An Unsupervised Method for Senses Clustering*

[1]Sofía Paniagua Rivera, [2]Héctor Jiménez-Salazar & [3]David Pinto

Faculty of Computer Science,
B. Autonomous University of Puebla,
14 sur y Av. San Claudio, C.U., Edif. 135
Puebla, Pue., México, C.P. 72570
Tel. (+52-222)2295500 Ext. 7212, Fax (+52-222)2295672
[1]sofiadp@terra.com.ve, {[2]hjimenez, [3]dpinto}@cs.buap.mx

Abstract. The difficulty of obtaining tagged corpora in order to perform Word Sense Disambiguation has led to diverse strategies. Clustering methods may be used as an initial step to discover regularities on instances, i.e. contexts of ambiguous words. In this work we evaluate a sense clustering method with a novel feature selection phase over Senseval-2 Spanish collection. The feature selection technique proposed is based on the sense of a syntagm. Purely unsupervised clustering methods using this feature selection technique shown good accuracy results.

1 Introduction

Word Sense Disambiguation (WSD) is, without a doubt, one of the most important problems in Natural Language Processing (NLP) area. Since the first Senseval contest in 1998, this problem has been treated in a more systematic way [1]. Senseval provides standard collections of information for comparing diverse algorithms of WSD. This problem consists on the recognition of the proper sense for an ambiguous word. For example, let be the word *interest*, in the sentence *an interest in philosophical research*, the goal is to determine for this specific sentence, that *interest* expresses mental excitement, avoiding the possibility of consider it as a finance or commerce sense (charge for borrowing money). The solution of this problem would represent a very important advance in systems that use words or sentences in natural language, for example, search robots for managing digital libraries, and other vast volumes of information like Internet, as Google, Altavista and so on.

Until now, algorithms for WSD have been categorized as supervised or unsupervised. Basically, supervised algorithms for WSD apply a training phase using a set of positive and negative examples (solved instances), in this case, sentences tagged with the correct sense of the ambiguous word. On the other hand, unsupervised algorithms for WSD should not use that kind of extra information. In literature does not exists an agreement about what an unsupervised algorithm for WSD should be; there exist some researchers of NLP that use methods relying (mainly) on knowledge drawn from machine readable dictionaries (MRD)

* This work was partially supported by BUAP-VIEP 3/G/ING/05.

Bhanu Prasad (Editor): IICAI-05, pp. 840-848, 2005.
Copyright © IICAI 2005

and/or raw text (we named this kind of methods as knowledge-based methods). In particular, we apply the term of unsupervised to those methods that do not use external knowledge sources. Certainly, supervised algorithms for WSD show a better performance than unsupervised ones, but for the former is required a training set, restricting its use to an specific domain. On the other hand, the unsupervised algorithms for WSD can be used for solving problems in any domain.

Furthermore, in 1997, H.T. Ng [7] addressed the necessity of focussing the gathering of training examples, in a different way, and therefore may be able to construct algorithms for WSD, since manual tagging of a sufficient big training set for real WSD applications could take more than fifteen years. Advances in WSD have been important in last years. The enrichment of corpora from Internet, in order to accumulate enough amount of examples of contexts that contain ambiguous words, and the use of others resources, like MRD, has led to compile resources to obtain better results (Rada Milhacea reports around a 63% for this kind of methods [10]). Notably, the results in Senseval-2 were about 14 percentage points lower than in Senseval-1 (for the English lexical task), even though the same evaluation methodology was used and many of the systems were improved versions of the same systems that participated in Senseval-1. This can be seen as evidence that WordNet sense distinctions are indeed not well-motivated, but more research is required to confirm this [9].

Nowadays, WSD problem entails high demand to elaborate various lexical resources, not only to construct collections of examples for training phase of the algorithms, but also, the attainment of specialized dictionaries [4] that are useful also for different applications. This fact implies to analyze one of the essential components of this problem: the prospection of the Web, discovering interrelations in that big corpus and reviewing some algorithms in order to improve their performance. Particularly, clustering algorithms provide means for performing this task.

