Phrases in Category-based Language Models for Spanish and Basque ASR

Raquel Justo, M. Inés Torres

Dep. of Electricity and Electronics, University of the Basque Country, Spain
webjulibr@lg.ehu.es, manes@we.lc.ehu.es

Abstract

In this work, we integrate phrases or segments of words into class n-gram language models in order to take advantage of two information sources: words and categories. Two different approaches to this kind of models are proposed and formulated. The models were integrated into an Automatic Speech Recognition system and subsequently evaluated in terms of word error rate. The experiments, carried out over two different databases and languages, demonstrate that a language model based on categories composed by phrases can outperform classical class n-gram language models.

Index Terms: speech recognition, language models, categorization, phrases or segments of words.

1. Introduction

Statistical Language Models (LM) are broadly used in Automatic Speech Recognition (ASR) systems, being word n-gram LMs the most widely used approach, because of their effectiveness when it comes to minimizing the word error rate [1]. Large amounts of training data are required to get a robust estimation of the parameters defining such models. However when dealing with an ASR application for which the amount of training material available is limited, alternative approaches as a class n-gram LM [2], could be used.

A class n-gram LM is more compact and generalizes better on unseen events. Nevertheless, it only captures the relations between the categories of words, while it assumes that the inter-word transition probability depends only on the word classes. This fact degrades the performance of the ASR system. To avoid the loss of information associated with the use of a class n-gram LM, other authors have proposed different approaches, e.g. model interpolation, aiming to take advantage of both the accurate modeling of word n-grams for frequent events, and the predictive power of class n-gram models for unseen or rare events [3, 4].

On the other hand, using phrases or word segments is a technique that has already successfully been used within both the speech recognition [5] and machine translation [6] frameworks. In this work a LM based on categories made up of segments is employed in order to combine the benefits of word-based and category-based models [7, 8]. Two different approaches to category-based models are proposed and formulated. In both we make use of segments of words within the categories, in order to capture the relations between words in each category. The proposed models, fully explained in Section 3, were integrated into an ASR system and evaluated in terms of word error rate.

Several series of experiments were carried out using two different languages and corpora, (Section 5). These experiments show that the integration of word segments into a category-based LM yields a better performance of the ASR system.

2. Categories and word segments or phrases

In this work, a set of segments and a set of categories formed by those segments were obtained from the selected corpora. Two different types of criteria were used.

Linguistic categories and segments: In this case, the set of segments and the set of categories were simultaneously obtained under a linguistic criterion by applying a rule based method. These categories are independent of the task and consist of word segments having the same linguistic function. This set of categories, as well as the segments the categories are made up of, were provided by ametzagaina (only for the Spanish corpus). Furthermore, they provided us with the segmented and categorized corpus.

Statistical categories and segments: In this case, we first obtained a set of segments using a statistical criterion. The most frequent n-grams of the corpus were selected as segments. In this sense, and in order to avoid rare or unimportant n-grams, a minimum number of occurrences was required. Then, a segmented training corpus was generated with the set of segments. Finally, different sets of statistical categories constituted by the defined segments were obtained with the aid of mkcls [9].

3. Word segments in category-based language models

Two different approaches to category-based LMs are formulated below. Both of them are generated introducing segments or sequences of words inside the categories that are going to be used. However, in the first approach, $M_w$, the words in a segment are separately studied and the transition probability among them is calculated. Therefore, the same vocabulary of a classical word-based LM is considered. In the second approach instead, $M_s$, the words in a segment are joined and the whole segment is treated as a unique word or lexical unit, so as new words need to be considered in the vocabulary.

3.1. LMs based on categories consisting of word sequences

Our goal is to estimate the probability of a sequence of N words $\bar{w} = w_1, w_2, \ldots, w_N$ in accordance with a LM based on categories consisting of segments of words, $M_w$.

Let us define a segmentation $(s)$ of the sequence of words into $M$ segments, as a vector of $M$ indexes, $s = (a_1, \ldots, a_M)$, such that $a_1 \leq \ldots \leq a_M = N$. The $\bar{w}$ sequence of words can be represented in terms of such segmentation as follows:

$$\bar{w} = w_1, \ldots, w_N = w_{a_0}^{a_1}, \ldots, w_{a_M-1}^{a_M} \quad (1)$$

$\dagger$Ametzaína R&D group, member of the Basque Technologic Network, http://www.ametzain.com
where \( w_{ai}^{a_{i-1}} = w_{ai-1}, \ldots, w_{ai} \). The set of all possible segmentations of a \( \bar{w} \) sequence of words is denoted as \( S(\bar{w}) \).

