A cross-comparison of two clustering methods

Olivier Ferret
CEA Saclay
DTI/SITI
91191 Gif-sur-Yvette Cedex
ferret@sphinx.cea.fr

Brigitte grau and Michèle Jardino
LIMSI CNRS
BP133
91403 Orsay, France
bg,jardino@limsi.fr

Abstract

Many Natural Language Processing applications require semantic knowledge about topics in order to be possible or to be efficient. So we developed a system, SEGAPSITH, that acquires it automatically from text segments by using an unsupervised and incremental clustering method. In such an approach, an important problem consists of the validation of the learned classes. To do that, we applied another clustering method, that only needs to know the number of classes to build, on the same subset of text segments and we reformulate our evaluation problem in comparing the two classifications. So, we established different criteria to compare them, based either on the words as class descriptors or on the thematic units. Our first results lead to show a great correlation between the two classifications.

1 Introduction

Among all the applications in Natural Language Processing (NLP), many require semantic knowledge about topics in order to be possible or to be efficient. These applications are, for example, topic segmentation and identification or text classification. As this kind of knowledge is not easy to build manually, we developed a system, SEGAPSITH (Ferret and Grau, 1998a), (Ferret and Grau, 1998b), to acquire it automatically. In this field, there are two classes of approaches. Supervised learning that requires to know a priori which topics have to be learned and to possess a tagged corpus as a learning set. It is the approach generally adopted by the different systems, as those participating to TREC or TDT. However, we wanted to design a system allowing us to work in open domain, without any restriction about the subjects to be represented and, thus, to be recognized in texts. SEGAPSITH is grounded on an unsupervised and incremental learning based on a conceptual clustering method. After a thematic segmentation of the texts that divides a text in segments made of lemmatized words, i. e. thematic units, the system aggregates similar enough thematic units. Aggregation consists of regrouping all the words of the different similar units and associating to them a weight according to their occurrence number. This weight represents the importance of a word relative to the described topic. The incremental aspect allows us to augment topic knowledge by treating successive corpora without reconsidering the knowledge already existing.

In such an approach, an important problem consists of the validation of the learned classes. As we do not possess an existing classification that agrees with the granularity level of our classes, we decided to accomplish this evaluation by using a second classification method on the same data and by comparing their results. This second method is an entropy-based method, and requires to know the number of classes to form. So, if both results are similar enough, although the methods applied are different, we could conclude that the learned classes are quite relevant and that the two methods are efficient.

After applying the second method, we possess two sets composed by the same number of classes. Each class regroups thematic units and is described by a set of words. So, we established

different criteria to compare them, based either on the words as class descriptors or on the thematic units they gather. After the presentation of the two methods, we shall present our tests and the first results we obtained.

2 Semantic domain learning

This description aims at showing the data used for learning, and the specificity of the learned classes.

2.1 The thematic segmentation: SEGCOHLEX

Studied texts are newspaper articles coming from two corpora: "Le Monde" and "AFP". Some of these texts have been used to build a lexical network where links between two words represent an evaluation of their mutual information to capture semantic and pragmatic relations between them, computed from their co-occurrence count. In order to build class of words linked to a same topic, we first realize a topic segmentation of the texts in thematic units (TU) whose words refer to the same topic, and learning is applied on these thematic units.

Text segmentation is based on the use of the collocation network. A topic is detected by computing a cohesion value for each word resulting from the relations found in the network between these words and their neighbors in a text. As in Kozima's work (Kozima, 1993), this computation operates on words belonging to a focus window that is moved all over the text.The cohesion values lead to build a graph and by successive transformations applied to it, texts are automatically divided in discourse segments. Such a method leads to delimit small segments, whose size is equivalent to a paragraph, i. e. capable of retrieving topic variations in short texts, as newswires for example. Table 1 shows an extract of the words belonging to a cohesive segment about a dedication of a book.