The lackness of training sets and other knowledge sources for specific domains requires to experiment with unsupervised clustering algorithms. In this paper, we present results of a new approach of senses clustering algorithm for ambiguous words over Senseval-2. We propose a novel method for features selection (our main contribution), based on the sense of a syntagm that obtains good results. We do not use any kind of external source of information, and we used Senseval senses of ambiguous words in order to measure the performance of our algorithm.

Feature selection methods may be crucial in the clustering task. Wrong results may be caused due to the methods focus problems in a general fashion. The method used here exploits a formulae on the sense of a syntagm. We hope using sense of a syntagm as a mean of feature selection, sense clustering be enhanced. The underlying hypothesis in this proposal is that sense of a syntagm is built making an especial combination of the sense of words that compound it [8]. Additionally, we work on the clustering algorithm tuning the size of groups, through of repeating the cluster step.

The description of our algorithm and feature selection process are presented in the next two sections, results of our experiments and conclusion of this work are presented later.

2 Clustering algorithms

We have tested a variety of supervised and unsupervised clustering algorithms, like K-Means, K-NN-Modified, Mod-SLC, etc. At the end, we have decided to use Mod-SLC, mainly because this algorithm have shown the best behavior, besides it is a totally unsupervised clustering algorithm.

The Mod-SLC method is based on a proposal of Hassan et al [5] for parts manufacturing. His method is based on concepts presented by Sneath in 1957 [11]. In its pure version, this algorithm calculates a similarity matrix for those contexts that are the clustering goal. The maximum value of similarity is found, consequently, a pair of contexts to be clustered is obtained, and depending of the following circumstances, an action is taken:

- If none of this pair of contexts have already been clustered, then a new cluster composed by such pair is created.
- If this pair of contexts have already been clustered in different groups, then those groups are merged in order to obtain a unique cluster.
- Finally, if only one context of this pair has been clustered, then the other context is added to the same group.

In our case, we are treating with contexts of ambiguous words, and after some experiments done, we observed that some features inside these contexts led Mod-SLC algorithm to generate only one group; which is a real problem for our goal, mainly if we do not use a threshold for the maximum similarity value.

In this way, we construe a modified algorithm of Mod-SLC which is shown as follows.

INPUT: Context collection of a word (CC)
OUTPUT: Clusters of contexts (CL)

Step 0 Construct an incidence matrix of pairs (context, word) using CC.
Step 1 Use the incidence matrix to create a contexts similarity matrix (SM).
Step 2 Calculate the similarity average (SA) in SM.
Step 3 Find the maximum similarity value in SM (SB_{ij}) equal or greater than SA. Consequently a pair of contexts, C_i and C_j, is found. If such value cannot be found, then go to step 6.
Step 4 Add C_i and C_j, to the same group under the following circumstances:
  Step 4.1 If both, C_i and C_j, have not already been assigned, then a new cluster with this pair of contexts is created.
  Step 4.2 If both, C_i and C_j, have already been assigned to different groups, then a merge operation is performed, obtaining only one cluster.
  Step 4.3 If C_i or C_j has not been assigned, then that context is added to the cluster that the other one context belong.

Step 5 Return to step 3.
Step 6 Write out clusters and terminates.

After the execution of clustering algorithm, the feature selection process for each cluster was carried out. In order to obtain the best clusters and features of each cluster, a double execution of the modified Mod-SLC algorithm was applied to the corpus. This means that the Mod-SLC algorithm was executed over an initial set of contexts and the same algorithm was executed over the results obtained by the previous process. This new algorithm was named SLC-SLC.

Features selection procedure for ambiguous words is the main contribution of this work. After the evaluation of various feature selection methods, we selected only two, which are described in the following section.

