Knowledge Acquisition from Texts : Using an Automatic Clustering Method Based on Noun-Modifier Relationship

Houssem Assadi
Electricité de France - DER/IMA and Paris 6 University - LAFORIA
1 avenue du Général de Gaulle, F-92141, Clamart, France
houssem.assadi@der.edfgdf.fr

Abstract

We describe the early stage of our methodology of knowledge acquisition from technical texts. First, a partial morpho-syntactic analysis is performed to extract "candidate terms". Then, the knowledge engineer, assisted by an automatic clustering tool, builds the "conceptual fields" of the domain. We focus on this conceptual analysis stage, describe the data prepared from the results of the morpho-syntactic analysis and show the results of the clustering module and their interpretation. We found that syntactic links represent good descriptors for candidate terms clustering since the clusters are often easily interpreted as "conceptual fields".

1 Introduction

Knowledge Acquisition (KA) from technical texts is a growing research area among the Knowledge-Based Systems (KBS) research community since documents containing a large amount of technical knowledge are available on electronic media.

We focus on the methodological aspects of KA from texts. In order to build up the model of the subject field, we need to perform a corpus-based semantic analysis. Prior to the semantic analysis, morpho-syntactic analysis is performed by LEXTER, a terminology extraction software (Bourigault et al., 1996) : LEXTER gives a network of noun phrases which are likely to be terminological units and which are connected by syntactical links. When dealing with medium-sized corpora (a few hundred thousand words), the terminological network is too voluminous for analysis by hand and it becomes necessary to use data analysis tools to process it. The main idea to make KA from medium-sized corpora a feasible and efficient task is to perform a robust syntactic analysis (using LEXTER, see section 2) followed by a semi-automatic semantic analysis where automatic clustering techniques are used interactively by the knowledge engineer (see sections 3 and 4).

We agree with the differential definition of semantics : the meaning of the morpho-lexical units is not defined by reference to a concept, but rather by contrast with other units (Rastier et al., 1994). In fact, we are considering "*word usage rather than word meaning*" (Zernik, 1990) following in this the distributional point of view, see (Harris, 1968), (Hindle, 1990).

Statistical or probabilistic methods are often used to extract semantic clusters from corpora in order to build lexical resources for ANLP tools (Hindle, 1990), (Zernik, 1990), (Resnik, 1993), or for automatic thesaurus generation (Grefenstette, 1994). We use similar techniques, enriched by a preliminary morpho-syntactic analysis, in order to perform knowledge acquisition and modeling for a specific task (e.g. : electrical network planning). Moreover, we are dealing with language for specific purpose texts and not with general texts.

2 The morpho-syntactic analysis : the LEXTER software

LEXTER is a terminology extraction software (Bourigault et al., 1996). A corpus of French texts on any technical subject can be fed into it. LEXTER performs a morpho-syntactic analysis of this corpus and gives a network of noun phrases which are likely to be terminological units.

Any complex term is recursively broken up into two parts : head (e.g. PLANNING in the term REGIONAL NETWORK PLANNING), and expansion (e.g. REGIONAL in the term REGIONAL NETWORK) [1].

This analysis allows the organisation of all the candidate terms in a network format, known as the

[1] All the examples given in this paper are translated from French.

"terminological network". Each analysed complex candidate term is linked to both its head (H-link) and expansion (E-link).

LEXTER also extracts phraseological units (PU) which are "informative collocations of the candidate terms". For instance, CONSTRUCTION OF THE HIGH-VOLTAGE LINE is a PU built with the candidate term HIGH-VOLTAGE LINE. PUs are recursively broken up into two parts, similarly to the candidate terms, and the links are called H'-link and E'-link.

3 The data for the clustering module

The candidate terms extracted by LEXTER can be NPs or adjectives. In this paper, we focus on NP clustering. A NP is described by its "terminological context". The four syntactic links of LEXTER can be used to define this terminological context. For instance, the "expansion terminological context" (E-terminological context) of a NP is the set of the candidate terms appearing in the expansion of the more complex candidate term containing the current NP in head position. For example, the candidate terms (NATIONAL NETWORK, REGIONAL NETWORK, DISPATCHING NETWORK) give the context (NATIONAL, REGIONAL, DISPATCHING) for the noun NETWORK.

If we suppose that the modifiers represent specialisations of a head NP by giving a specific attribute of it, NPs described by similar E-terminological contexts will be semantically close. These semantic similarities allow the KE to build conceptual fields in the early stages of the KA process.