2.2 Semantic Domain learning in SEGAPSITH

Learning a semantic domain consists of aggregating all the most cohesive thematic units, TUs, that are related to a same subject, i. e. a same kind of situation. We only retain segments whose cohesion value is higher than a threshold, in order to ground our learning on the more reliable units. Similarity between a thematic unit and a semantic domain is evaluated from their common words. When the similarity value exceeds a threshold, the thematic unit is aggregated to the semantic domain, otherwise a new domain is created. Each aggregation of a new thematic unit increases the system's knowledge about one topic by reinforcing recurrent words and adding new ones. Weights on words represent their importance relative to the topic and are computed from the number of occurrences of these words in the TUs.

Units related to a same topic are found in different texts and often develop different points of view of a same type of subject. To ensure a better similarity between them, SEGAPSITH enriches a particular description given by a text segment by adding to these units those words of the collocation network that are particularly linked to the words found in the segment. Table 2 gives an extract of the words added to the segment of Table 1.

This method leads SEGAPSITH to learn specific topic representations (see Table 3) as opposed to (Lin, 1997) for example, whose method builds general topic descriptions as for economy, sport, etc. Moreover, it does not depend on any a priori classification of the texts.

We applied the learning module of SEGAPSITH on one month (May 1994) of AFP newswires, corresponding to 7823 TUs. In our experiments (Ferret and Grau, 1998a), (Ferret and Grau, 1998b), we showed that domains reach a stability at 15 to 20 aggregations, and that words having a weight below 0.1 are rarely related to the domain. Thus, we only selected domains resulting from at least 15 aggregations for our cross-comparison, i.e. 71 domains regrouping 4935 TUs and 4380 words having a weight upon 0.1. A lot of domains share common words, and are close enough to be considered as different representations of specific points of view of a general topic, as economy, sport, etc.

3 Entropy-based clustering

The second clustering method gives an optimal partition of the 4935 thematic units in 71 non-

words	weight	words	weight
strider	0.683	entourer (to surround)	0.368
toward	0.683	signature (signature)	0.366
dédicacer (to dedicate)	0.522	exemplaire (exemplar)	0.357
apposer (to append)	0.467	page (page)	0.332
pointu (sharp-pointed)	0.454	train (train)	0.331
relater (to relate)	0.445	centaine (hundred)	0.330
boycottage (boycotting)	0.436	sentir (to feel)	0.328
autobus (bus)	0.435	livre (book)	0.289
enfoncer (to drive in)	0.410	personne (person)	0.267

Table 1: Extract of a segment about a dedication

inferred words	weight	inferred words	weight
paraphe (paraph)	0.522	imprimerie (press)	0.418
presse_parisien (parisian-press)	0.480	éditer (to publish)	0.407
best_seller (best_seller)	0.477	biographie (biography)	0.406
maison_d'édition (publishing_house)	0.450	librairie (bookshop)	0.405
libraire (bookseller)	0.447	poche (pocket)	0.389
tome (tome)	0.445	éditeur (publisher)	0.363
Grasset (a publisher)	0.440	lecteur (reader)	0.355
rééditer (to republish)	0.428	israélien (Israeli)	0.337
parution (appearance)	0.427	édition (publishing)	0.333

Table 2: Extract of words selected in the collocation network for the segment of Table 1

words	occurrences	weight
juge d'instruction (examining judge)	58	0.501
garde_à_vue (police custody)	50	0.442
bien_social (public property)	46	0.428
inculpation (charging)	49	0.421
écrouer (to imprison)	45	0.417
chambre_d'accusation (court of criminal appeal)	47	0.412
recel (receiving stolen goods)	42	0.397
présumer (to presume)	45	0.382
police_judiciaire (criminal investigation department)	42	0.381
escroquerie (fraud)	42	0.381

Table 3: The most representative words of a domain about justice

overlapping clusters according to the word distributions in the units. It is realized with an algorithm which looks like K-means (here K=71). Each cluster is the merge of several thematic units and is represented by its centroid. We search for the partition which minimizes the Kullback-Leibleir divergence (Cover and Thomas, 1991) between the word distributions of the thematic units and those of of their centroids. This entropy-based measure is convex (Jardino, 2000), this propriety permits to get an optimal partition whatever the initial conditions.

