Multiple text interview clustering with the help of a hierarchical classification algorithm.

Stéphane Lapalut
ACACIA project, INRIA Sophia Antipolis,
BP 93, 06 902 Sophia Antipolis cedex, France
Stéphane.Lapalut@sophia.inria.fr

Abstract

Besides the factual knowledge written in books, lot of knowledge can be obtained by questioning experts. This produces a large amount of rich but ill-structured texts. To deals with them, we propose the use of a statistical method, the top-down hierarchical classification and a new interpretation of its results. The analysis gives two kinds of results: first a segmentation of texts that reflects their "semantic contexts", which we use to raise structures of texts, and second, classes of significant terms related to these contexts that help to interpret classes meaning. We briefly describe the method with its formal background and then we applied it on an example borrowed from the knowledge acquisition field, the processing of a corpus of text interview from several experts. We give the main conclusion from the analysis. We conclude with the usefulness and developments of the method.

Keywords

multiple text processing, complex text processing, statistical analysis, text clustering, text comparison, top-down hierarchical classification.

1 Introduction

For the human intelligence, it is difficult to cope with a huge corpus of texts, such as texts relating interviews from experts. We have adapted a statistical method that deals with semantic contexts, to facilitate the work of the knowledge engineers. As our method and the algorithm we use are blind with respect to the text meaning, we can applied them to process all kinds of texts.

The statistical method, called "top-down hierarchical classification", handles both self contained texts and sets of chosen texts. It detects groups of terms of the corpus strongly distinguishable, according to the statistical occurrence of meaningful terms or pairs of them in small text units. These groups are called classes and have been identified as relevant semantic contexts (Reinert, 1979). They lead to a partition of the corpus that reflects its structure. After interpreting each class with the help of the related terms, the analyst knows the subject of each part of the structure of the corpus. We experimentally find out that in the case of a multiple text corpus, parts of the raised structure are closely related to the initial texts if there are strongly distinguishable, or characterize groups of similar texts. In both case, the class characterization allows a quick and safe tagging of texts, raising their main topics.

We recall in the next section the method fully described in (Lapalut, 1995), and we briefly expose the algorithm principle. The validity of our method is discussed in the third section and a case study, a multiple text analysis, performed in the fourth section. Then we report related work in the fifth section. We conclude in the last section with research directions to carry on the method one step further.

2 The top-down hierarchical classification method

The initial objective addressed was to discover semantic contexts characterized by groups of terms. A principle derived from the Huyghens decomposition formula is used: *"to find a partition of a set that minimizes the intra-class variance, it is sufficient to find dichotomies that maximize the inter-class variance"*. This section give a taste of the method that has been developed by Max Reinert since 1984 and implemented in a tool called ALCESTE (Reinert, 1992).

We refine the initial cycle to adapt it for our purpose of discovering the structure of a corpus. We have found a correlation between natural articulations of the corpus and the structure given by the classes. The classes give a structure that are refined by the user with the help of class interpretations. Formally, we can consider the corpus processing as the research of correspondences between sets from a triplet $\{U_1, TU_1, S\}$ where U_1 is the elementary segmentation of the corpus (the set of elementary context units (ECU)), TU_1 the objective clustering of U_1 into classes (found by the top-down classification) and S the implicit organization of the corpus according to the interrelated topics in experts' interview: this organization is to be discovered. The whole analysis permits the clarification of S, which is the correspondence between the corpus viewed as a sequence of ECU, U_1, and semantics contexts defined by classes in TU_1.

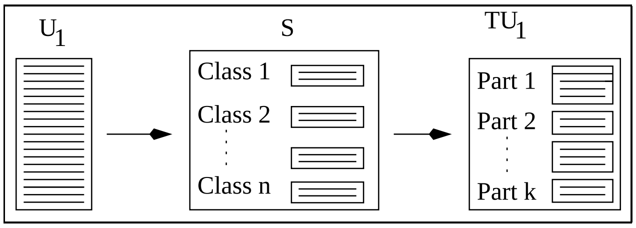

Figure 1: Evolution of ECU sets from the raw corpus to the structured corpus.

