Semantic tagging of a corpus using the Topic Navigation Map standard

Helka Folch*,***

* Electricité de France (Division Recherche et Dévelopement) 92141 Clamart, France
** Ecole Normale Supérieure de Fontenay/Saint-Cloud UMR8503 : Analyses de corpus linguistiques, usages et traitements, 92211 Saint-Cloud, France
Helka.Folch@edf.fr

Abstract

Our work carried out as part of the Scriptorium project has confronted us with a variety of problems which shed light on important issues related to corpus architectural design, such as the definition of fine-grained textual units, extraction of relevant subsections of the corpus, and in particular linking techniques enabling , text annotation with arbitrary and at times conflicting meta-data. The need for greater flexibility and expressive power for our annotation schemes has led us to apply linking techniques as defined in HyTime (ISO 10744:1992) and Topic Maps (ISO13250). We have used these techniques in particular to construct a semantic map of the corpus which enables hypertext navigation in terms of the topics inductively acquired through text-mining software. Navigation is aided by a 3D geometric representation of the semantic space of the corpus.

1. What is Scriptorium ?

Scriptorium (Lahlou et al., 1995a), is a project developed in the Research & Development Division of *EDF* (*Electricité de France*) in collaboration with *ENS* (*Ecole Normale Supérieure)*. The aim of this project is to identify prominent and emerging topics from the automatic analysis of the discourse of the company's (EDF) different social players (managers, trade-unions, employees, etc. by way of textual data analysis methods. The corpus under study in this project has 8 million words and is very heterogeneous (it contains book extracts, corporate press, union press, summaries of corporate meetings, transcriptions of taped trade union messages, etc. All documents are SGML tagged following the TEI (Dunlop, 1995) recommendations. Each document is provided with a descriptive header and is segmented into minimal textual units or chunks (which correspond to paragraphs).

Scriptorium is a modular architecture which provides an open framework where different text-mining software can be plugged in. At present the following software has been used : ALCESTE (Reinert, 1987) and ZELLIG (Habert et al. 1999). The results of these text mining tools are integrated into the architecture as structured annotation layers pointing at the relevant locations in the corpus.

2. Current approaches in semantic annotation

Increasing availability of textual data in electronic form has made semantic access to document collections a crucial issue. Semantic annotation of corpora is mainly performed today on the basis of pre-existing categories which are defined in external resources and then projected onto the corpus. These resources include lexical resources such as dictionaries, thesauri or domain-specific terminologies and conceptual resources such as general-purpose or application-specific ontologies.

Work in the field of Word Sense Disambiguation (WSD) for instance, exploits lexical resources with the aim of selecting the relevant meaning of a word in a given context from a pre-established list of all the word's potential meanings. Reducing ambiguity by tagging each word of a corpus with its contextual meaning is then exploited in Information Retrieval (IR) to increase the precision of the results generated in response to a query.

Semantic ontologies and networks are also exploited in view of enhancing navigation or IR performance. To date, the most widespread lexical semantic network is WordNet (Miller et al., 1990), developed at Princeton University since 1985. The nodes of this network are lists of synonyms or *synsets*. Each synset represents a meaning which can be lexicalized through a set of synonymous words. A meaning is defined differentially in terms of the relations a given synset has with other synsets of the network. The network is partitioned into different areas which correspond to parts of speech (POS) categories (nouns, adjectives, verbs). Different types of relations structure the network which depend on the POS category. For instance, hyponomy relations are defined for nouns, antonomy for adjectives, implication for verbs, etc. The network is essentially conceptual as the relations are established between meanings and not between lexical items. Corpus annotation with WordNet categories has been used to optimize navigation and IR, in particular to expand queries by generating synonyms that can be inferred through the relations of the network. This is aimed at improving recall in response to a query.

Semantic annotation can also be performed on the basis of an application-specific conceptual model. Given the growing merging between XML and database technologies and the increasing use of XML as a standard format for exchanging data among heterogeneous computer systems, document markup is used for annotating metadata whose structure is described in an external model comparable to a database schema. This model specifies the entities and relations of a given application domain. In this approach too, semantic annotation is based on a pre-defined categorization scheme, which is in this case similar to the fields and records of a database table.

In contrast to the approaches described above, the focus is different within the context of document-based

systems aimed at technology watch (or social watch as is the case for Scriptorium), given that the aim is not to access the corpus by means of a pre-defined list of topic categories, but to identify emerging topics in the corpus.

Furthermore, the nature of these topics is also different, as they are defined inductively by text mining tools based on data-driven methods. They are in consequence contextual and endogenous to a given corpus.