On the other hand, let \( C = \{ c_i \} \) be a previously defined set of categories, selected using any classification criteria (statistical, linguistic, \ldots). Each category consists of a set of segments previously defined as well (selected as described Section 2), whereas each segment within a given category is made up of a sequence of several words. If the \( \bar{w} \) sequence of words is categorized using the \( C \) set of categories, the corresponding category sequence is written as \( \bar{c} = c_1, c_2, \ldots, c_T \) where \( T \leq N \).

In this work, the only segmentations that will be taken into account will be those compatible with the possible sequences of categories \( (c) \) associated to each sequence of words. This set of segmentations is denoted by \( S_c(\bar{w}) \). That is, only segmentations having the following form will be considered

\[
\bar{w} = w_1, \ldots, w_N = w_{a_0}^{a_1}, \ldots, w_{a_{N-1}}^{a_N}
\]  
(2)

where \( w_{a_i-1}^{a_i} \) must be a segment belonging to the \( c_i \) category.

The segmentation of a sequence of words can be understood as a hidden variable. In this way, the probability of a sequence of words \( \bar{w} \), according to a \( M_w \) LM can be obtained by means of Equation 3

\[
P_{M_w}(\bar{w}) = \sum_{\bar{c} \in C} \sum_{\bar{a} \in S_c(\bar{w})} P(\bar{w}, \bar{c}, \bar{a}) = \sum_{\bar{c} \in C} \sum_{\bar{a} \in S_c(\bar{w})} P(\bar{w}, \bar{c}) P(\bar{c}) = \sum_{\bar{c} \in C} \sum_{\bar{a} \in S_c(\bar{w})} P(\bar{w}|s, \bar{c}) P(s|\bar{c}) P(\bar{c})
\]  
(3)
being \( C \) the set of all the possible category sequences, given a predetermined set of categories \( C \).

The probability of a given sequence of categories, \( p(\bar{c}) \), can be calculated as a product of conditional probabilities; that is:

\[
P(\bar{c}) = \prod_{i=1}^{T} P(c_i|c_{i-1}^{i-1})
\]  
(4)

In order to estimate the \( P(c_i|c_{i-1}^{i-1}) \) probability, a set of \( k_a \) \( k \)-TSS models, where \( k = 1, \ldots, k_a \), has been used, as Equation 5 shows:

\[
P(c_i|c_{i-1}^{i-1}) \approx P(c_i|c_{i-1}^{i-1})
\]  
(5)

\( k \)-TSS models [10] can be considered to be a syntactic approach to the well-known n-gram LMs, where \( n \) is represented by \( k \) in the \( k \)-TSS model. However, this approach permits us to obtain a unique Stochastic Finite State Automaton (SFSA) that integrates \( K \) \( k \)-TSS models (where \( k = 1, \ldots, K \)) and allows for back-off smoothing [11]. Acoustic models are then easily incorporated to this network into an ASR system.

The term \( P(s|\bar{c}) \), on the other hand, could be estimated using different approaches: zero or higher-order models, etc. However, let us assume, here, the segmentation probability to be constant \( P(s|\bar{c}) = \alpha \), as proposed in several phrase-based statistical machine translation works [12].

Finally, \( P(\bar{w}|s, \bar{c}) \) is estimated in accordance with zero-order models. Thus, given a sequence of categories \( \bar{c} \), and a segmentation \( s \), the probability of a segment given its category \( c_i \) only depends on this \( c_i \) category, but not on the previous ones, so that:

\[
P(\bar{w}|s, \bar{c}) = \prod_{i=1}^{T} P(w_{ai}^{a_{i-1}}|c_i)
\]  
(6)

To estimate the \( P(w_{ai}^{a_{i-1}}|c_i) \) probability \( k_a \) smoothed \( k \)-TSS models, where \( k = 1, \ldots, k_a \), have been used for each category. Thus, a LM that studies the inter-word transition probability in each category is generated.

\[
P(w_{ai}^{a_{i-1}}|c_i) \approx \prod_{j=a_{i-1}+1}^{a_i} P(w_j|w_{j-k_a-1}^{j-k_a+1}, c_i)
\]  
(7)

Summing up, Equation 3 can be rewritten as the following Equation:

\[
P_{M_w}(\bar{w}) = \alpha \sum_{\bar{c} \in C} \sum_{\bar{a} \in S_c(\bar{w})} P(\bar{w}, \bar{c}) P(\bar{c}) = \sum_{\bar{c} \in C} \sum_{\bar{a} \in S_c(\bar{w})} P(\bar{w}|s, \bar{c}) P(s|\bar{c}) P(\bar{c})
\]  
(8)

Under this approach, several SFSA need to be integrated into the ASR system: a SFSA representing the transition probabilities among categories as well as one additional SFSA for each category, representing the transition probabilities among the words contained in the segments of the category. These SFSA were integrated “on-the-fly” [13] in the search network when needed.

