Segment-Based Classes for Language Modeling Within the Field of CSR

Raquel Justo and M. Inés Torres

Dept. of Electricity and Electronics, University of the Basque Country. Spain
raquel.justo@ehu.es, manes@we.lc.ehu.es

Abstract. In this work, we propose and formulate two different approaches for the language model integrated in a Continuous Speech Recognition System. Both of them make use of class-based language models where classes are made up of segments or sequences of words. On the other hand, an interpolated model of a class-based language model and a word-based language model is explored as well. The experiments carried out over a spontaneous dialogue corpus in Spanish, demonstrate that introducing segments of words in a class-based language model a better performance of a Continuous Speech Recognition system can be achieved.

Keywords: language model, classes, segments of words.

1 Introduction

Within the field of Continuous Speech Recognition (CSR) the use of a language model (LM) is required in order to represent the way in which the combination of words is carried out in a specific language. Nowadays, Statistical Language Models (SLMs), based on n-grams, are the most commonly used approach in CSR [1]. They learn the frequency of occurrence of word sequences from a training corpus. Specifically, word n-gram LMs have demonstrate their effectiveness when it comes to minimizing the *word error rate* (WER) [2]. Alternatively, some formalism based on regular grammars and context free grammars have also been used in language modeling [3]. Language constraints, such as long-term dependencies, could be better modeled under this kind of syntactic approaches. However, they still present difficulties of learning and integrating, e.g. into a Continuous Speech Recognition system, when dealing with complex, real tasks.

In this work, we take advantage of both approaches by using *k-testable in the strict sense* (k-TSS) LMs. k-TSS languages are a subclass of regular languages and can be inferred from a set of positive samples by an inference algorithm [4]. k-TSS LMs are considered as the syntactic approach of the well-known n-gram models, where n is represented by k in the k-TSS model. This syntactic approach leads to the use of a Stochastic Finite State Automaton (SFSA) to represent the LM at decoding time. Moreover, the required smoothing, needed to deal with unseen events, is carried out by interpolating K k-TSS models, where $k = 1, \ldots, K$, into

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 714–723, 2007.

a unique smoothed SFSA under a backing-off structure [5]. Then, acoustic models can easily be incorporated to this network into a CSR system.

Large amounts of training data are required to get a robust estimation of the parameters defining the mentioned models. However, there are numerous CSR applications, e.g. human-machine dialogue tasks, for which the amount of training material available is rather limited. One of the ways to deal with sparseness of the data is to cluster the vocabulary of the application into a smaller number of classes. Thus, an alternative approach, as a class n-gram LM, could be used [6,7].

A class n-gram LM is more compact and generalizes better on unseen events. Nevertheless, it only captures the relations between the classes of words, while it assumes that the inter-word transition probability depends only on the word classes. This fact degrades the performance of the CSR 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 [6,8,9].

On the other hand, using phrases or word segments is a technique that has already successfully been used in language modeling for speech recognition [10,11,12] and machine translation [13]. In this work, a LM based on classes made up of segments of words is employed in order to combine the benefits of word-based and class-based models. That is, a class n-gram (k-TSS in our case) LM is generated to deal with the sparseness of the data. However, the proposed classes consist of sequences or segments of words, instead of being made up of isolated words. Therefore, the relations between words can be captured inside each class.

We propose and formulate in this work two different approaches to class k-TSS LMs based on word segments. Both are fully explained in Section 2. On the other hand an interpolated model is proposed as well. Such a model is defined as a linear combination of a word-based and a class-based LM, where classes are made up of segments of words.

The proposed models were integrated into a CSR module in a dialogue system application. The task consists of telephone queries about long-distance train timetables, destinations and fares uttered by potential users of the system. Several series of experiments were carried out on a spontaneous dialogue corpus in Spanish, in order to asses the proposed models (Section 5). These experiments show that the integration of word segments into a class-based LM yields a better performance of the CSR system.

2 Word Segments in Class-Based Language Models

Two different approaches to class-based LMs are formulated below. Both of them are generated introducing segments or sequences of words inside the classes of a class-based LM. However, in the first approach, M_{sw}, the words in a segment are separately studied and the transition probability among them is calculated.

