Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 197-205 recibido 04-05-2011 aceptado 23-05-2011

A Part-of-Speech Tag Clustering for a Word Prediction System in Portuguese Language*

Agrupamiento de Categorías para un Sistema de Predicción de Palabras en Portugués

Daniel Cruz Cavalieri
Teodiano Freire Bastos Filho
Mário Sarcinelli Filho
Universidade Federal do
Espírito Santo
Av. Fernando Ferrari, 514
Campus Universitario, Vitória
Espírito Santo, Brasil
{daniel,tfbastos,sarcinel}@ele.ufes.br

Sira Elena Palazuelos Cagigas
Javier Macías Guarasa
José L. Martín Sánchez
Universidad de Alcalá
Ctra. Madrid-Barcelona, Km. 33,600, s/n.
Campus Universitario, Alcalá de Henares
Madrid, España
{sira,macias,jlmartin}@depeca.uah.es

Resumen: Este trabajo presenta un método automático para reducir el conjunto de categorías de palabras que será utilizado por un sistema de predicción de palabras en Portugués. El método se basa en una medida de similitud que se aplica a una matriz de asociación, generada mediante el empleo de una medida de disparidad (odds ratio) aplicada sobre la matriz de distribución de probabilidades de bigramas de categorías (*bipos*) presentes en un corpus. Los resultados presentados en este trabajo muestran que la utilización del método de agrupamiento propuesto, con un umbral adecuado de similitud, tiene potencial para mejorar el sistema de predicción de palabras. Además posibilita la utilización de nuevas técnicas de agrupamiento de categorías como agrupamiento borroso. Los resultados también muestran que cuando se utiliza un sistema de predicción de palabras basado en un modelo sintáctico, la agrupación no se puede realizar entre las categorías sintácticas más importantes, aunque los grupos generados parezcan correctos desde el punto de vista lingüístisco.
Palabras clave: Agrupamiento de categorías de palabras, sistema de predicción de palabras, modelo del espacio vectorial, optimización, portugués.

Abstract: This paper presents an automatic method for reducing the part-of-speech tagset to be considered by a word prediction system in Portuguese. The method is based on a similarity measure applied to a association matrix, generated by employing a odds ratio association measure in the bigrams of parts-of-speech (*bipos*) probability distribution in a corpus. The results reported in this paper show that using the proposed clustering method with an appropriate threshold value over the similarity has the potential to improve the word prediction system. Moreover, it makes possible to use new clustering techniques such as fuzzy clustering. The results also show that when using a word prediction system based on a syntactic model, the clustering cannot be performed between the major syntactic categories, even if the clusters generated seem correct from a linguistic point of view.
Keywords: Part-of-speech clustering, word prediction system, vector space model, optimization, Portuguese language.

1 *Introduction*

Formerly, word prediction methods have been developed in order to increase message composition rate for people with severe motor and speech disabilities. Nowadays, text prediction methods, if adequately integrated within the user interface of an application, can benefit anyone trying to produce text messages or commands. In general, prediction refers to those systems that guess which letters, words, or phrases are likely to follow

* This work was supported by the Spanish Ministry of Science and Innovation under projects VISNU (Ref. TIN2009-08984) and SD-TEAM (Ref. TIN2008-06856-C05-05) and by CAM-UAH under FUVA project (CCG10-UAH/TIC-5988).

ISSN 1135-5948

in a given segment of a text (Ghayoomi y Momtazi, 2009). In all the cases, the main goal of all writing assistance systems is increasing the KeyStroke Saving (KSS), which is the percentage of keystrokes that the user saves by using word prediction systems, besides ensuring a good quality in the text produced.

There are several word prediction systems that have been developed and are being developed with different methods for different languages (Palazuelos-Cagigas, 2001; Ghayoomi y Momtazi, 2009; Cavalieri et al., 2010). Traditionally, word predictors have been based on n-gram statistical language modeling. A major drawback of this kind of predictors is that they do not consider the syntactic and semantic structure of the sentence, and, therefore, there is a possibility of predicting words which are syntactically and/or semantically inappropriate to write the desired sentence. In contrast, using Part-of-Speech (POS) tags of words in prediction systems makes then more appropriate. In other words, the aim of grammatical prediction is to reduce the envelope of search used by more conventional methods for providing predictions. This envelope should only contain words syntactically correct given the current sentence structure.