3.2. LMs based on categories consisting of joined sequences of words

In a second approach, a LM based on categories consisting of joined word sequences, \( M_s \), is considered. In this approach, each segment \( w_{a_i}^{a_{i+1}} \) in the corpus is considered as a new lexical unit that cannot be divided into different words. Let us denote each lexical unit by \( l_i \), where \( i \in \{ \Sigma \} \) being \( \{ \Sigma \} \) the set of all the possible segments previously defined (the same mentioned in 3.1 but assuming now that they cannot be separated in different words). Thus, a sequence of lexical units \( \bar{l} = l_1, \ldots, l_M \) corresponds to a specific segmentation \( (s) \) of the sequence of words \( \bar{w} = w_{a_0}^{a_1}, \ldots, w_{a_M}^{a_{M+1}} \).

\[
\bar{w} = \underbrace{w_{a_0}^{a_1}, \ldots, w_{a_M}^{a_{M+1}}}_{\bar{l} = \{ l_i \}}
\]  
(9)

Assuming again only that segmentations compatible with a given category sequence \( \bar{c} = c_1, \ldots, c_T \) are to be considered; the possible sequences of lexical units, for a given segmentation of words, have the following form \( \bar{l} = l_1, \ldots, l_T \).

A sequence of lexical units involves a specific segmentation itself, thus, in this case, \( \bar{l} \) is considered as a hidden variable and the probability of a sequence of words is given by Equation 10

\[
P_{M_s}(\bar{w}) = \sum_{\bar{c} \in C} \sum_{\bar{l} \in L_c(\bar{w})} P(\bar{w}, \bar{c}, \bar{l}) = \sum_{\bar{c} \in C} \sum_{\bar{l} \in L_c(\bar{w})} P(\bar{w}|\bar{c}, \bar{l}) P(\bar{c}) P(\bar{l}|\bar{c}) = \sum_{\bar{c} \in C} \sum_{\bar{l} \in L_c(\bar{w})} P(\bar{w}|\bar{c}, \bar{l}) P(\bar{c}) P(\bar{l}|\bar{c}) P(\bar{l}|\bar{c})
\]  
(10)

being \( C \) the set of all the possible category sequences, given a predetermined set of categories \( C \) and \( L_c(\bar{w}) \) the set of all the possible sequences of lexical units compatible with the given sequence of words and the possible sequences of categories.

The third term in Equation 10, \( P(\bar{c}) \), is estimated as stated in Equations 4 and 5 (see previous Section).

The second term in Equation 10 is the probability of a sequence of lexical units given a sequence of categories. Assuming again zero-order models, this probability is calculated as:

\[
p(\bar{l}|\bar{c}) = \prod_{i=1}^{T} P(l_i|c_i)
\]  
(11)
A $k$-TSS model, with $k = 1$, i.e. a unigram, has been used to estimate this kind of probability for each category.

Finally, the first term in Equation 10, $P(\bar{w}[i,c])$ is equal to 1 when the sequence of lexical units, $\bar{l}$, and the sequence of categories, $\bar{c}$, are compatible with the sequence of words, $\bar{w}$, and 0 otherwise.

Summing up Equation 10 can be rewritten as follows:

$$P_{M_{sl}}(\bar{w}) \simeq \sum_{\bar{l} \in L} \sum_{\bar{c} \in C} \prod_{i=1}^{T} \left[ P(\bar{l}[i,c])P(c[i,c]_{i-k_a+1}) \right]$$

(12)

Here, smoothed $k$-TSS models are used again to represent the category based LM. The corresponding SFSAs are integrated in the search network represented by Equation 12 “on-the-fly” only when required.

4. Task and corpus

The experiments, presented in the next section, were conducted over two different corpora. The first one, DIHANA [14], consists of human-machine dialogues in Spanish. This corpus 225 speakers ask by telephone for information about long-distance train timetables, fares, destinations and services. A total of 900 dialogues were acquired using the Wizard of Oz technique. This task has intrinsically a high level of difficulty due to the spontaneity of the speech and the problematic derived from the acquisition of large amount of transcriptions, of human-machine dialogues, for training purpose. The other one, METEUS [15], is a read speech corpus in Basque. It consists of wether forecast reports picked up from the internet. Basque is a minority language that shares official status with Spanish in a community of 2.5 million of inhabitants. It is a very inflected language both in nouns and verbs and regarding its syntactic structure is very different from Spanish. Due to their specific features, the two corpora are well-suited in order to study improvements derived from segmentation and categorization within the LM. The features of the corpora are detailed in Table 1.