In the second approach instead, M_{sl}, the words in a segment are joined and the whole segment is treated as a unique new "word" or lexical unit. Finally, a hybrid model is proposed as a linear combination of a word based and a class-based LM.

2.1 LMs Based on Classes of Word Segments: M_{sw}

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 classes consisting of segments.

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:

$$\bar{w} = w_1, \ldots, w_N = w_{a_0=1}^{a1}, \ldots, w_{a_{M-1}+1}^{a_M=N} \tag{1}$$

where $w_{a_{i-1}+1}^{a_i} = w_{a_{i-1}+1}, \ldots, w_{a_i}$. 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 classes, selected using any classification criteria. Each class consists of a set of segments previously defined as well. Each segment within a given class is made up of a sequence of several words. If the words in $\bar{w}$ are classified using the C set of classes, the corresponding sequence of classes is written as $\bar{c} = c_1, c_2, \ldots, c_T$ where $T \leq N$.

In this work, only segmentations compatible with the possible sequences of classes ($\bar{c}$) associated to each sequence of words are considered. This set of segmentations is denoted by $\mathcal{S}_{\bar{c}}(\bar{w})$. That is, only segmentations having the following form will be considered

$$\bar{w} = w_1, \ldots, w_N = w_{a_0=1}^{a1}, \ldots, w_{a_{T-1}+1}^{a_T=N} \tag{2}$$

where $w_{a_{i-1}+1}^{a_i}$ must be a segment belonging to the c_i class.

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 LM based on classes made up of segments (M_{sw}), can be obtained by means of Equation 3

$$\begin{aligned} P_{M_{sw}}(\bar{w}) = \sum_{\forall \bar{c} \in \mathcal{C}} \sum_{\forall s \in \mathcal{S}_{\bar{c}}(\bar{w})} P(\bar{w}, \bar{c}, s) s = \sum_{\forall \bar{c} \in \mathcal{C}} \sum_{\forall s \in S_{\bar{c}}(\bar{w})} P(\bar{w}, s|\bar{c}) P(\bar{c}) = \\ = \sum_{\forall \bar{c} \in \mathcal{C}} \sum_{\forall s \in \mathcal{S}_{\bar{c}}(\bar{w})} P(\bar{w}|s, \bar{c}) P(s|\bar{c}) P(\bar{c}) \end{aligned} \tag{3}$$

being $\mathcal{C}$ the set of all the possible class sequences, given a predetermined set of classes C.

The probability of a given sequence of classes, $p(\bar{c})$, can be calculated as a product of conditional probabilities, as Equation 4 shows. The history (c_1^{i-1}) is usually truncated to the n most recent categories, when classical n-grams are used, or to the k_c most recent categories under the k-TSS approach, where k_c is the maximum length of the considered class history.

$$P(\bar{c}) = \prod_{i=1}^{T} P(c_i|c_1^{i-1}) \simeq \prod_{i=1}^{T} P(c_i|c_{i-k_c+1}^{i-1}) \tag{4}$$

The term $P(s|\bar{c})$, on the other hand, could be estimated using different approaches: zero or higher-order models, assuming that all the segmentations have the same probability, etc. Let us assume, in this work, the segmentation probability to be constant $P(s|\bar{c}) = \alpha$, as proposed in several phrase-based statistical machine translation works [14].

Finally, $P(\bar{w}|s, \bar{c})$ is estimated in accordance with zero-order models. Thus, given a sequence of classes $\bar{c}$, and a segmentation s, the probability of a segment given a class c_i only depends on this c_i class, but not on the previous ones, as Equation 5 shows.

$$P(\bar{w}|s, \bar{c}) \simeq \prod_{i=1}^{T} P(w_{a_{i-1}+1}^{a_i}|c_i) \tag{5}$$

The term $P(w_{a_{i-1}+1}^{ai}|c_i)$ represents the probability of a sequence of words, which must be a segment, given the class of this segment and is estimated using a k-TSS model as shown below.

$$P(w_{a_{i-1}+1}^{ai}|c_i) \simeq \prod_{j=a_{i-1}+1}^{a_i} P(w_j|w_{j-k_w+1}^{j-1}, c_i) \tag{6}$$

where k_w stands for the maximum length of the word history that is considered in each class c_i.