In order to build a system that uses grammatical knowledge to derive predictions, the manner in which the grammar itself is to be represented must first be considered. Generally, the class set of language models based on POS tags have usually been defined by linguistics experts, according to linguistics aspects. Language models based on a small number of POS tags show higher perplexity, while language models based on more detailed linguistic classes present a lower perplexity (Bahrani et al., 2008; Momtazi y Sameti, 2009). However, using all the lexicon features available results in a large number of classes and in a great amount of text material and memory needed to train and use the system. Thus, it is necessary to reduce the number of classes trying to balance the relationship between number of POS tags, the perplexity of the language model and the keystroke saving.

Most word prediction systems are mainly focused on non-inflected languages, like English. These types of languages have a small amount of word variations and it is possible to include all of them in the dictionary used in the word prediction (Garay-Vitoria y Gonzalez-Abascal, 1997). Since our focus in this work is the Portuguese language, which has a reasonable amount of inflections, it may be difficult to store all of them.

Within this context, an initial POS tagset was first derived from linguistics guidelines. For this task, a lexicon extracted from the Portuguese corpus CHAVE (Santos y Rocha, 2004), tagged by the morphological analyzer PALAVRAS developed by (Bick, 2000), was used. Then, an automatic clustering method based on the similarity between the POS tags was used. This similarity can be interpreted as a *distance measure* applied over an *association matrix*, generated from the application of an *association measure* in a *co-occurrence matrix* composed by the frequencies of the POS tags bigrams in a corpus (see section 3 for details about each matrix, the variables used, and the procedure). When POS tags with very similar frequency distributions are combined, there are very small changes in the results, and advantages such as smaller training time and resources needed, for example.

The POS-based model language and its features are introduced in section 2. Section 3 deals with the clustering approach proposed in this work. The experimental results and discussions can be found in section 4. Finally, section 5 contains the concluding remarks and section 6 the future works.

2 *Language Model*

2.1 POS-based Language Model

The general task of a word prediction system is to estimate the a priori probability $P(\mathbf{W})$ for a given word chain $\mathbf{W} = w_1 \dots w_k$. For the word bigram model, for example, $P(\mathbf{W})$ is approximated by:

$$P(w_1 \dots w_k) \approx \prod_{i=1}^{k} P(w_i|w_{i-1}). \qquad (1)$$

By modeling the language with POS tags, the system predicts the next POS tag to be produced in the current sentence and narrows the amount of possible next words when each letter of the word is entered. In other words, a syntactic predictor has access to the following sequence of words and POS tags to predict the current word:

$$\cdots\ w_{i-2}/t_{w_{i-2}}\ w_{i-1}/t_{w_{i-1}}\ cw_i,$$

where $t_{w_{i-2}}$ and $t_{w_{i-1}}$ are the POS tags of the previous words w_{i-2} and w_{i-1}, respectively. cw_i is the current word prefix typed by the user.

There are different methods for incorporating the statistical POS tag information into the word predictor (Fazly, 2002). In this work the syntactic predictor was estimated by the equation 2.1, as follows:

$$\begin{aligned} P(W) \approx \quad & \textstyle\prod_{i=1}^{k} \sum_{t_i^j \in T(w_i)} P(t_{w_i}^j | t_{w_{i-1}}) \cdot \\ & \cdot P(w_i | t_{w_i}^j) \cdot P(t_{w_{i-1}} | w_{i-1}), \end{aligned} \qquad (2)$$

where $t_{w_i}^j$ is the jth tag for w_i, that varies from 1 to $|T(w_i)|$. $T(w_i)$ is the set of all possible POS tags that may be assigned to the word w_i. $P(t_{w_i}^j|t_{w_{i-1}})$ is the bigram POS tag probability and $P(w_i|t_{w_i}^j)$ is the conditional probability of the word w_i to be the current word of the sentence given $t_{w_i}^j$ as its POS tag. $P(t_{w_{i-1}}|w_{i-1})$ is the conditional probability of the word w_{i-1} to be tagged with the tag $t_{w_{i-1}}$ (this is obtained from a general dictionary: see section 4.4 for details).