3.1 Entropy

We assume that each thematic unit is represented by one quantitative vector whose components are

the relative occurrences of a selection of words related to the unit. The advantages of this normalization is that the representation of the thematic units does not depend on the length of the units and can be modelized in the frame of the information theory (Cover and Thomas, 1991).

Assuming that $O_{w,tu}$ is the occurrence of the word labelled w in the thematic unit labelled tu and that O_{tu} is the occurrence of all the words in the thematic unit tu, such that $O_{tu} = \sum_w O_{w,tu}$, each thematic unit vector component , $p(w/tu)$, is :

$$p(w/tu) = \frac{O_{w,tu}}{O_{tu}} \quad (1)$$

When the thematic units are unclassed, their entropy is given by (Cover and Thomas, 1991):

$$H_{TU} = -\sum_{w,tu} p(w,tu) * ln[p(w|tu)] \quad (2)$$

with $p(w,tu) = \frac{O_{w,tu}}{\sum_{w,tu} O_{w,tu}}$

When the thematic units are gathered in K clusters, labelled k, the cluster entropy is, H_K:

$$H_K = -\sum_{w,k} p(w,k) * ln[p(w|k)] \quad (3)$$

where $p(w|k)$ is defined as :

$$p(w|tu \in k) = p(w|k) = \frac{O_{w,k}}{O_k} \quad (4)$$

with $O_{w,k} = \sum_{tu \in k} O_{w,tu}$, $O_k = \sum_{tu \in k} O_{tu}$ and $p(w,k) = \frac{O_{w,k}}{\sum_{w,k} O_{w,k}}$

The cluster entropy is always higher than or equal to the unit entropy (log-sum rule (Cover and Thomas, 1991)), so that the Kullback-Leibleir divergence defined as:

$$D_{KL} = H_K - H_{TU} \quad (5)$$

is always higher than or equal to 0.

3.2 Clustering algorithm

Minimizing the Kullback-Leibler divergence amounts to minimize the entropy H_K because H_{TU} does not depend on the clusters.

The number of possible partitions is huge, roughly 4935^{71}. We have observed that a random search is faster than a systematic one (Jardino, 2000), and we have used this paradigm to build the algorithm described below:

1- Define a priori, K, the cluster number, here K=71.

2- Initialize: put all the thematic units in one cluster, calculate the entropy H_K (equation 3). The remaining K-1 clusters are empty.

3- Do the random selection of one thematic unit and of another cluster for this unit.

4- Move the unit from its former cluster to the new randomly selected one, calculate the new entropy.

5- If the new entropy is lower, leave the unit in its new cluster, otherwise move it back to its initial cluster.

6- Repeat 3 to 5 until there is no more change.

The optimal clustering of the 4935 thematic units in 71 clusters is performed on a workstation (SGI Indy) within twenty minutes.

4 Comparing two classifications

We established different criteria for comparing the two classifications, based on the elements used to describe the classes. First, each class is a set of words with an occurrence number for each of them; second it is also a set of thematic units. So, the comparison can be done along these two points of view.

In order to evaluate the overlapping of the classes of words, we applied each classification method on the two classification results: the classes of words resulting from the second method are classified relative to the semantic domains. For comparing the classes of TUs, we applied the entropy measure on one hand to measure the overlapping of the classes, and κ and Mantel tests on the other hand to evaluate the differences in the repartition of all the TUs.