Summing up, the probability of a sequence of words can be computed from Equation 7:

$$P_{M_{sw}}(\bar{w}) \simeq \alpha \sum_{\forall \bar{c} \in \mathcal{C}} \sum_{\forall s \in \mathcal{S}_{\bar{c}}(\bar{w})} \prod_{i=1}^{T} \left[\left[\prod_{j=a_{i-1}+1}^{a_i} P(w_j|w_{j-k_w+1}^{j-1}, c_i) \right] P(c_i|c_{i-k_c+1}^{i-1}) \right] \tag{7}$$

Under this approach, several SFSAs need to be integrated into the CSR system: a SFSA representing the transition probabilities among classes as well as one additional SFSA for each class, representing the transition probabilities among the words contained in the segments of the class. Moreover, acoustic models should also be integrated in the search network. A static full integration of all these models is computationally prohibitive, thus, each SFSA is integrated "on the fly" [15] in the search network only when needed.

2.2 LMs Based on Classes of Linked Words: M_{sl}

In a second approach, we propose a LM based on classes consisting of joined sequences of words. In this approach each segment, $w_{a_{i-1}+1}^{a_i}$, will be considered as a new lexical unit that cannot be divided into different words. Let us denote each lexical unit by l_i, where $l_i \in \{\Sigma\}$, being $\{\Sigma\}$ the previously defined set of all the possible segments that will be inside the classes. The same hypothetical sets of segments and classes of 2.1 are considered here but assuming now that the segments 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 $\bar{w}$.

$$\bar{w} = \underbrace{w_{a_0=1}^{a1}}_{l_1}, \ldots, \underbrace{w_{a_{M-1}+1}^{a_M}}_{l_M} \tag{8}$$

Assuming again that only segmentations compatible with a given class sequence ($\bar{c} = c_1, \ldots, c_T$) are to be considered; the possible sequences of lexical units, for a given sequence of words, will have the following form $\bar{l} = l_1, \ldots, l_T$, where l_i is a segment belonging to c_i.

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 9.

$$\begin{aligned} P_{M_{sl}}(\bar{w}) = \sum_{\forall \bar{c} \in \mathcal{C}} \sum_{\forall \bar{l} \in \mathcal{L}_{\bar{c}}(\bar{w})} P(\bar{w}, \bar{c}, \bar{l}) = \sum_{\forall \bar{c} \in \mathcal{C}} \sum_{\forall \bar{l} \in \mathcal{L}_{\bar{c}}(\bar{w})} P(\bar{w}, \bar{l}|\bar{c}) P(\bar{c}) = \\ = \sum_{\forall \bar{c} \in \mathcal{C}} \sum_{\forall \bar{l} \in \mathcal{L}_{\bar{c}}(\bar{w})} P(\bar{w}|\bar{l}, \bar{c}) P(\bar{l}|\bar{c}) P(\bar{c}) \end{aligned} \tag{9}$$

being $\mathcal{C}$ the set of all the possible class sequences, given a predetermined set of classes C. $\mathcal{L}_{\bar{c}}(\bar{w})$ is the set of all the possible sequences of lexical units compatible with the given sequence of words and the possible sequences of classes.

The third term in Equation 9, $P(\bar{c})$, is estimated as stated in Equations 4 (see previous Section).

The second term in Equation 9 is the probability of a sequence of lexical units given a sequence of classes. Assuming again zero-order models, this probability is calculated as:

$$p(\bar{l}|\bar{c}) = \prod_{i=1}^{T} P(l_i|c_i) \tag{10}$$

A k-TSS model, with $k = 1$, i.e. an unigram, has been used to estimate this kind of probability for each class.

Finally, the fist term in Equation 9, $P(\bar{w}|\bar{l}, \bar{c})$ is equal to 1 when the sequence of lexical units, $\bar{l}$, and the sequence of classes, $\bar{c}$, are compatible with the sequence of words, $\bar{w}$, and 0 otherwise. Taking into account that the restriction $\bar{l} \in \mathcal{L}_{\bar{c}}(\bar{w})$ has been established, the term $P(\bar{w}|\bar{l}, \bar{c})$ is equal to 1 in all the cases we have considered.