3 Clustering Approach

As mentioned in (Wood, 1996), the majority of linguistic theories examine language from an analytic point of view. That is, they wish to understand all levels of meaning, from contextual concepts through semantic and syntactic structure to phonetic relations. However, when looking at prediction this is not necessarily the case. One of the aims of syntactic prediction is to ensure that the system does not offer to the user grammatically incorrect words. However, a required feature in a word prediction system, especially when used by people with disabilities, is that it must be as fast as possible. Thus, the aim of this research is to find a reduced POS tagset trying to maintain a good relationship between the accuracy of the prediction system and the speed at which it generates the predicted word list to the user. Furthermore, the less tags the POS tagset has, the better the word prediction algorithms parameters are estimated and less severe is the sparse data problem.

In order to obtain a reduced POS tagset we have not followed the strategy already used by (Sánchez-Martínez, Pérez-Ortiz, y Forcada, 2005) for machine translation. Instead, we have tried to reduce the number of Portuguese POS tags clustering the ones that lie close together in a vector space model. Commonly, this kind of strategy is used in word clustering and in speech recognition tasks (Velldal, 2003). One advantage of this clustering algorithm is that the number of clusters (reduced POS tags) to discover is automatically determined by providing the algorithm with a distance threshold. We first select, based on linguistic guidelines, the fine POS tags and, in each step, those POS tags that are closer are merged into a single cluster, only if their distance is larger than the specified threshold.

There is an important difference concerning the POS tags we use within this paper and the ones used in works like (Yarowsky, 1992; Lin, 1997; Resnik, 1999; Velldal, 2003). The main difference is the reduction of the size of the vector space model. While all of them work with word context features (word classes, word bigrams, etc.), our vector space model is based on the POS bigram probability distribution in a corpus. In this case, the basic idea is similar to that proposed by (Sánchez-Martínez, Pérez-Ortiz, y Forcada, 2005) who exploits the fact that some of the POS tags have very similar frequency distribution and, if we combine the most similar ones, there should be only a small change in the results.

3.1 The Vector Space Model

In (Velldal, 2003), the space model was formally defined as a triple $\{\mathbf{F}, \Theta, s\}$, corresponding to a *co-occurrence matrix*, an *association measure* and a *similarity function*, respectively. $\mathbf{F}$, the *co-occurrence matrix*, is the bipos matrix: the value of each element in the matrix $\mathbf{F}$ is given by the number of times that the predicted POS tag occurs given the previous POS tag. In others words, $\mathbf{F}$ is defined by the vectors $\{\mathbf{f}_{m1}, \ldots, \mathbf{f}_{mn}\}$, where the m is the index of the predicted POS tag and the nth coordinate corresponds to the previous POS tag. The association measure Θ is a weighting function that maps each element $f_{mn} \in \mathbf{F}$ to a salience score, where the cluster analysis will be performed using the proximity function s.

3.1.1 The Association Matrix and Association Measure

As can be seen in (Velldal, 2003), the raw frequencies alone may not always be very informative. For this reason, an *association measure* Θ was also applied to each component of the co-occurrence matrix $\mathbf{F}$. In this case, each component $f_{mn} \in \mathbf{F}$ will have its own 2-dimensional cross-classification table, as illustrated in Table 1. This table of observed cell frequencies is also known as a *contingency table*.

The association measure Θ is based on the *odds ratio*, estimated applying the equation 3 on the values of the table 1. The odds ratio is a way of comparing whether the probability of a certain event is the same for two groups. An odds ratio of 1 implies that the event is equally likely in both groups. An odds ratio greater than one implies that the event is more likely in the first group. An odds ratio less than one implies that the event is less likely in the first group. Thus, the result is called the *association matrix* and was defined as $\mathbf{X} = \Theta(\mathbf{F})$. In the case of unobserved or negatively correlated co-occurrence pairs in $\mathbf{X}$, the elements are assumed to have zero association.

	$= t_{w_{i-1}}$	$\neq t_{w_{i-1}}$
$= t_{w_i}$	$f(t_{w_i}, t_{w_{i-1}})$	$\sum f(t_{w_i}, \neg t_{w_{i-1}})$
$\neq t_{w_i}$	$\sum f(\neg t_{w_i}, t_{w_{i-1}})$	$\sum f(\neg t_{w_i}, \neg t_{w_{i-1}})$

Table 1: Contingency table of observed frequencies.

$$\Theta = \frac{f(t_{w_i}, t_{w_{i-1}}) \cdot \sum f(t_{w_i}, \neg t_{w_{i-1}})}{\sum f(\neg t_{w_i}, t_{w_{i-1}}) \cdot \sum f(\neg t_{w_i}, \neg t_{w_{i-1}})}. \quad (3)$$

Where, for example, $f(t_{w_i}, \neg t_{w_{i-1}})$ refers to the frequency of the predicted POS tag t_{w_i} not involving the previous POS tag t_{w_i}, and so forth.