3 Features selection

Once the clusters were created, we proceed to choose the features that will identify each cluster. Due to the double execution of the SLC algorithm, a double feature selection should be done. The second execution of the clustering algorithm will require as input, a set of clusters conformed by the features selected in the previous execution of SLC algorithm. However, as we will see, a second feature selection process, removes important features from clusters.

We applied two methods: *efficacy solution*, and other method based on the sense of a syntagm, that we named SR. The description of both techniques are presented as follows.

3.1 Efficacy solution

Although, Hassan et al. [5] used the efficacy formula to measure the quality of clusters generated by its algorithm, we applied it to the features selection process. The efficacy formula allows to verify, what cluster a word belongs to. In our case, we understand as a word, one term in the vocabulary of the clusters set, conformed by the SLC-SLC algorithm.

Let be a set of clusters generated by some clustering algorithm, applied over a set of contexts ($Context_1$, $Context_2, \ldots Context_m$) using the words of a vocabulary (see figure 1). Originally, efficacy was proposed to be calculated as follows.

$$Efficacy = \frac{(e - e_0)}{(e + e_1)} \tag{1}$$

where, e is the number of *one's* in the clustering matrix. e_0 is the number of exceptional elements (outside the groups generated), and e_1 is the number of *zero's* inside the clusters. Thus, this formula will obtain a value of 1 when exceptional elements (EE) or zero's (Zs) does not exist inside the groups generated. On the other hand, the efficacy value will decrease to zero, proportionally to the increment of these values (EE and Zs).

	$Word_1$	$Word_2$	$Word_3$	$Word_4$	$Word_5$	$Word_6$	$Word_7$		$Word_n$
$Context_1$	1	1	1	1	0	0	1	...	1
$Context_2$	1	0	1	1	0	0	0	...	0
$Context_3$	0	1	0	1	1	1	0	...	0
$Context_4$	0	0	0	1	0	1	1	...	1
:	:	:	:	:	:	:	:		:
$Context_m$	1	0	1	0	1	0	0	...	0

Cluster 1 Cluster 2 ...

Fig. 1. Example of clustering evaluation

In our case, for each ambiguous word, each term of the vocabulary found in is analyzed, evaluating its efficacy with respect to each cluster conformed. The term is assigned, as a feature, to the cluster that obtains the best efficacy value. Logically, each term of the vocabulary will be assigned strictly to one cluster, which could get worse the quality of the ambiguous words representation [12].

3.2 *SR* solution

This technique obtains features of each cluster, by the union of the intersection of pair contexts. Two options were considered for the feature selection process. The first one was applied after the first execution of SLC for our SLC-SLC algorithm, the second one was applied after every execution of SLC.

Given a set of r contexts $SC = \{ctx_1, ctx_2, ..., ctx_r\}$ of a cluster C_i, the SR solution of C_i is a set of terms belonging to the same cluster that fulfill the following formula:

$$\text{Set of features} = \bigcup_{j \neq k} \left(ctx_j \bigcap ctx_k \right) \tag{2}$$

The idea is based on a proposal to obtain the sense of a syntagm. In [8] is said that the sense of a syntagm is determined by a combination of properties of the words that appear in it. In [6] a formula that uses this idea was proposed. They considered the properties of words under the use of a dictionary, given the following function which associates those properties to some word w.

$$\Upsilon : V \rightarrow V^*, w \mapsto \Upsilon(w) = Prop(w). \tag{3}$$

Denoting $\psi(z_1, z_2) = \Upsilon(z_1) \cap \Upsilon(z_2)$, the sense of the syntagm $w_1, \ldots, w_n, w$ is built iteratively as follows:

$$\xi(w_1, \ldots, w_n, w) = \xi(w_1, \ldots, w_n) \cup \bigcup_{i=1}^{n} \psi(w_i, w), \tag{4}$$

initially $\xi(w_1, w_2) = \psi(w_1, w_2)$.

In our application, we have taken the properties, as the ambiguous word context. This idea is based on the sense relationships that are associated to one word wd (every word related with wd). In this manner we obtain equation 2.