It is essential for text mining software to run on homogeneous corpora in order to yield relevant results. As we have pointed out above, our document collection is extremely heterogeneous. The concepts underlying the documents cannot be represented in one single semantic space. Therefore we do not attempt to build a semantic representation of the entire corpus through one only analysis. Our approach therefore consists in building sub-corpora of exploitable size which are homogeneous with respect to a given parameter. To this aim, we use an extractor developed using the XML Python libraries which dynamically assembles subsets of text chunks in response to a query. This extractor runs queries concerning the descriptive parameters stored in each document's header as well as fulltext searching constraints. These sub-corpora are then analyzed by the text mining tools, which at present include ALCESTE and ZELLIG. In the following two sections we show how we construct semantic classes with ALCESTE and exploit them to construct a navigable topic map overlaid on the corpus.

3. Generating semantic classes with ALCESTE

ALCESTE is a textual data analysis software developed by M. Reinert at CNRS. This program performs a classification of the text chunks of a given corpus in order to produce classes of statistically related chunks. The notion of a chunk corresponds to that of an elementary textual sequence of variable size. In our case we have chosen the paragraph, as that is our basic segmentation unit.

Similarity of text chunks is based on a distributional criteria. It depends on the number of word stems that chunks have in common, taking into account the frequency of occurrence of the words in the corpus.

Classification is based on a hierarchical descending clustering algorithm in which successive dichotomies are carried out using the first axis of a factor analysis. Details on the algorithm can be found in (Reinert, 1985).

The important point is that the method is contrastive, in other words a class of chunks is formed in opposition to the rest, not in terms of an absolute criteria.

After these classes have been determined, the program seeks the list of words most characteristic or specific to each class according to a chi2 metric. This also produces a geometric representation of classes and words in the space, based on the axis yielded by the factor analysis. We use the coordinates of the classes and words on the first three axis to obtain a 3D representation of the semantic space of a given ALCESTE analysis.

The resulting ALCESTE classes reveal underlying representations or concepts which are lexicalized through a set of related chunks and a characteristic vocabulary. These classes cannot be interpreted as absolute semantic categories but rather as points of view or semantic contexts relative to the sub-corpus under study. More on the interpretation of ALCESTE classes in (Lahlou, 1995b).

4. The Topic Navigation Map Standard

Topic Navigation Maps (TNM) is an international standard (ISO/IEC 13250) providing a language expressed in SGML and HyTime to construct a layer of topics and relations aimed at classifying and semantically annotating a collection of documents. A topic map is implemented as a set of independent links. A topic map can be compared to a semantic network overlaid on a pool of textual data. The important point is that there is a separation between the textual data and the topic map structure. As an independent structure, a topic map can be associated to different textual resources; likewise, and more relevant to our aims, different topic maps can be layered over the same document collection providing thus multiple views on the same corpus. A topic map structure is optimized for navigation, but in contrast to HTML provides the mechanisms for explicitly defining a semantic model for this structure, aimed at for instance, describing the relation expressed by a link.

The main element of a Topic Map is a topic which is an independent multi-headed link pointing to a set of occurrences. A topic link groups all textual occurrences (of any granularity) which are related to the topic. A topic has a topic type (or multiple topic types), which correspond to link types in HyTime.

A topic occurrence is any addressable textual unit in the document collection. Topic occurrences can be described by the role they play. This corresponds to the facility provided in HyTime for describing the role of anchors in independent links. Occurrences can then be grouped in relation to a common role.

The topic map standard allows the description of relationships between topics by way of a construct called a topic association. A topic association is itself an independent link which specifies relationships between topics. Topic associations can also be typed thus allowing an explicit description of the relation and the definition of relation taxonomies.

An important characteristic of a topic is its scope. The scope defines the limit of validity of a topic. In other words, the characteristics associated to a certain topic (its name, its occurrences, its type) are only valid within a certain context. The limits of validity are defined by the scope. The scope is defined explicitly by a topic or group of topics.

Another construction defined in the TNM standard is a facet. Facets provide a mechanism to associate attribute-value pairs with to data chunks. This can include properties such as date, accessibility, affiliation, etc. Facets are properties that can be projected onto textual data chunks and are orthogonal to the topic map, as the

facets a given chunk exhibits are independent of the topics it is associated to. They can be used to filter chunks having a given value for a given property and to create restricted subsets of chunks in response to a query. Facets are also implemented as independent links.

Next we will describe how we use this formalism to semantically tag our corpus with the categories inductively acquired by ALCESTE.