3.1.2 The Proximity Matrix

The notion of proximity can be seen as a relation of the *distance* between each mth vector of the vector space model. A commonly used measure of similarity is the cosine of the angle between two vectors, defined as

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}, \quad (4)$$

with a value ranging from *zero* for orthogonal vectors, to *one* for vectors that point in the same direction.

Thus, as the similarity function s is specified the *proximity matrix* $\mathbf{S} = m \times m$ can be constructed. In other words, a component $s_{ij} \in \mathbf{S}_{ij}$ represents the similarity between the vectors $\mathbf{x}_i$ and $\mathbf{x}_j \in \mathbf{X}$.

3.2 The New POS Tagset

Given the matrix $\mathbf{S}$, the similarity between POS tags vectors can be ranked according to their similarity scores. Table 2 gives a example of such a list for the POS tag (noun, masculine and singular).

Rank	POS tag	Similarity
1	(noun, neuter, singular)	0.8752
2	(adjective, masculine, singular)	0.5954
3	(indefinite pronoun, masculine, singular)	0.5607
4	(ordinal)	0.4731
5	(possessive pronoun, masculine, singular)	0.4466

Table 2: The 5 most similar POS tags of the (noun, masculine and singular) tag.

The display of such similarity rankings can be useful and interesting in their own right, as it summarizes the most common and distinguishing usage patterns of a POS tag at a quick glance. After computing the above scores, we can automatically reduce the number of POS tags.

4 Evaluation

It is difficult to evaluate a word prediction system. In particular, since the user objectives and problems may vary, a specific metric may be more appropriate than another one to evaluate the advantages the prediction may produce. It is also necessary to consider that a change in any parameter of the experiment setup (the language, training texts, etc.) can lead to quite significant variations in the results.

4.1 The Initial Portuguese POS Tagset

Before applying the proposed clustering algorithm, an initial Portuguese POS tagset was generated by selecting the most functional Portuguese POS tags delivered by the morphological analyzer PALAVRAS. As showed in (Culleto, 2007; Aliprandi et al., 2007), the use of some major part-of-speech of words (noun, verb, adjective, etc,) along with some inflections like gender (masculine, feminine, neuter), number (singular, plural, neuter) and person (1th, 2nd, 3rd, 1th/3rd) can generate accurately POS-based word predictors with a relatively low speed list of predicted words. However, this causes the number of initial tags to be relatively large: 81 fine POS tags in our Portuguese lexicon. As mentioned in (Brants, 1995) It is important to notice that the larger the tagset the worse the data sparseness problem.

4.2 Experimental Setup

Firstly, to generate the vector space model (**F**) a training set consisting of approximately 2,000,000 words was used. They were extracted from newspapers and texts documents in the Portuguese tagged corpus CHAVE, which contains about 26 millions of words. The same corpus was used to train the different prediction methods. In this case, the training set was composed by approximately 1,200,000 words, from texts that do not overlap with the ones used to generate the vector space model. Finally, in order to evaluate the reduced POS-based word prediction system, we took four test texts (also distinct from both previous training sets) detailed in Table 3.

Topic	Domains	#words	#keystroke needed
a	Belief and thoughts	10714	66005
b	World news	12775	79811
c	Politics	13473	84189
d	Commerce and finance	13324	81228

Table 3: Domains, number of words and keystroke needed to write each text in test set without the help of the word prediction.

As the difference in the number of POS tagset generated by each method, each dictionary (or vocabulary) used to test the word prediction system is also different. However, this difference is significantly small when compared to the size of each dictionary (about 140,000 entries, each composed of a word, its frequency and POS tag). Moreover, since we are working on a method to reduce the number of tags in a POS tagset, we can consider the possible reduction (or not) in the size of the dictionary as a factor more in the comparison of the systems.

4.3 Performance Measures

The POS-based word prediction system was evaluated according three different criteria: Perplexity (PP), Keystroke Saving (KSS) and Hit Rate (HR). The KSS is referred to the percentage of keystrokes that the user saves by using the word prediction system and is calculated by comparing two kinds of measures: the total number of keystrokes needed to type the text (K_T) without the help of the word prediction and the effective number of keystrokes saved using word prediction (K_E). Hence,

$$KSS = \frac{K_T - K_E}{K_T} \times 100. \qquad (5)$$

A higher value for keystroke saving implies a better performance.