The SR solution improved the results obtained by the *efficacy solution* for feature selection task [12], and therefore it was chosen for final tests.

4 Experimental results

4.1 Dataset

In our experiments, we extracted a set of ambiguous words from a dictionary (named MiniDir) developed by CLIC (`http://clic.fil.ub.es`) specifically for Senseval, obviously to be used by systems of semantic disambiguation. We used a subset of MiniDir, and the words used can be seen in table 1 (each word has an English translation for one of its senses). MiniDir offers a lexical resource with coarse grain of senses, with an average of four senses for nominal entries and adjectives, and six senses for verbs, unlike EuroWordNet where words, present a fine grain, i.e., detail in the word sense. The criteria used in the elaboration of MiniDir.2.1 are listed in [2].

Word	**POS**	**#Senses**	**#Contexts**
corona (*crown*)	Noun	4	297
hermano (*brother*)	Noun	3	593
apoyar (*to support*)	Verb	4	316
apuntar (*to appoint*)	Verb	9	39
subir (*to rise*)	Verb	5	288
tocar (*to touch*)	Verb	13	117
vencer (*to conquer*)	Verb	7	383
verde (*green*)	Adj	5	342

Table 1. Ambiguous words used in the experiment.

After selected the set of ambiguous words, we used a preprocessed Spanish Senseval-2 Training Set (SSTS) for extracting those contexts that contains such ambiguous words. The number of contexts found for each word is also reported in table 1. The size of this corpus is approximately 41 Megabytes.

The preprocessing phase applied to SSTS concerns the elimination of stopwords and non alphabetic strings, and lowercase conversion. Meanwhile the preprocessing of the contexts found concerns the elimination of non representative words. In order to complete the last task, we used an association measure named *mutual information* (see Manning [3]) that is presented as follows.

Given w_1, an ambiguous word, and w_2, a word of some context of w_1,

$$MI(w_1, w_2) = log_2 \left(\frac{N * Fr(w_1, w_2)}{Fr(w_1) * Fr(w_2)} + 1 \right), \quad (5)$$

where $Fr(w_1)$ and $Fr(w_2)$ are the frequencies of w_1 and w_2, respectively, in the set of contexts, $Fr(w_1, w_2)$ is the frequency of the pair of words w_1 and w_2 in the set of contexts, and N is the number of contexts. The use of mutual information allowed to remove non-influential words from the contexts, and this proposal had a positive impact on the quality of features selected.

4.2 Evaluation

In order to evaluate our results, we used a baseline proposed by Ted Pedersen [14] for unsupervised clustering and used by him in WSD. Calculation of accuracy, for the set of clusters obtained, requires a set of classes; each class must contain examples using one sense for each ambiguous word considered. Clustering of sentences of known senses, $s_1, \ldots, s_n$ in $c_1, \ldots, c_m$ clusters may be seen as shown in table 2 (example given by T. Pedersen).

Cluster/Sense	s_1	s_2	s_3	Total
c_1	10	30	5	45
c_2	20	0	40	60
c_3	50	5	10	65
Total	80	35	55	170

Table 2. Example of baseline computation for 3 senses and 3 clusters.

Assuming the worst case, i.e., all contexts were clustered in c_i, we can conform a baseline. For example, baseline for c_3 (see, table 2) should be 55/170=0.32, if c_3 is associated to the sense s_3, etc. On the other hand, the average accuracy for a word w may be calculated as:

$$Accuracy(w) = \frac{1}{NS} * \sum_{i=1}^{NS} \frac{Gc_i}{Gt_i} \tag{6}$$

where Gc_i is the number of groups correctly assigned; Gt_i, the total number of groups found for sense i, and NS the real number of senses for word w.

The evaluation of our algorithm is presented in the next section. We compare accuracy against baseline for a set of ambiguous words taken from Senseval-2.