2.1 Data and preprocessing

The algorithm developed by Reinert (Reinert, 1986; Reinert, 1992) takes two inputs : firstly the corpus and secondly a set of parameter called the analysis plan. The corpus is a set of one or more French texts. It can be roughly structure into initial context units (ICU), to delimit well-known parts, such as section or single text in case of a multiple text corpus. The analysis plan specifies how the statistical analysis will be carrying on. The main parameters are the maximum number of desired classes, the presence of a double classification and the length of UCE, which is fixed to validate the application of the underlying statistical theory. In case of a double classification, two different ECU lengths are needed to allow the determination of stable classes.
The first stage of the process performs a morphological reduction of words. Then the corpus is splited into ECU. An ECU is a suite of a fixed number of plain-word (or non-noisy-word). A plain-word is a word that it is not recognize as a tool-word (or noisy-word), according to a glossary of tool-word. At the same time, the list of the plain-word (terms) is established. The list of ECU and the list of term serve as the rows and columns entries of a binary matrix. A "1" means that the considered term occurs in the considered ECU. So the matrix will be sparse, due to the reduced number of term in a ECU (10 to 20) and the great number of different term in a corpus. This matrix is processed by the hierarchical top-down classification algorithm, as presented in the next section.

2.2 The algorithm

Given the boolean table, which uses terms as column entries and ECU as row entries, we define classes as particular sets of rows. A given row belongs to a single class, but an overlap of the set of term between two classes is allowed, even if some terms are specific to a single class at the end of the analysis. The final set of classes is an incomplete partition of the initial set of ECU.

To distinguish classes, the algorithm calculates a χ^2 distance between the margins of two given sub-tables, determined by successive dichotomies. A margin is a row vector formed by the sum of all values in each column. At each step of the algorithm, the dichotomy that maximizes the χ^2 rate of the margin tables is found. The algorithm stops when the specified number of classes is reached.
For our application, we have used an analysis that performs a double classification. Then a crossing between these two classifications determines a set of stable classes. For each resulting final classes, sets of both term and couple of term attached to the classes, say the profiles, are calculated from the set of ECU for this class, together with meaningless (less correlated) term and couples of term, say the anti-profiles. This is these profiles and the set of ECU that characterize a class, which we use in our new extension and interpretation of the analysis result. We describe it in the next section.