5. Construction of a navigable topic map layer over the corpus

In our model, we define ALCESTE classes as topics. The name of the topic is attributed manually as labelling an ALCESTE class requires an interpretation process aimed at finding an underlying concept which can be characterized by the list of typical chunks and specific vocabulary of the class in question. As described in the preceding section, an ALCESTE class denotes a representation or concept which is lexicalized through a set of related chunks and a characteristic vocabulary. Thus, the topic representing a class assembles all the occurrences of text chunks specific to the class. It also groups pointers towards the occurrences of the words that constitute the characteristic vocabulary of the class. The two groups of occurrences are distinguished by the role they play, the role of the first group is typed as "typical chunks" , the role of the second group is typed as "typical words".

As we have already mentioned, the semantic classes produced by ALCESTE are only valid within the context of the sub-corpus under study. We have used the notion of scope as defined in the Topic Maps standard to express this. In our application, the scoping topic of a semantic class is the sub-corpus analyzed by ALCESTE. We have, in consequence, defined another type of topic, namely "sub-corpus", which is a link pointing at all the chunks which have been extracted from the primary text collection and compose the sub-corpus under analysis. It is this topic that is used to define the scope of any semantic class. The name of a sub-corpus topic is the extraction query that has been used to build the given sub-corpus.

Another building block of the Topic Map Standard which we use in our application, is that of a facet. We use facets to project external properties on chunks of the corpus. Such properties include information such as affiliation, date, division or department within the company, etc. They are pre-existing classification schemes as opposed to the inductive information (ALCESTE classes) which composes the topic map. The schemes themselves are coded as facets in our model, whilst the value codes of these schemes (a particular date, for instance) are coded as facet values.

5.1. Application interface

We have used the python XML libraries and *tmproc*, an implementation of python classes for topic map processing to construct this topic map structure in the XML interchange format. We also produce an implementation which simulates HyTime linking facilities in an HTML format enabling navigation of the topic map in an ordinary browser (Netscape or Internet Explorer).

The elements of the topic map (ALCESTE classes, sub-corpora) provide semantic access to the corpus. Navigation can start from a word index or a list of ALCESTE classes. Clicking on an ALCESTE class, for instance, initiates traversal among the occurrences of the 'typical chunks' of that class. A list of occurrences (of either 'typical words' or 'typical chunks') thus define a navigation path through the corpus.

Each chunk is at the intersection of several navigation paths, as it can be a typical occurrence of different classes resulting from the analysis of different sub-corpora.