4.3 Results

Evaluation of results for eight ambiguous words are shown in table 3. Part (a) of the table shows the accuracy of clustering using SLC-SLC, with feature selection in each execution of SLC. Part (b) of the table, shows the accuracy of clustering using SLC-SLC with feature selection only in the first execution of SLC.

	(a)		(b)	
Word	Accuracy	Baseline	Accuracy	Baseline
corona	0.35	0.25	0.46	0.16
hermano	0.06	0.33	0.46	0.15
apoyar	0.32	0.30	0.43	0.27
apuntar	0.11	0.11	0.11	0.11
subir	0.1	0.1	0.16	0.07
tocar	0.38	0.76	0.076	0.038
vencer	0.012	0.14	0.051	0.03
verde	0.125	0.20	0.20	0.12

Table 3. Accuracy values for SLC-SLC with: (a) double SR, (b) single SR.

5 Conclusions

We have presented an exploration of senses clustering using a feature selection method based on the sense of a syntagm. The result of applying double selection, had a worse performance with respect to the baseline, (-27%), however, by using single selection, the enhancement of baseline is positive and high (102%). These values show that syntagm's sense formula selects good features, and by applying twice this process, not only eliminates noise data, but additional rich features belonging to the sense of the words. Results encourage other applications of syntagm's sense approach in Natural Language Processing, like Information Retrieval, and Text Representation.

References

1. Adam Kilgarriff, Senseval: an exercise in evaluating word sense disambiguation, *Euralex 98*, pag. 176-174, 1998.
2. Castelló, G., N. Artigas, M. A. Martì y M. Taulé, Guía Diccionario CLIC-The Xtract2–WP–03/Barcelona: CLiC-Thera, 2003.
3. Chris Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA: May 1999.
4. German Rigau, Desambiguación automáatica del sentido de las palabras, en *Tecnologías del Texto y del Habla*, M. Antonia, Martì & Joaquim, Llisterri (Eds.). Edicions Universitat de Barcelona, Fundación Duques de Soria, 2004.
5. Hassan M. S., Reda M. S. A. A., Araby I. M., Formation of Machine Groups and Part Families: A modified SLC Method and Comparative Study, *Integrated Manufacturing Systems*, pp. 123-137, 2003.
6. Héctor Jiménez-Salazar, Guillermo Morales Luna, Domain Membership Degrees and Classification Methods, *Computación y Sistemas* Vol. 5 No. 4 pp 288-295, México, 2002.
7. Hwee Tou Ng, Getting Serious about Word Sense Disambiguation, *SIGLEX*, 1997.
8. Josefina Garcia F, Estructura conceptual y comunicacion, *Dimension Antropologica*, Año 2 Vol. 3 pag 75-84, Mexico, 1995.
9. Phillip Edmonds: SENSEVAL: The evaluation of word sense disambiguation systems, *ELRA Newsletter*, Vol. 7 No. 3, 2002.

10. Rada Mihalcea, Timothy Chklovski and Adam Killgariff, The Senseval-3 English Lexical Sample Task, in *Proceedings of ACL/SIGLEX Senseval-3*, Barcelona, Spain, July 2004.
11. Sneath, P.H., Some Thoughts of Bacterial Classification, *Journal of General Microbiology*, Vol. 17, pp. 184-200, 1957.
12. Sofía Paniagua, Héctor Jiménez & David Pinto, Pruebas con algoritmos de agrupamiento para generar una base de datos léxica, *Avances en la Ciencia de la Computación*, pag. 304-310, México, 2004.
13. Sonia Vázquez, Rafael Romero, Armando Suárez, Andrés Montoyo, Iulia Nica, Antonia Martí, The University of Alicante system at SENSEVAL-3, *SENSEVAL-3, ACL*, 2004.
14. Ted Pedersen, A Baseline Methodology for WSD, *Proc. of 3rd. Int. Conf. on Intelligent Text Processing and Computational Linguistics*, CICLing, México, 2002.