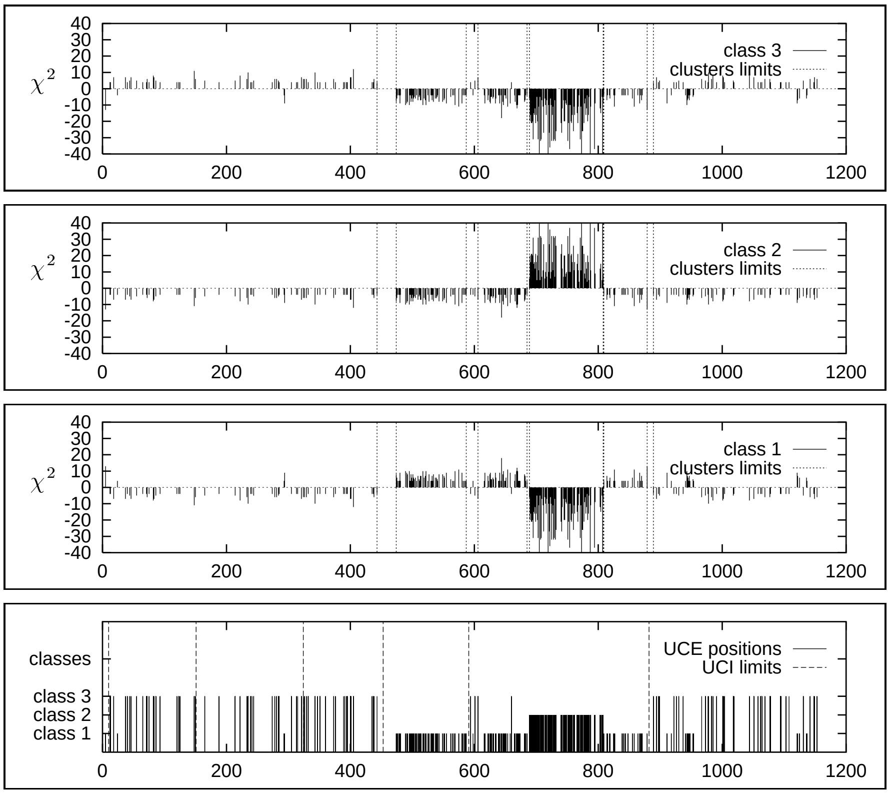

Figure 2: Comparison between the found segmentation and the initial structure of a well-formed text.

2.3 Exploitation of the results

The main analysis result used to produce the text segmentation is the sets of ECU that characterize each classes. For each of these ECU, a χ^2 rate is calculated. We represent these data within some diagrams (see figure 2). The x-axis represent the ECU in the order they appear in the corpus. We can represent all ECU for one class using their χ^2 rate along the positive y-axis and the χ^2 rate of the ECU of the other classes along the negative y-axis. This allows a direct detection of the cluster, and then the deduction of a raw segmentation, which is precised by the user. Another representation to check the cluster repartition use all the ECU characterizing classes together with the class number along the y-axis.

For each class, with the help of the four profile and anti-profile files, the user is able to tag them. From the term and couple of term profiles, we produce two possible tags, which are suppose to sum up what the class is. Then we do the same with the anti-profiles, which produces two tags about what the class isn't. These four tags are compared and cross-checked. This last step leads to a good interpretation of the relevant class subject. For each cluster, its related class together with a careful reading of the corresponding corpus part, help to validate the segmentation it suggests. Finest interpretations can be drawn from the statistical results, but this is out of the scope of this paper.

3 Validation

As our method is a semi-automatic one, its validity relies partially on the user's interpretations. A critical data is the length of the corpus: the bigger its size, the more precise the statistical analysis results. The main point of these results is the corpus segmentation according to ECU clusters. Each cluster can be characterized by the number of ECU belonging to the same class, their χ^2 rates, and the length of the cluster in term of ECU. We haven't found yet a formal characterization of these clusters, even if all their characteristics seem to be correlated. Anyway, the segmentation is given by the articulations between clusters, which are precisely located in the text by a human reader. This reading allows a checking of the clustering, together with a determination of the subject relevant to each cluster. As we have seen it in the result interpretation, the class types of the clusters allow, at last, a checking of both the analysis and the interpretation.

We have test our method on various corpora, from single to multiple experts' interviews, and from hill-structured to well-structured corpus. The size of the corpus seems to directly influence the level of the segmentation. With a big corpus, we obtain a precise but a large grain size segmentation. With a middle size corpus, we obtain a fine grain sized segmentation but with less accuracy in the χ^2 rates. In (Lapalut, 1995), we have shows this phenomena with a single text corpus analysis and the analysis of a multiple text corpus including this single text.
The uniqueness of the segmentation is a relevant question. From hill-structured corpus, such as text interview, experimentation are difficult and sometimes, the human reader is unable to raise such a segmentation. With the help of our method, he uses the results as a guideline and can validate/adjust the proposed segmentation. From well-structured corpus, we have a way to compare the segmentation with an objective one. As an example, we have process the well-known "Discours de la méthode", from the French philosopher René Descartes[1] (see figure 2). This text is initially structured into six parts. We have made the corpus with one ICU per part, and process them in various order, with different ECU lengths, asking for six to twelve classes. We have also process the corpus without ICU, which leads to the same segmentation. These minor parameter variations do lightly affect the final segmentation, even if the precision of the results suffer from the approximative ECU lengths. The determined classes remain stable, together with the correlated segmentation. Three of the initial segmentation are rediscovered by the analysis. One other is splited into three clusters and the three ICU at the beginning of the corpus assimilated into one existing but weakly characterized cluster (see figure 2). We remark that this corpus reaches the lowest size for the statistical analysis to give accurate results, even if the class crossing carries out three well separated classes (92.65 % of correctness for the ECU/class correlation in figure 2).