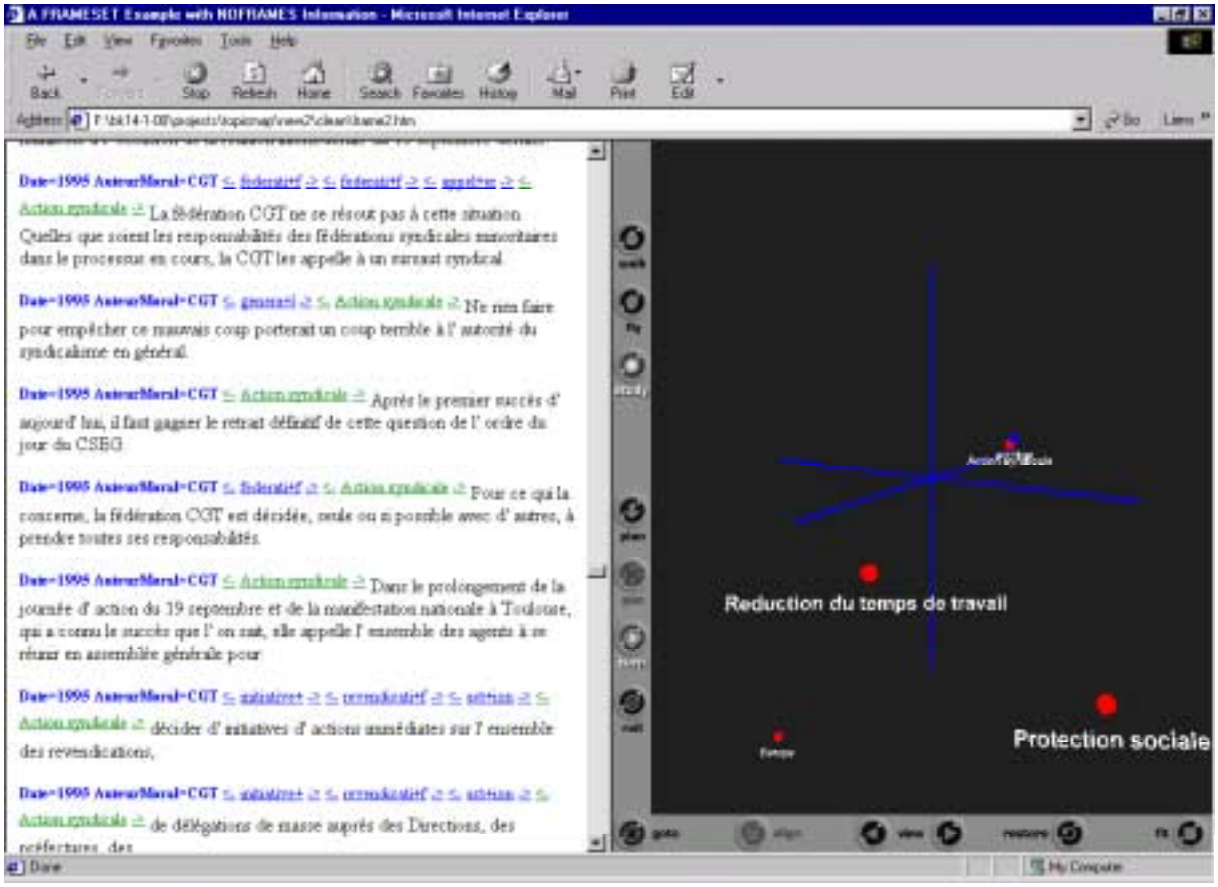

Figure 1: Application interface

Figure 1 shows an extract from the corpus, where each chunk is preceded by a header containing, at the left, the facet values associated to the chunk, next, the typical word occurrences contained in the chunk and (in green) the different ALCESTE classes associated to the class. Each class is defined in a different scope.

The left half of the browser window can be described as the "navigation compass". It is a VRML scene which gives a 3D representation of the semantic space of a sub-corpus. The coordinates of the objects in this space (classes are represented as red spheres, 'typical words' as blue spheres and 'typical chunks' as green spheres) are calculated by ALCESTE as described in the preceding section.

6. Conclusion

We are exploring issues related to content-based access of very large, heterogeneous corpora and the semantic tagging of these corpora with inductively-acquired categories. The problems encountered are very different from those posed by semantic annotation of corpora on the basis of pre-defined categorization schemes.

We use statistical methods to inductively acquire semantic classes from the corpus. However due to the contrastive nature of these methods, the relevance of their results are extremely dependent on the homogeneity of the data on which they operate. Our approach consists of constructing homogeneous sub-corpora by extracting text chunks relevant to a certain criteria, and running the textual analysis software (in particular ALCESTE) on them. The semantic classes produced by this software are then relative to the sub-corpus under analysis.

We have used the Topic Navigation Map standard to construct a navigable semantic map of the global corpus from the results of this statistical software. We believe this standard is flexible for the following reasons. Firstly, TNM allows structuring multiple views over the same textual pool; this is adapted to modeling concepts underlying heterogeneous corpora, which cannot be represented in one single semantic space. Secondly, TNM allows the possibility of explicitly describing the relations between text chunks. Thirdly, the notion of "scope" in the TNM standard provides a mechanism to model the contextual nature of inductively-acquired semantic classes. Fourthly, TNM provide a way of articulating inductively acquired categories with information external to the corpus (pre-existing classification schemes) through the notion of 'facet'. Finally, the fact that every construct in the TNM standard can be implemented as an independent link, greatly simplifies software design.

7. References

Dunlop, D. (1995). *Practical considerations in the use of TEI headers in large corpora*. Computers and the Humanities, (29), 85–98. Text Encoding Initiative. Background and Context, edited by Nancy Ide and Jean Véronis.

Habert, B., Fabre, C. (1999). *Elementary dependency trees for identifying corpus-specific semantic classes*. Computers and the Humanities.

Lahlou, S., Aubert, C., Piat, G. (1995a). *Scriptorium : le projet*. EDF-Direction des Etudes et Recherches. HN 51/95/010.

Lahlou, S. (1995b). *La construction du sens dans l'analyse statistique de données textuelles*. EDF-Direction des Etudes et Recherches. HN 51/95/012.

Marchortorchino, F. (1989). *Liaison analyse factorielle – Analyse relationnelle. (I) : Dualité 'Burt-Condorcet'*. IBM-France, F 142, Paris.

Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K. *Introduction to WordNet : An on-line lexical database*. Journal of Lexicography, 3, 1990, 235-244.

Reinert, M. (1985). *Classification descendante hierarchique : Un algorithme pour le traitement des tableaux logiques de grandes dimensions.* 4ème journées internationales 'Analyse de données et informatique'. INRIA 1985.

Reinert, M. (1987). *Un logiciel d'analyse de données textuelles 'ALCESTE'*. 5ème journées internationales 'Analyse de données et informatique'. INRIA 1987.