Experiments with Linguistic Categories for Language Model Optimization

A. Casillas, A. Varona, I. Torres

Dpt. de Electricidad y Electrónica, Facultad de Ciencias
Universidad del País Vasco (UPV-EHU)
{arantza,amparo,manes}@we.lc.ehu.es

Abstract. In this work[1] we obtain robust category-based language models to be integrated into speech recognition systems. Deductive rules are used to select linguistic categories and to match words with categories. Statistical techniques are then used to build n-gram Language Models based on lexicons that consist of sets of categories. The categorization procedure and the language model evaluation were carried out on a task-oriented Spanish corpus. The cooperation between deductive and inductive approaches has proved efficient in building small, reliable language models for speech understanding purposes.

1 Introduction

Nowadays, Automatic Speech Understanding (ASU) systems require a Language Model (LM) to integrate the syntactic and/or semantic constraints of the language. Thus, ASU systems can allow sequences of words in accordance with the previously defined syntaxis of the application task, as generated by a Language Model. The most reliable way to integrate LM into an ASU system is to infer statistical models, typically n-grams, from large training corpora. Statistical LMs estimate the *a priori* probability $P(\Omega)$ of a sequence of words $\Omega \equiv \omega_1\omega_2 \ldots \omega_{|\Omega|}$ being pronounced. Under the n-grams formalism, the estimation of $P(\Omega)$ is based on the estimation of the probability of observing a word given the $n-1$ preceding lexical units, $P(\omega_i/\omega_1 \ldots \omega_{n-1})$, for every word ω_i in the lexicon and for every potential n-gram, i.e. combination of n words appearing in the application task. However, when the size of the lexicon is high, the size of the training corpora needed to obtain well trained statistical LM's is prohibitive. In fact, the lack of training samples for building reliable LM's is an open problem when designing ASU systems for current application tasks such as dialogue systems, speech to speech translation, spontaneous speech processing, etc. The most common solution is to reduce the size of the lexicon by grouping the different words into categories. The aim of this work is to obtain reliable, small, robust Language Models to be integrated into an ASU system. We combine natural language

[1] This research is supported in part by the Spanish Research Agency, project HERMES (TIC2000-0335-C03-03), DIHANA (TIC2002-04103-C03-02) and by the University of the Basque Country (9/UPV 00224.310-13566/2001).

techniques with traditional speech recognition techniques to tackle our objective. Deductive rules are used to define the set of categories and to determine the category of each word. Statististical techniques are then applied to generate category-based LM's.
A task-oriented Spanish corpus is used for category definition and language model generation and evaluation. This corpus, called BDGEO, represents a set of queries to a Spanish geography database. This is a medium-size vocabulary-specific task designed to test integrated systems for Speech Understanding. The lexicon consists of 1,459 words and the training material used for this work includes 82,000 words.

2 Description of Categories

Two sets of categories (or word equivalence classes) are selected and evaluated in the work. We employ the MACO [1] part of speech tagger to determine the set of categories. The MACO toolkit applies deductive techniques to determine the category of each word. In this case, each word belongs to exactly one word class. In general, the number of word classes is smaller than the number of words, so the size of the LM is reduced and it is better trained.
The first set of categories is based on the main classes recognized by the MACO morphological analyzer. From this initial set, three classes are removed since they have no sense in a Speech Understanding framework: abbreviations, dates and punctuation marks. Nevertheless, we need to add the class *sentence beginning*, required by the statistical language model to independently parse each sentence, i.e. to reset the history of the n-gram model. Table 1 shows the 12 classes that constitute the Reduced Set of Categories (RSC). Table 1 also shows the number of different words that correspond to each category in BDGEO database. This database contains a large number of geographical entities (mount, river, etc.) and proper names (Madrid, Ebro, etc.). Thus, the largest category is *Name*, accounting for 46% of the lexicon: 669 words. The number of different words in the category *Article* almost matches the number of different possibilities of a Spanish article.
The second set of categories includes an extended set of the first one. Categories *name*, *pronoun* and *verb* are now expanded to *proper name*, *interrogative pronoun*, *relative pronoun*, *auxiliar verb*, *main verb*, etc. We have also taken into account the number of each word (singular (S), plural (P) or common (C)) and its gender (masculine (M), feminine (F) or invariable (I)). Table 2 shows the whole Extended Set of Categories (ESC), including characteristic expansion when considered (TFP: article, feminine, plural; TMS: article masculine, singular, etc.). This table also shows the number of different words of BDGEO vocabulary corresponding to each category. This number is significatively reduced for previous *Name* and *Verb* categories.