For the purpose of the ill-structured text processing, and with only a minimal linguistic analysis, these results are good enough and helpful when dealing with the knowledge acquisition task (Lapalut, 1996a; Lapalut, 1996b).

[1] Thanks to the Association des Bibliophiles Universels for providing this text to the world-wide community (ftp://ftp.cnam.fr/pub/ABU).

4 Multiple text interview clustering

The whole analysis process is fully describe in (Lapalut, 1995). The main hypothesis is that in case of a corpus gathering well-distinguishable texts, if we attribute one ICU per text, the obtained segmentation must conform to this initial structure (hence, we must obtained one cluster per ICU). If this is not the case, some texts are not well distinguishable and we have to interpret in what sense they share some similarities. If the texts are well-distinguishable, we can characterize their differences.

We present the results we obtain for a corpus composed by ten experts' interview texts[2] (about 15000 lines) within the same corpus. The obtained classes presented in figure 3 lead to a corpus structure very close to the existing ICU partition, e.g. the initial contextual units, one text putted in the corpus being one ICU. Precisely the program points out 11 classes. Each of them is directly relevant for one or two ICU:

(a) 5 ICU are relevant for a single class (ICU 2, 5, 7, 9 and 10),

(b) two ICU for two classes (ICU 1 and 4),

(c) one ICU for three classes (ICU 6),

(d) and the last two ICU (3 and 8) are not directly characterized by a class, but they are weakly related to two classes.

	classes	*yve-int	*pie-int	*man-int	*jlo-int	*fra-int	*dan-jlo-24	*dan-man-002	*dom-fra-24	*man-pie-jlo-003	*man-solo-003
Yve's accident studies	1	X	~	~							
impact, trace, deformation	2						**X**	X	~		~
*pie-int	3		X								
crash, kinematics analysis	4				X		**X**				
road infrastructure	5			~		~					
perception of infrastructure	6	~									
the grip tyres on the road	7					X					
central lane and indicator	8									X	X
infrastructure and approach	9						**X**	~	~		
the whole accident analysis	10				X						
the accident situation	11	X									
class interpretations		1	2	3	4	5	6	7	8	9	10

*<label> names each experts' interviews.

X states a strong correlation between the class and the text (more than half of the ECU that characterized the class belong to these texts).

~ states a weak but existing correlation between the class and the text.

Figure 3: Correlation between texts and classes from a multiple text corpus.

In the case (a), we have verified that the semantic contexts described by the classes are in accordance with the related texts. Case (b) leads to the same conclusion. The two classes related to ICU 1 and 4 correspond to the same domain with a slight point of view difference. Case (c) is a very interesting one and we investigate it in further details in (Lapalut, 1995). Anyway, classes related to ICU 6 are in accordance with an analysis of that text we have made before. At that time, we have no well-established explanation for the case (d). So, we already can conclude that

[2]These texts come from a research supported by the French "Ministère de la Recherche et de l'Espace" under contract n.92 C 0757 and the French "Ministère de l'Equipement, des Transports et du Tourisme" under contract n. 93.0003.

this kind of analysis makes precise similitudes and differences between the texts entered as ICU in the corpus. For each of them we obtained a synthetic point of view of their main subject (cases (a) and (b)).

The huge number of terms and ECU contained in this sort of multiple text corpus allows a very contrasted class characterization. χ^2 rates are higher than in the case of single text that makes the interpretation of classes easier. For each class, a small set of terms are preponderant. But each of them is very precise and specific, such as the class 8: “the accident factors for the case 003, central lane and indicator ambiguity”. This class is relevant for the text 9 and its characterization is exactly the conclusion of a manual analysis we made before.