5. Experiments and Results

The LMs proposed in this work were fed into an ASR system, which was subsequently evaluated in terms of word error rate (WER). The ASR system makes use of the Viterbi Algorithm to search for the best sequence of uttered words for a given sequence of acoustic observations. Thus, the decoder finds the best sequence of states through a probabilistic network, combining categories, segments, words and acoustic models. The acoustic models are continuous Hidden Markov Models.

5.1. Experiments over the corpus DIHANA

Two series of experiments were carried out in order to evaluate the two proposed approaches in Section 3.

Firstly, LMs based on categories consisting of word sequences, $M_{sw}$, was fed into the ASR system, according to the equations in Section 3.1. Making use of this LM, different experiments were carried out, choosing for all of them a value of $k_a = 3$ and $k_b = 2$. First of all, a set of 57 linguistic categories was employed. For this set of categories 3,851 different segments were extracted from the segmented corpus provided by ametzaz, as described in Section 2. Then, 50, 100, 200, 300 and 400 statistical classes were used, all of them with previously defined 1,289 different statistical segments, obtained with the statistical technique described in Section 2.

On the other hand, LMs based on categories consisting of joined sequences of words, $M_{sl}$, were integrated into the ASR system according to the equations in 3.2. A value of $k_a = 3$ was established. Experiments were carried out using the same sets of linguistic and statistical categories described above. The same sets of segments were also employed here.

Table 2 illustrates WER results using the proposed LMs and a classical word-based LM (with a value of $k = 3$, i.e. a trigram) as a baseline.

The results obtained in Table 2 were also compared with the values of WER obtained in another work [16], over the same task and using a classical class n-gram ($k$-TSS in our case) model with categories made up of isolated words. As shown in the mentioned work, classical class-based LMs using 50, 75 and 100 statistical classes achieve WER values of 24.20, 23.05 and 22.22 respectively. It can be concluded from this, that better results are obtained when using word segment based categories (in both $M_{sw}$ and $M_{sl}$ models), than when employing classical class n-gram LMs.

Looking at the results in Table 2 it can be concluded that statistical categories yield better results than linguistic ones. On the other hand, regarding the experiments carried out with the $M_{sl}$ model, a significant drop of the WER is observed compared to the previous model ($M_{sw}$) for all of the selected sets of categories (e.g. a 6.7% with 300 classes). Furthermore, the result obtained with 300 statistical classes and an $M_{sl}$ model improves the WER values obtained with the word n-gram ($k$-TSS in our case) LM ($M_{w}$) by a 2.8%.

5.2. Experiments over the corpus METEUS

Experiments carried out over DIHANA were repeated over METEUS, using again the two proposed LMs ($M_{sw}$ and $M_{sl}$). However, in this case, only statistical classes were employed.
Table 3: WER results using METEUS, for a classical word n-gram LM \( (M_w) \), a classical class n-gram LM \( M_c \) and the two proposed category-based LMs containing segments of words \( (M_{sw} \) and \( M_{sl} \)). Different sets of 300, 400, 500 and 600 statistical classes were employed in all category-based LMs.

<table>
<thead>
<tr>
<th>stat. cat.</th>
<th>( M_{sl} )</th>
<th>( M_{sw} )</th>
<th>( M_c )</th>
<th>( M_w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>6.17</td>
<td>6.68</td>
<td>6.66</td>
<td>5.91</td>
</tr>
<tr>
<td>400</td>
<td>6.01</td>
<td>6.59</td>
<td>6.53</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>5.81</td>
<td>6.37</td>
<td>6.47</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>6.02</td>
<td>6.51</td>
<td>6.62</td>
<td></td>
</tr>
</tbody>
</table>

6. Concluding remarks and future work

In this work, we propose a LM that combines the benefits of both word-based and category-based models. For this purpose, segments of words have been integrated into the classes of a class n-gram LM \( M_w \), with categories made up of isolated words, as shown in Table 3. The differences among all the LMs seem to be less significant than in the experiments over the corpus in Spanish. That could be due to the specific features of the task. Since we are dealing with a read corpus, the performance of the ASR system with the baseline LM is much better, therefore improvements over the obtained results are more difficult to notice. Nevertheless, the results provided by \( M_{sl} \) are better than the results obtained by \( M_{sw} \) and \( M_c \) (with 500 classes 8.8% and 10.2% respectively), so as the same tendency shown in the experiments over the previous corpus is also observed here. Furthermore, using the \( M_{sl} \) and 500 statistical classes a slightly improvement (1.7%) is achieved regarding the baseline word based LM.

7. References