Table 1. Reduced Set of Categories (RSC) proposed.

category[tag](number)	category[tag](number)	category[tag](number)
Determinant[D00](40)	Adverb[R00](31)	Article[T00](10)
Conjunction[C00](12)	Name[N00](669)	Pronoun[P00](44)
Numeral[M00](47)	Adjective[A00](231)	Interjection[I00](1)
Verb[V00](348)	Preposition[S00](25)	Line begging[S](1)

Table 2. Extended Set of Categories (ESC) proposed. Each cell represents the extension of the corresponding category of Table 1.

tag(number)	tag(number)	tag(number)
DMP(9) DFS(8) DCS(3) DCP(2)	R00(31)	TMP(2) TMS(4) TFS(2) TFP(2)
C00(12)	N00(305) NFS(140) NFP(45) NMS(124) NMP(41) NMN(3) NCS(6) NCP(5)	PCP(5) PCS(9) PFS(5) PMS(7) PT0(2) PCN(7) PFP(3) PMP(4) PRO(2)
[M00](47)	AFS(64) AMS(47) ACS(38) ACP(17) AFP(28) ACN(4) AMP(33)	[I00](1)
VMG(13) VMS0(159) VMSM(13) VMN(33) VMSF(8)VAN(2) VMP0(90) VMPF(11) VMPM(3) VAS0(11) VAP0(5)	[S00](25)	[S](1)

3 Language Model Evaluation

The test set perplexity (PP) is typically used to assess the quality of the LM. Perplexity can be interpreted as the (geometric) average branching factor of the language according to the model. It is a function of both the task and the model. The test set Perplexity (PP) is based on the mean log probability that an LM assigns to a test set ω_1^L of size L. It is therefore based exclusively on the probability of words which actually occur in the test as follows:

$$PP = P(\omega_1^{L)})^{-1/L} = e^{-\frac{1}{L}\sum_{i=1}^{L} log(P(\omega_i/\omega_1^{i-1}))} \quad (1)$$

The test set perplexity depends on the size of the lexicon (classically the number of different words in the task). Actually, the highest value of perplexity that could be obtained is the size of the lexicon when all combinations of words are equally probable. Low perplexity values are obtained when high probabilities are assigned to the test set events by the LM being evaluated, i.e., when "good" LM's are obtained.

Several n-gram models, $n = 2, ..., 4$, are obtained using the CMU toolkit [2]. In each case, the proposed set of categories, RSC ($l = 12$ categories in Table 1) and ESC ($l = 52$ categories in Table 2), is considered to build the language model. For comparison purposes, language models based on the lexicon consisting of the whole set of words ($l = 1459$), are also considered. In such cases, N Categorization (NC) is carried out. Table 3 shows the size of each model measured by the number of different n-grams appearing in the training set. Each model is then evaluated in terms of test set perplexity (PP). The whole database is used to train and test models, maintaining training-test set independency by using the

well-known *leaving-one-out* partition procedure. As the three sets of categories lead to very different lexicon sizes (l), the PP cannot be directly compared in these experiments. Thus, a new measure, PP/l, is also included in Table 3.

Table 3. Perplexity (PP) evaluation of n-grams with $n = 2, ..., 4$ for three different lexicons (l): reduced set of categories (RSC) (Table 1), extended set of categories (ESC) (Table 2 sets of categories and no categorization (NC). The number of different n-grams (size) as well as pp/l measure are also provided.

sets of categories		n=2			n=3			n=4		
	lexicon (l)	size	PP	PP/l	size	PP	PP/l	size	PP	PP/l
NC	1459	7971	21.96	0.02	21106	14.99	0.01	36919	14.05	0.01
ESC	52	972	9.48	0.18	5043	6.61	0.13	13439	5.96	0.11
RSC	12	133	5.05	0.42	808	3.94	0.33	2962	4.04	0.34

Table 3 shows important reductions in the size of the model and in PP when linguistic sets of categories are considered. Both measures decrease with the size of the lexicon, leading to smaller, better trained, more efficient Language Models. However, when the number of categories is too small (RSC) the number of words corresponding to each category can be very high (see Table 1), making recognition work more difficult. The aim of an ASU system is to provide the most probable sequence of words according to the acoustic sequence uttered. Therefore, the most probable word in each category has to be selected when the LM is based on categories. This is measured to some extent by PP/l, which expresses a perplexity per lexicon unit. This value is lowest when each category consists of a single word and higher for small sets of categories. Up to a opint, it therefore gauges the difficulty if decoding a task sentence. A good agreement between this measure, PP and model size and trainability is represented by the extended set of categories (ESC).

4 Conclusions

The objective of our experiments is to reduce LM complexity to get a set of well trained LM's. Thus, two sets of linguistical categories (or word classes) are evaluated in terms of perplexity. Both sets lead to small, low-perplexity language models that can be trained with reduced training corpora. Experimental comparison carried out in this work enables us to propose the extended set of categories, which includes the number of the word (singular and plural), the gender of the word (masculine and feminine), etc., as an adequate lexicon for building statistical language models for this task. However, these language models need to be integrated into the ASU system to be compared in terms of final word error rates.

References

[1] "The MACO Morphological Analyzer." http://www.lsi.upc.es/ nlp
[2] "The CMU-Cambridge Statistical Language Modeling toolkit." http://svr-www.eng.cam.ac.uk/ prc14/toolkit-documentation.html