The HR is defined as the percentage of correct words that appear in the suggestion list without entering any letter of the following word. In others words, it is the relation between the number of times that a word is guessed and the number of written words. Again, a higher HR also implies a better performance.

The PP can be usually defined as the average number of choices at each word prediction (Trnka et al., 2006). So if the PP is low i.e. the probability of the prediction is high, meaning a better language model. The PP of the bipos and tripos models can be computed, respectively, by the equations 6 and 7, as follows:

$$PP_{2gram} = 2^{-\frac{1}{m}\sum_{i=1}^{m} log_2(P(w_i|w_{i-1}))} \qquad (6)$$

$$PP_{3gram} = 2^{-\frac{1}{m}\sum_{i=1}^{m} log_2(P(w_i|w_{i-1}w_{i-2}))}, \qquad (7)$$

where $P(w_i|w_{i-1})$ and $P(w_i|w_{i-1}w_{i-2})$ are computed by the equation 2.1.

4.4 Word Prediction Engine

To evaluate our reduced POS-based method, the software PREDWIN, developed by (Palazuelos-Cagigas, 2001) for Spanish language and adapted to Portuguese language by (Cavalieri et al., 2010), was used. This system utilizes different kinds of algorithms to realize the word prediction. In this paper was used only the basic statistical models *bi-POS* and *triPOS* in order to find the 5 best predicted words.

The Figure 1 shows the general architecture of the system followed by a briefly description of the main blocks. A further detail of the system can be founded in (Palazuelos-Cagigas, 2001).

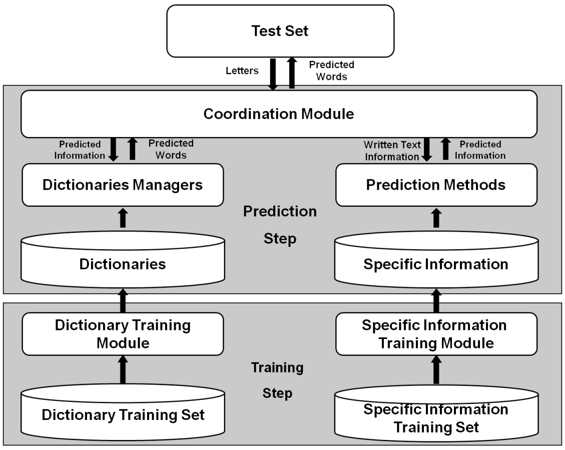

Figure 1: General Architecture of the word prediction system PREDWIN. Adapted from (Palazuelos-Cagigas, 2001).

- **Test Set:** The texts used to evaluate the reduced POS-based word prediction methods used in this work.
- **User Model:** The automatic algorithm used by the word prediction system to emulate a real user. For each letter in the test text the prediction system shows a list of predicted words. If the desired word was in this list, the user model selects the word. If not, the prediction system goes to the next letter until the test text ends.
- **Dictionary:** Contains the words and the information required to support each of the word prediction methods, as the POS tags and word frequencies.
- **Prediction Methods:** Algorithms responsible of calculating, based on the information provided by the coordination module, the probability set of categories that can be assigned to the next word.
- **Specific Information Training Module:** Defined as the procedures, automatic or manual, needed to generate the information used by each prediction method. The automatic reduced POS tag method presented in this work is an example of such procedure.

4.5 Results

In order to find the best reduced Portuguese POS tagset we have performed a clustering on a vector space model composed by the bigram probabilities of the POS tag in a specific corpus. Once the clustering has been performed the reduced POS tagset was used in a POS-based word prediction system. Note that the final number of tags is indirectly determined because the clustering algorithm is provided with a distance threshold.

The POS-based word prediction system was evaluated in the test set by employing three methods of POS tag construction:

Baseline Method The first POS tagset was manually defined following linguistics guidelines and generated by the PALAVRAS tagger. In this case we have achieved a baseline POS tagset with 81 tags.

Method A The second method generates two POS tagsets derived from the baseline POS tagset. Again based on linguistic characteristics, two POS tagsets were manually obtained with 63 and 42 tags, respectively. In these cases the major syntactic categories of words (noun, adjective, pronoun, etc) were kept, collapsing some of the word inflections (gender, number, person) and some punctuation marks.