4.1 The word point of view

4.1.1 Classification by similarity

The classification of the clusters relative to the semantic domains exploits the same similarity measure than the one used for the learning phase. In a first step, some domains are selected according to the value of the $activ$ function:

$$activ(d) = \sum_i W_{d,i} * W_{c,i} \quad (6)$$

where $W_{d,i}$ is the weight of the word i in the semantic domain d and $W_{c,i}$ is the weight of the same word in the cluster c. This first step was used in the learning phase because the number of semantic domains was increasing rapidly and this measure leads to a first fast selection of interesting domains before evaluating an in-depth similarity. We kept this step, even if it was not necessary, in order to apply exactly the same method in the evaluation phase. Afterwards, each selected domain can be compared to the cluster by using the similarity measure given below. If one of these similarity values is greater than a given threshold, fixed to 0.25 in our tests, the cluster is linked to the domain that is the most similar to it. The similarity measure is:

$$sim(d,c) = \sqrt[4]{\frac{\sum_w W_{d,w}}{\sum_t W_{d,t}} \frac{\sum_w O_{d,w}}{\sum_t O_{d,t}} \frac{\sum_w W_{c,w}}{\sum_t W_{c,t}} \frac{\sum_w O_{c,w}}{\sum_t O_{cl,t}}} \quad (7)$$

where the w index is used for indicating common words between the cluster c and the semantic domain d and the t index, for indicating all the words of the cluster or the domain. W is the weight of a word and O its occurrence number.

The similarity measure is only based on the common words. As learning is unsupervised and incremental, differences at time t might disappear at time $t+1$. The comparison is based on the proportion of common words relative to the total of words of each entity to be compared. The evaluation of the common words in each entity is done according to their occurrence number and their weight. So, we avoid to obtain a high similarity value between two entities that only share very few words having a high weight. We combine these criteria in a geometrical mean for evaluating each entity and for computing the global similarity from the evaluation of the two entities in order to smooth the effect of few recurrent words when the domains are in their formation phase, words that would act as attractors otherwise.

4.1.2 Entropy-based classification

For each of the 71 clusters, we have searched for the nearest domain obtained with the same kind of entropy-based measure defined above. We assume that we have a probabilistic model which gives the predictions of the words according to the domains. In order to avoid the null value, non-learned events are infered using the Witten-Bell interpolation (Witten and Bell, 1991). The interpolated value of the prediction of a word w, knowing the domain d is $p'(w|d)$ such that:

$$p'(w|d) = \frac{O(w,d) + n_{sw}(d)/V}{O(d) + n_{sw}(d)} \quad (8)$$

where $n_{sw}(d)$ is the number of words seen in each domain and V the size of the vocabulary. Each cluster is also defined by a set of words and we compare the distribution of the words in the cluster with the distributions of the words in the domains (equation 8) with the Kullback-Leibler divergence. Each cluster is associated with the nearest domain.

4.1.3 Comparison

The results of the two classifications described above are given in Table 4. For the similarity-based classification, only 3 clusters do not match with any domain and 47 different domains are selected for the 68 remaining clusters with 34 links that are one cluster-one domain. For the entropy-based classification, 44 clusters have been associated to the 71 domains.

Several clusters are linked to the same domain. This can been explained by the closeness of some of the domains. This is shown when they are hierarchically classified; we obtain then 34 general domains that regroup 1 to 5 domains each. We also observe that most clusters are only linked to one domain. The two methods give almost the same results and show that the two classifications are similar.

domain⇒cluster link	links (similarity)	links (entropy)
no link	3	
1→1	34	29
1→2	8	8
1→3	3	4
1→4	1	2
1→5	1	
1→6		1

Table 4: Number of links between one domain and the clusters

4.2 The TU point of view

4.2.1 A simple comparison

One domain and one cluster are associated to each thematic unit. It is then possible to calculate the number of domains and clusters which partially or fully overlap. Table 5 represents the intersection between the two partitions. For each domain we chose the cluster which has the highest intersection with the domain. Then we calculated the percentage of thematic units of this domain which are both in the domain and in this chosen cluster.

coverage	number of clusters
cov=100%	8
80%≤cov¡100%	16
60%≤cov¡80%	19
40%≤cov¡60%	19
20%≤cov¡40%	8
cov<20%	1

Table 5: Coverage rates of UT which are common to each domain and those of the associated clusters which correspond to the highest intersection

Height domains are identical to height clusters. The lowest coverage (18%) is obtained for one domain. The other seventy domains cover more than 20% of the clusters.