5 Related Work

Statistical approach is far from being a new one in text processing. It has been widely used in text categorization(Jacobs, 1993; Register and Kannan, 1992), as well as in language analysis (Wu et al., 1992; Charniak, 1993). Some applications have been made to the knowledge acquisition field. Statistical methods are mostly used to strengthen a conceptual method (Register and Kannan, 1992; Fall et al., 1989; Jacobs, 1993). Tools such as NLDB (Jacobs, 1993) and SKIS (Register and Kannan, 1992) combine both conceptual and statistical approaches within a hybrid system. Most applications tend to provide a set of tools such as NLDB, which is a "set of statistical method", or to offer an assistant to a human expert, like INLEN (Kaufman et al., 1991) that "performed a sophisticated data analysis", KITTEN (Shaw and Gaines, 1987) where "the knowledge acquired is being feedback to facilitate the intelligence of people", MOCA (Fall et al., 1989) that "helps analysts to cope with large quantities of intelligence data" and SADC (Moulin and Rousseau, 1992). A few of them pretend to perform automated analysis, such as NLDB and SKIS, but they require the user to build lexicons (NLDB) or well defined sets of criteria (keyword↔category correspondences in SKIS) adapted to the knowledge engineer purpose and to the considered domain.

Texts in natural language format serve as entries in KITTEN, NLDB, SADC and SKIS. Corpora are used in various ways, from "a set of entities that are sentences whose features are words they contain" (IMS project, (Gaines and Shaw, 1994)) to sequences of keywords (SKIS). The ALCESTE segmentation method is close to the one used in KITTEN, but with the constraint of the ECU fixed length. Statistical treatments mostly use term weighing. Some algorithms, such as TEXAN in KITTEN or the "similarity measuring component" in SKIS implement a distance measurement too. The only algorithm with a pure distance measurement is TEXAN. But it is a "simple distance-in-text measure" that lacks to take into account the whole corpus. ALCESTE process the whole corpus, by the way of the crossed entries of non-noisy words and ECU in the matrix , which allows a measurement distance based on a χ^2 criteria. It meets the requirements proposed in MOCA, as "it does not need to input the number of clusters desired", it allows "overlapping cluster" on non-noisy words and it analyses the corpus as a whole to produce clusters.

The intended results of discovering the corpus structure can be compared to the one of the text logical structure as in SADC method, despite that this latter uses a purely conceptual approach. The work of Hearst (Hearst, 1994) addresses this goal too. We must take into consideration this latter approach as its goal is nearly the same as ours, but with some differences, both in the used method and in the obtained results. Her goal was to "partition expository texts into coherent multi-paragraph discourse units which reflect the subtopic structure of the text". This is the first difference with our method, as expository texts are a bit more structured than texts from expert interviews. She uses a "texttiling" algorithm based on similarity determination. She processes a corpus segmentation into "token-sequences" of a fixed length. A token-sequence includes a predetermined number of consecutive tokens (around 20) and it is similar to ECU in our method. Then she uses a window including a fixed number of token-sequence (around 6) to perform "comparison of adjacent pairs". So this algorithm performs a *local* analysis of the corpus. It only gives similarity indications without any other detail than the token occurrence frequency. She does not seem to use the pre-existent structure of the corpus.

The algorithm we presented in this paper performs an *overall* analysis and a correspondence between text units. We think it is more suitable for ill-structured text, such as interviews, than Hearst's texttiling algorithm. With our method, we obtain more than the overall corpus structure. For instance, detailed indications about term distribution and correlation enable us to tag the structure and can be used later for other purposes, such as ontology building. To conclude, the presented method and algorithm meet previous work. They are candidate to be a "more sophisticated text analysis technique" (Shaw and Gaines, 1987), but till now with a human assistance. The corpus organization (ICU), its segmentation (ECU) and the clustering algorithm contribute to enhance and refine the set of statistical tools and methods commonly used; then the overall extension reaches a useful feature to help the knowledge acquisition process.