Method B In the third method, the reduced POS tagset was obtained automatically based on the similarity coefficient s applied to the association matrix $\mathbf{X}$, generated from the application of the association measure Θ in each element of the co-occurrence matrix $\mathbf{F}$, as explained earlier.

In order to evaluate the POS tag clustering method proposed, two thresholds values were chosen: 0.7 and 0.85. As mentioned earlier, these values are interpreted as the similarity between two POS tags. When the threshold value of 0.7 was used, the number of tags (clusters) was 54. Otherwise, when

Methods	#POS tags	Topic	biPOS PP	biPOS KSS(%)	biPOS HR(%)	triPOS PP	triPOS KSS(%)	triPOS HR(%)
Method A	63	a	90.41	38.66	28.90	83.23	38.80	29.20
		b	67.59	35.50	27.30	58.03	35.56	27.20
		c	69.90	36.64	28.10	64.29	36.60	28.10
		d	67.02	36.49	31.20	62.44	36.61	30.90
	42	a	91.90	38.33	28.80	83.34	38.63	29.00
		b	68.51	34.81	27.10	59.14	35.00	27.10
		c	72.25	36.24	27.60	64.62	36.37	27.70
		d	68.42	35.97	30.60	62.72	36.25	30.80
Method B	72	a	94.04	38.10	28.60	86.14	38.25	28.90
		b	73.76	35.30	27.40	64.83	35.44	27.40
		c	76.67	36.17	28.40	70.50	36.15	28.30
		d	73.39	36.51	30.60	68.62	36.63	30.90
	54	a	104.78	37.16	28.00	96.97	37.36	28.40
		b	80.43	34.96	26.80	70.89	35.13	27.10
		c	91.54	35.54	27.50	83.81	35.89	27.60
		d	81.58	35.36	30.10	75.51	35.56	30.20

Table 4: The results generated by incorporating different POS tagsets in the word prediction system.

the threshold value of 0.85 was selected, 72 clusters were produced.

As can be seen in Table 4, when comparing the KSS of each method the results achieved by the POS-based word prediction system with the reduced POS tagsets are not better than the others. However, when comparing the HR values the 72-tag tagset shows better results in almost all the test texts, which can be explained by the fact that, in this case, the word prediction system works well to words like prepositions and articles. Furthermore, when compared to the 54-tag tagset, the 72-tag system shows better results, which can be attributed to the fact that in this case there are only clustering in the same major syntactic category (for example, feminine plural noun and neuter plural noun), as we expected.

It can also be seen in Table 4 that the perplexity of the language models increases with the decreasing in the number of previous POS tags, which is consistent with previous works. Furthermore, the better is the perplexity, the better is the KSS. This may seem a bit obvious, but since most of the previous research was carried out for machine translation or categorization of words, it seems interesting to evaluate the effects of the language model perplexity in a POS-based word prediction system.

Table 4 also shows the data sparseness problem when tested the text in topic *c*. As can be seen, in both 72-tag and 63-tag tagsets, when using biPOS the word prediction system shows better results in the KSS than when using triPOS.

It is also important to note that in the experiments reported in this paper we have used a smoothing technique to avoid null transition and emission probabilities for those unseen events in the training corpus.

5 Conclusions

We have shown a method for reducing an original POS tagset in order to optimize a POS-based word prediction system. The results reported in this paper show that using the automatic clustering method proposed with an appropriate threshold value of similarity can slightly increase the POS-based prediction system, but, when compared to the same system with a manually reduced POS tagset the results were worst.

We also see that, when working with a language model to predict words, it is necessary to pay a special attention in the construction of the POS tagset, taking in account the relationship between categories of words. In others words, even if the clustering seems intuitively correct, as in this work, it can not be performed between the major syntactic categories of words. However, this kind of clustering can work for another kind of language application like language translation or categorization of words.

6 Future Works

In this work, the clustering method uses a distance between clusters. We have used the cosine distance to measure the similarity between fine POS tags, but other distance measures could also be suitable. Furthermore, only the association measure based on odds ratio was used. We can also use the *log likelihood ratio* and the *mutual information* as association measures, as proposed by (Velldal, 2003).