Summing up Equation 9 can be rewritten as follows:

$$P_{M_{sl}}(\bar{w}) \simeq \sum_{\forall \bar{c} \in \mathcal{C}} \sum_{\forall \bar{l} \in \mathcal{L}_{\bar{c}}(\bar{w})} \prod_{i=1}^{T} \left[P(l_i|c_i) P(c_i|c_{i-k_c+1}^{i-1}) \right] \tag{11}$$

Here, smoothed k-TSS models are used again to represent the class based LM. The corresponding SFSAs are integrated in the search network represented by Equation 11 "on-the-fly" only when required.

2.3 Interpolating an M_{sw} Model and a Word-Based LM, M_h

The interpolation of a class-based and a word-based LM has demonstrated to outperform both mentioned models. In this work a hybrid model (M_h) is defined

as a linear combination of a a word-based LM, M_w, and a LM based on classes made up of word segments, M_{sw}.Using such a model the probability of a word sequence is given by Equation 12.

$$P_{M_h}(\bar{w}) = \lambda P_{M_w}(\bar{w}) + (\lambda - 1)P_{M_{sw}}(\bar{w}) \tag{12}$$

In the above equation, the term $P_{M_w}(\bar{w})$ is the probability of a word sequence using a classical word-based language model, and in this work, a k-TSS model was used to estimate this probability, as Equation 13 shows.

$$P_{M_w}(\bar{w}) = \prod_{i=1}^{N} P(w_i|w_1^{i-1}) \simeq \prod_{i=1}^{N} P(w_i|w_{i-k+1}^{i-1}) \tag{13}$$

The term $P_{M_{sw}}$ is the probability given by Equation 7 in Section 2.1.

3 Classes and Word Segments

In order to deal with the proposals presented in the previous Section, a set of segments and a set of classes formed by those segments needed to be obtained from the selected corpus. Two different types of criteria were used.

Statistical classes 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. In the experiments shown in Section 5 the n-grams (where $1 \leq n \leq 5$) appearing in the corpus a number of times above a prefixed threshold were included in the set of the defined segments. Then, a segmented training corpus was generated with the set of segments. Finally, different sets of statistical classes constituted by the defined segments were obtained with the aid of *mkcls* [16].

Linguistic classes and segments: In this case, the set of segments and the set of classes were simultaneously obtained under a linguistic criterion by applying a rule based method. These classes are independent of the task and consist of word segments having the same linguistic function in the sentence. This set of classes, as well as the segments the classes are made up of, were provided by *ametzagaina*[1]. Furthermore, they provided us with the segmented and classified corpus. An example of some employed classes and segments appears below:

- **IZ** (stands for a noun phrase, NP): “el próximo viernes”, “un billete de ida y vuelta”, “el de las once”, ...
- **LO-que** (stands for any phrase ending with the word “que”): “el que”, “los que”, “un euromed que”, “dígame los que”, ...
- **PR-despues** (stands for a prepositional phrase, PP, beginning with the word “después”): “después de las dos”, “después de las quince”, ...

[1] Ametzagaiña R&D group, member of the Basque Technologic Network, http://www.ametza.com

4 Task and Corpus

The experiments were carried out over a task-oriented corpus that consists of human-machine dialogues in Spanish, DIHANA (acquired with a Consortium of Spanish Universities) [17]. In this corpus, 225 speakers ask by telephone for information about long-distance train timetables, fares, destinations and services.

Table 1. Features of the corpus

		DIHANA
Training	Sentences	8,606
	Different sent.	5,590
	Words	77,476
	Vocabulary	865
Test	Sentences	1,348
	Words	12,365
	Vocabulary	503
	OOV	72
	PP ($k = 3$)	14.59

Table 2. Different sets of classes and segments

	linguistic	**statistical**				
$\|C\|$	57	50	100	200	300	400
$\|\Sigma\|$	3,851	1,289				
total no. cat.	55,053	57,078				
total no. seg.	55,053	57,078				

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. Therefore, it is well-suited to study the improvements associated to modifications in the LM. The features of the corpus are detailed in Table 1.