6 Conclusion and discussion

From the analysis of single and multiple text corpora we conducted, both inter-text and intra-text comparisons are possible. The validation tests show the robustness of the method. The segmentation remains stable within little variations of the main parameters.
From a single text analysis, the gained segmentation together with the type of each clusters provide helpful guidelines to the knowledge engineer for focusing his work and becoming acquainted with the text in a quicker and safer way. The multiple text analysis allows to draw comparisons between texts. Above a given number of different kinds of text, e.g. ICU in the corpus, the overall text differences are given back by classes. Texts strongly characterized by specific subjects lead to peculiar classes. Side effect phenomena reveal some specific features of very precise parts of texts. So, according to the way the classification is carried out, different grains and structure levels can be reached. The main remaining problem is to put relevant texts together within a corpus to highlight a desired level of detail or some desired comparisons between texts and so, experts' knowledge particularities.

References

Charniak, E. (1993). *Statistical language learning.* Bradford books. The MIT Press, Cambridge, Mass.

Fall, S., Crawford, T., Souders, S., and Rabin, M. (1989). Automated knowledge acquisition technics for intelligence analysts. *Applications of Artificial Intelligence*, 1095 of SPIE:66–77.

Gaines, B. and Shaw, M. (1994). Using knowledge acquisition and representation tools to support scientific communities. In *Proceedings of the Twelve National Conference on Artificial Intelligence*, volume 1, pages 707–712. AAAI, AAAI Press/The MIT Press.

Hearst, M. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistic*, Las Cruces, NM.

Jacobs, P. (1993). Using statistical methods to improve knowledge-based news categorization. *IEEE Expert*, 8(2):13–23.

Kaufman, K., Michalsky, R., and Kershberg, L. (1991). Knowledge extraction from database: design principles of the INLEN system. In *Proceedings of the 6th International Symposium on Methodology for Intelligent Systems*, number 542 in LNCS, pages 152–161, Berlin. Springer-Verlag.

Lapalut, S. (1995). Text clustering to support knowledge acquisition from documents. Technical Report RR-2639, INRIA U.R. de Sophia Antipolis, BP 93, 06902 Sophia Antipolis Cedex. ftp://ftp.inria.fr/INRIA/publication/RR/RR-2639.ps.gz.

Lapalut, S. (1996a). How to handle multiple expertise from several experts: a general text clustering approach. In Maurer, F., editor, *2nd Knowledge Engineering Forum*, number SFB 501 Bericht 01/96, Kaiserslautern University, Germany.

Lapalut, S. (1996b). Text Clustering to Help Knowledge Acquisition from Documents. In Shadbolt, N., O'Hara, K., and Schreiber, G., editors, *Advances in Knowledge Acquisition, 9th European Knowledge Acquisition Workshop, Nottingham, U.K.*, number 1076 in LNCS, pages 115–130. Springer-Verlag.

Moulin, B. and Rousseau, D. (1992). Automated knowledge acquisition from regulatory texts. *IEEE Expert*, 7(5):27–35.

Register, M. and Kannan, N. (1992). A hybrid architecture for text classification. In *Fourth International conference on Tools with Artificial Intelligence, TAI'92*, pages 286–92, Arlington, VA, USA. IEEE Compu. Soc. Press.

Reinert, M. (1979). *Classification descendante hiérarchique pour l'analyse de contenu et traitement statistique de corpus.* PhD thesis, Université Pierre et Marie Curie, Paris 6, Paris.

Reinert, M. (1986). Classification descendante hi rarchique : un algorithme pour le traitement des tableaux logiques de grandes dimensions. In *Data Analysis and Informatics*, North-Holland, Amsterdam.

Reinert, M. (1992). *Notice du logiciel ALCESTE, version 2.0.*

Shaw, M. and Gaines, B. (1987). Kitten: Knowledge initiation and transfert tools for experts and novices. *Int. J. Man-Machine Studies*, 27:251–280.

Wu, Z., Hsu, L., and Tan, C. (1992). A survey on statistical approaches to natural language processing. technical report TRA4/92, Departement of Information Systems and Computer Science, National University of Singapore, Kent Ridge, Singapore 0511.