4.2.2 Comparison with the κ coefficient

In order to compare more precisely our two classifications, we used the κ coefficient as it was done by Dietterich in (Dietterich, 2000) and as it is often done in the field of remote sensing for example. The κ coefficient measures the degree of agreement among several judgements and is expressed as follows:

$$k = \frac{\theta_1 - \theta_2}{1 - \theta_2} \tag{9}$$

where θ_1 is the proportion of times that the judgments agree and θ_2 is the proportion of times that we could expect the judgments to agree by chance. As we are in a case of unsupervised classification whereas Dietterich's work was about supervised classification (building of decision trees), we have first set a one-to-one mapping between the semantic domains and the clusters. This was done by a very simple procedure: we computed the size of the intersection between each cluster and each domain; then we iteratively mapped the cluster and the domain that had the largest intersection until each cluster was mapped with a domain. Of course, this is not an optimal procedure in order to ensure that the intersection of each couple of classes is the largest one but it can be considered as a baseline.

Then, the evaluation of the κ coefficient was done by building a matrix $K * K$, with K, the number of classes (clusters or domains), such that each element $k_{i,j}$ is equal to the number of TUs assigned to the class i by SEGAPSITH and to the class j by the entropy-based clustering. θ_1, which estimates the probability that the two classifications agree, is defined by:

$$\theta_1 = \frac{\sum_{i=1}^{K} k_{i,i}}{N} \tag{10}$$

where N is the total number of TUs. It evaluates the proportion of TUs that were put in the same classes by the two clustering algorithms.

θ_2, which estimates the probability that the two algorithms agree by chance, is given by:

$$\theta_2 = \sum_{i=1}^{K} (\frac{k_{i+}}{N} * \frac{k_{+i}}{N}) \tag{11}$$

where $\frac{k_{+i}}{N}$ and $\frac{k_{i+}}{N}$ are the marginal distributions.

The κ coefficient that results from the evaluation of θ_1 and θ_2 is equal to 0 when the two clustering algorithms agree only by chance and to 1 when they really agree for each TU. Negative values occur when there is a systematic disagreement.

For the 71 classes of our test set, we computed the κ coefficient in two cases. First, with a random mapping of the clusters and domains. We got K = -0.013, which is very close to the agreement by chance. Second, we applied the above mapping procedure and got K = 0.484, which indicates a significant correlation between the two classifications. We think that with a more complex mapping procedure, the κ would be higher.

4.2.3 Application of the Mantel Test

In this paradigm, each classified thematic unit, tu, is described according to its position in the

classification in relation to all the classified elements. This position is characterized by a distance between tu and each other element. In the work we present here, we choose a simple distance: $dist(tu_1, tu_2) = 0$ if tu_1 and tu_2 are part of the same class; otherwise, $dist(tu_1, tu_2) = 1$. However more complex distances may be used when the classifications are hierarchical ones for example. After this first step, each tu_i of the two classifications to compare is characterized by a vector, each element of which, d_{ij}, is equal to the distance between tu_i and tu_j. Hence, each classification is characterized by a distance matrix, which is a square symmetric matrix of size $N^2 = 4935^2$. Comparing the two classifications amounts to compare their distance matrices. In the ecology field, such kind of comparison is achieved by a statistical test, called the Mantel test (Mantel, 1967). In (Legendre, 2000), Legendre defines the Mantel test as " a procedure to test the hypothesis that the distances among objects in a matrix A are linearly independent of the distances among the same objects in another matrix B. The result of this test may be used as support for or against the hypothesis that the process that generated the first set of distances is independent of the process that generated the second set. The unique feature of the Mantel test is the use of a linear statistic to assess the relationship between two distance matrices". The basic statistic used in the Mantel test is the Z statistic:

$$ZS = \sum_{i=1}^{i=N} \sum_{j=1}^{j=N} x_{ij} y_{ij}$$

As the elements of a distance matrix are not independent, the significance of ZS, the Z statistic that is computed for the two distance matrices to compare, is evaluated by comparing this value to the Z statistic that is computed for matrices whose rows and columns are randomly permuted. A distribution of random values is obtained by computing the statistic for many permuted matrices and if ZS is significantly above this distribution, the hypothesis that the two matrices are independent is rejected [1].

[1]The Z statistic is maximal when the two distance matrices are identical: the $x_{ij}y_{ij}$ term is not equal to zero only if x_{ij} and y_{ij} are equal to 1. Hence, each difference that could be introduced between the two matrices, decreases its value.

As an exploratory step, we applied the Mantel test in order to compare the results of the two classification methods we presented in this article. We used the software developed by Adam Liedloff (Liedloff, 1999). As the number of TUs is too large in comparison with the capabilities of this software, we experimented the Mantel test only on a subset of 1000 TUs. With the distance matrix computed from the results of the two classification methods, we got a Z statistic (ZS) equal to $940,894$. The maximum value of ZS is $978,460$ for the domains and $948,608$ for the clusters. The random distribution was built from 99 permuted matrices and its ZS value is $937,708 \pm 232$. As the proportion of the values from the random distribution that are above ZS is equal to zero, we can reject the hypothesis that the two matrices are independent and as a consequence, we can think that the two compared classifications are globally similar. However, as the results of the Mantel test are not easy to interpret, further tests must be performed to see what are the relations between these results and those of the other comparing methods and to determine if this test is actually suited for comparing such kind of classifications.

5 Conclusion

We presented in this paper an approach for evaluating the results of an unsupervised learning method, when no human evaluation is possible or when no classification exists as a reference. As a result, this method builds classes of weighted words that regroup thematic units. We defined in a previous work a stability threshold of these classes, thus we aim at evaluating this subset of classes. To do that, we applied another clustering method that only needs to know the number of classes to build on the same subset of TUs and we reformulate our evaluation problem in comparing the two classifications. Our first results lead to show a great correlation between the two results. We now have to develop other tests, for example on a different number of classes, to verify our first results. A second step will be to evaluate the methods on the same task, as a classification task for example, whose protocol has to be defined.

References

T. Cover and J. Thomas. 1991. *Elements of Information Theory*. Wiley & sons, New York.

T. G. Dietterich. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40:139–158.

O. Ferret and B. Grau. 1998a. A thematic segmentation procedure for extracting semantic domains from texts. In *ECAI*, Brighton, UK.

O. Ferret and B. Grau. 1998b. Structuration d'un réseau de cooccurrences lexicales en domaines sémantiques par analyse de textes. In *NLPIA*, Moncton, Canada.

M. Jardino. 2000. Unsupervised non-hierarchical entropy-based clustering. In H.A.L.Kiers, J.-Rasson, P.J.F.Groenen, and M.Schader, editors, *Data Analysis, Classification, and Related Methods*. Springer.

H. Kozima. 1993. Text segmentation based on similarity between words. In *ACL (Student Session)*, Ohio, USA.

P. Legendre. 2000. Comparison of permutation methods for the partial correlation and partial correlation and partial mantel tests. *Statistical Computation and Simulation*, 67:37–73.

Liedloff. 1999. Mantel nonparametric test calculator. *http://www.sci.qut.edu.au/nrs/mantel.htm*.

C.-Y. Lin. 1997. *Robust Automated Topic Identification*. Ph.D. thesis, University of Southern California.

N. Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.*, 27:209–220.

I.T. Witten and T.C. Bell. 1991. The zero-frequency problem: estimating the probabilities of novel events in adaptative text compression. *IEEE Transactions on Information Theory*, 37(3):1085–1094.