As already mentioned in Section 3, different sets of classes were obtained using two different classification criteria: a linguistic criterion and a statistical one. Furthermore, two different sets of segments were obtained, also using two different criteria and the techniques described in Section 3. Table 2 shows the statistics of the resulting groups of classes and segments, as well as the total number of classes and segments that are in the training corpus once it has been segmented or classified.

5 Experiments and Results

The LMs proposed in this work were fed into an CSR system, which was subsequently evaluated in terms of WER. The CSR 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 classes, segments, words and acoustic models (The acoustic models are continuous Hidden Markov Models).

Three series of experiments were carried out in order to evaluate the proposed approaches in Section 2.

Firstly, the **LM based on classes consisting of word sequences, $\mathbf{M_{sw}}$**, was fed into the CSR system, according to Equation 7. Making use of this LM, different experiments were carried out, choosing for all of them a value of $k_c = 3$ and $k_w = 2$. First of all, the set of linguistic classes was employed. Then, five experiments were carried out using 50, 100, 200, 300 and 400 statistical classes respectively.

On the other hand, **LMs based on classes consisting of linked words, $\mathbf{M_{sl}}$**, were integrated into the CSR system according to Equation 11. A value of $k_c = 3$ was established. Experiments were carried out using the same sets of linguistic and statistical classes described above. The same sets of segments were also employed here.

Finally the **hybrid model** was integrated into the CSR system according to Equation 12. For the M_{sw} model a value of $k_c = 3$ and $k_w = 2$ was established, whereas for the classical word-based model, M_w, a value of $k = 3$ was employed. On the other hand, the λ parameter was selected to obtain the best WER result ($\lambda = 0.1$). The same experiments with the same mentioned sets of classes and segments were repeated with this model.

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

First of all, looking at the results in Table 3 it can be concluded that statistical classes yield better results than linguistic ones, even when the number of classes is similar (50 statistical classes vs. 57 linguistic classes).

The results obtained in Table 3 were also compared with the values of WER obtained in another work [18], over the same task and using a classical class-based model with classes made up of isolated words. As shown in the mentioned work, 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 classes (in both M_{sw} and M_{sl} models), than when employing classical class-based LMs using classes made up of isolated words.

Regarding the results obtained with the M_{sw} model, when 50 classes were used the results improve by 7%, whereas for 100 classes the corresponding improvement equals 4.5%. Nevertheless, using a word based LM (M_w), WER values are lower

Table 3. WER results for a classical word based LM (M_w) and for the proposed LMs (M_{sw}, M_{sl} and M_h) using different sets of classes: 57 linguistic classes and 50, 100, 200, 300 and 400 statistical classes respectively

WER (%)					
no. cat.		M_{sl}	M_{sw}	M_h	M_w
ling.	**57**	22.78	25.97	20.04	19.84
statis.	**50**	20.96	22.52	19.23	
	100	19.83	21.21	18.84	
	200	19.42	20.79	18.14	
	300	19.27	20.66	18.22	
	400	19.63	21.38	18.52	

than those obtained for the M_{sw} model and the selected sets of classes and segments. This could be due to some strong assumptions made in the definition of the model.

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 classes. The best result is obtained for 300 statistical classes, achieving an improvement of a 6.7% with respect to the value obtained in the same conditions for the M_{sw} model. Furthermore, the result obtained with 300 statistical classes and an M_{sl} model improves the WER values obtained with the word based LM (M_w) by a 2.8%.

However, the use of a hybrid model, interpolating the M_{sw} and the M_{sl} models, outperforms the results obtained with all the previous proposals. Moreover, the best result is obtained for 200 statistical classes where an improvement of a 8.56% is observed with respect to the word-based LM.