In this paper only the bipos probabilities were used to calculate the similarity between POS tags. Tripos probabilities associated with a variable latent method could also be used to ensure a finest measure in the similarity.

Besides, only two values of distance thresholds were ised and, as can be seen in (Sánchez-Martínez, Pérez-Ortiz, y Forcada, 2005), the optimal threshold value can be found varying the threshold with small values (0.025 in his case).

References

Aliprandi, Carlo, Nicola Carmignani, Paolo Mancarella, y Michele Rubino. 2007. A word predictor for inflected languages: system design and user-centric interface. En *Proceedings of the Second IASTED International Conference on Human Computer Interaction*, IASTED-HCI '07, páginas 148–153, Anaheim, CA, USA. ACTA Press.

Bahrani, Mohammad, Hossein Sameti, Nazila Hafezi, y Saeedeh Momtazi. 2008. A new word clustering method for building n-gram language models in continuous speech recognition systems. En *Proceedings of the 21st international conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems: New Frontiers in Applied Artificial Intelligence*, IEA/AIE '08, páginas 286–293, Berlin, Heidelberg. Springer-Verlag.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. tesis, Aarhus University, Aarhus, Denmark, November.

Brants, Thorsten. 1995. Tagset reduction without information loss. En *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, páginas 287–289, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cavalieri, Daniel C., Sira E. Palazuelos-Cagigas, Teodiano F. Bastos-Filho, y Mário Sarcinelli-Filho. 2010. Evaluation of machine learning approaches to portuguese part-of-speech prediction. En António Teixeira Vera Lúcia Strube de Lima Luís Caldas de Oliveira, y Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 9th International Conference, Proceedings (PROPOR 2010)*, Porto Alegre, Brasil, 27-30 de Abril.

Culleto, Thomas. 2007. Prediction of liaison in french by measures of information theory. En Oxford, editor, *Proceedings of LingO*, páginas 59–67. Oxford University.

Fazly, Afsaneh. 2002. The use of syntax in word completion utilities. Master's thesis, University of Toronto, Department of Computer Science.

Garay-Vitoria, N. y J. Gonzalez-Abascal. 1997. Intelligent word prediction to enhance text input rate (a syntactic analysis based word prediction aid for people with severe motor speech disability). En *Annual International Conference on Intelligent User Interfaces*, páginas 241-247.

Ghayoomi, M. y S. Momtazi. 2009. An overview on the existing language models for prediction systems as writing assistant tools. En *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, páginas 5083–5087, San Antonio, Texas, 11-14 October. ISSN: 1062-922X.

Lin, Dekang. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. En *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, EACL '97, páginas 64–71, Stroudsburg, PA, USA. Association for Computational Linguistics.

Momtazi, Saeedeh y Hossein Sameti. 2009. A possibilistic approach for building statistical language models. En *Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications*, ISDA '09, páginas 1014–1018, Washington, DC, USA. IEEE Computer Society.

Palazuelos-Cagigas, S. E. 2001. *Contribution to word prediction in Spanish and its integration in technical aids for people with physical disabilities.* Ph.D. tesis, Universidad de Alcalá de Henares, Alcalá de Henares, Madrid, Spain.

Resnik, Philip. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

Sánchez-Martínez, Felipe, Juan Antonio Pérez-Ortiz, y Mikel L. Forcada. 2005. Target-language-driven agglomerative part-of-speech tag clustering for machine translation. En *Proceedings of the International Conference RANLP - 2005 (Recent Advances in Natural Language Processing)*, páginas 471–477, September.

Santos, Diana y Paulo Rocha. 2004. The key to the first clef in portuguese: Topics, questions and answers in chave. En *5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, páginas 821-832, Bath, UK, September 15-17.

Trnka, Keith, Debra Yarrington, Kathleen McCoy, y Christopher Pennington. 2006. Topic modeling in fringe word prediction for aac. En *Proceedings of the 11th international conference on Intelligent user interfaces*, IUI '06, páginas 276–278, New York, NY, USA. ACM.

Velldal, Erik. 2003. Modeling word senses with fuzzy clustering. Cand.philol. thesis, University of Oslo.

Wood, Matthew E. J. 1996. *Syntactic Pre-Processing in Single-Word Prediction for Disabled People.* Ph.D. tesis, Department of Computer Science, University of Bristol, June.

Yarowsky, David. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. En *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, páginas 454–460, Stroudsburg, PA, USA. Association for Computational Linguistics.