 2000 Satellite Workshop*, pages 1–6.

Zue, V., S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington. 2000. Jupiter: A telephone-based conversational interface for weather information. In *IEEE Trans. on Speech and Audio Proc.*, pages 8(1):85–96.