6 Concluding Remarks and Future Work

In this work, we propose and formulate two different approaches to language models, which are based on classes made up of segments of words. On the other hand, an interpolated LM was explored as well. The proposed models were integrated into a CSR system in order to evaluate them in terms of WER. The experiments carried out show that using a LM based on classes consisting of segments of words instead of a classical class n-gram (or k-TSS) LM, a better performance of a CSR system can be achieved. On the other hand, although some of the results attained with the class-based models in this work, outdo those obtained with a classical word-based LM, the observed improvement is not very significant. Therefore, the interpolation of a word-based LM and a LM based on classes made up of segments of words was employed. Using such a model a better performance of a CSR system can be achieved compared to a word-based LM.

However, since the M_{sl} model provides better results than the M_{sw} one, it could be interesting, for further work, to explore the interpolation of the M_{sl} model and a LM based on the same words or lexical units that M_{sl} employs.

Acknowledgments. We would like to thank the Ametzagaina group and Josu Landa, in particular, for providing us with the linguistic classification and segmentation of the corpus.

This work has been partially supported by the University of the Basque Country under grant 9/UPV00224.310-15900/2004 and by CICYT under grant TIN2005-08660-C04-03.

References

1. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge (1998)
2. Gupta, V., Lenning, M., Mermelstein, P.: A language model for very large-vocabulary speech recognition. Computer Speech and Language 6(2), 331–344 (1992)

3. Benedí, J.M., Sánchez, J.A.: Estimation of stochastic context-free grammars and their use as language models. Computer Speech and Language 19(3), 249–274 (2005)
4. García, P., Vidal, E.: Inference of k-testable languages in the strict sense and application to syntactic pattern recognition. IEEE Trans. Pattern Anal. Mach. Intell. 12(9), 920–925 (1990)
5. Torres, I., Varona, A.: k-tss language models in speech recognition systems. Computer Speech and Language 15(2), 127–149 (2001)
6. Brown, P.F., Pietra, V.J.D., Souza, P.V.d., Lai, J.C., Mercer, R.L.: Class–based n-gram Models of Natural Language. Computational Linguistics 18(4), 467–480 (1992)
7. Niesler, T.R., Woodland, P.C.: A variable-length category-based n-gram language model. In: IEEE ICASSP 1996, Atlanta, GA, vol. I, pp. 164–167. IEEE, Los Alamitos (1996)
8. Niesler, T., Whittaker, E., Woodland, P.: Comparison of part-of-speech and automatically derived category-based language models for speech recognition. In: ICASSP 1998, Seattle, pp. 177–180 (1998)
9. Zitouni, I.: Backoff hierarchical class n-gram language models: effectiveness to model unseen events in speech recognition. Computer Speech and Language 21(1), 99–104 (2007)
10. Deligne, S., Bimbot, F.: Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. In: Proc. ICASSP 1995, Detroit, MI, pp. 169–172 (1995)
11. Ries, K., Buo, F.D., Waibel, A.: Class phrase models for language modelling. In: Proc. ICSLP 1996, Philadelphia, PA, vol. 1, pp. 398–401 (1996)
12. Kuo, H.K.J., Reichl, W.: Phrase-based language models for speech recognition. In: Proceedings of EUROSPEECH 99, Budapest, Hungary, vol. 4, pp. 1595–1598 (September 1999)
13. Marcu, D., Wong, W.: A phrase-based, joint probability model for statistical machine translation (EMNLP), Philadelphia, PA (July 6-7, 2002)
14. Zens, R., Ney, H.: Improvements in phrase-based statistical machine translation. In: Proc. of the Human Language Technology Conf (HLT-NAACL), pp. 257–264 (2004)
15. Caseiro, D., Trancoso, I.: Transducer composition for on-the-fly lexicon and language model integration. In: Proceedings ASRU 2001 - IEEE Automatic Speech Recognition and Understanding Workshop, Madonna di Campiglio, Italy (December 2001)
16. Och, F.J.: An efficient method for determining bilingual word classes. In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics, pp. 71–76 (1999)
17. Benedí, J., Lleida, E., Varona, A., Castro, M., Galiano, I., Justo, R., López, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: Proc. of LREC 2006, Genoa, Italy (May 2006)
18. Justo, R., Torres, M.I., Benedí, J.M.: Category-based language model in a spanish spoken dialogue system. Procesamiento del Lenguaje Natural 37(1), 19–24 (2006)