The links around a NP within a PU are also interesting. Those candidate terms appearing in the head position in a PU containing a given NP could denote properties or actions related to this NP. For instance, the PUs LENGTH OF THE LINE and NOMINAL POWER OF THE LINE show two properties (LENGTH and NOMINAL POWER) of the object LINE; the PU CONSTRUCTION OF THE LINE shows an action (CONSTRUCTION) which can be applied to the object LINE.

This definition of the context is original compared to the classical context definitions used in Information Retrieval, where the context of a lexical unit is obtained by examining its neighbours (collocations) within a fixed-size window. Given that candidate terms extraction in LEXTER is based on a morpho-syntactical analysis, our definition allows us to group collocation information disseminated in the corpus under different inflections (the candidate terms of LEXTER are lemmatised) and takes into account the syntactical structure of the candidate terms. For instance, LEXTER extracts the complex candidate term BUILT DISPATCHING LINE, and analyses it in (BUILT (DISPATCHING LINE)); the adjective BUILT will appear in the terminological context of DISPATCHING LINE and not in that of DISPATCHING. It is obvious that only the first context is relevant given that BUILT characterises the DISPATCHING LINE and not the DISPATCHING.

To perform NP clustering, we prepared two data sets : in the first, NPs are described by their E-terminological context; in the second one, both the E-terminological context and the H'- terminological context (obtained with the H'-link within PUs) are used. The same filtering method [2] and clustering algorithm are applied in both cases.

Table 1 shows an extract from the first data set. The columns are labelled by the expansions (nominal or adjectival) of the NPs being clustered. Each line represents a NP (an individual, in statistical terms) : there is a '1' when the term built with the NP and the expansion exists (e.g. REGIONAL NETWORK is extracted by LEXTER), and a '0' otherwise ("national line" is not extracted by LEXTER).

	NATIONAL	DISPATCHING	REGIONAL
LINE	0	1	0
NETWORK	1	1	1

Table 1: example of the data used for NP clustering

In the remainder of this article, we describe the way a KE uses LEXICLASS to build "conceptual fields" and we also compare the clusterings obtained from the two different data sets.

4 The conceptual analysis : the LEXICLASS software

LEXICLASS is a clustering tool written using C language and specialised data analysis functions from Splus^{TM} software.

Given the individuals-variables matrix above, a similarity measure between the individuals is calculated [3] and a hierarchical clustering method is performed with, as input, a similarity matrix. This kind of methods gives, as a result, a classification tree (or dendrogram) which has to be cut at a given level in order to produce clusters. For example, this method, applied on a population of 221 NPs (data set 1) gives

[2] This filtering method is mandatory, given that the chosen clustering algorithm cannot be applied to the whole terminological network (several thousands of terms) and that the results have to be validated by hand. We have no space to give details about this method, but we must say that it is very important to obtain proper data for clustering

[3] similarity measures adapted to binary data are used - e.g. the Anderberg measure - see (Kotz et al., 1985)

21 clusters. figure 1 shows an example of such a cluster.

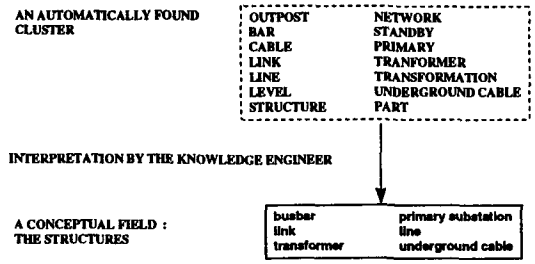

Figure 1: a cluster interpretation

The interpretation, by the KE, of the results given by the clustering methods applied on the data of table 1 leads him to define *conceptual fields*. Figure 1 shows the transition from an automatically found cluster to a conceptual field : the KE constitutes the conceptual fields of "the structures". He puts some concepts in it by either validating a candidate term (e.g. LINE), or reformulating a candidate term (e.g. PRIMARY is an ellipsis and leads the KE to create the concept `primary substation`). The other candidate terms are not kept because they are considered as non relevant by the KE. The conceptual fields have to be completed all along the KA process. At the end of this operation, the candidate terms appearing in a conceptual field are validated. This first stage of the KA process is also the opportunity for the KE to constitute synonym sets : the synonym terms are grouped, one of them is chosen as a concept label, and the others are kept as the values of a generic attribute labels of the considered concept (see figure 2 for an example).

```
line
        //conceptual field// : structure
        //type// : object
        //labels// : LINE, ELECTRIC LINE,
        OVERHEAD LINE
```

Figure 2: a partial description of the concept "line"

5 Discussion

- Evaluation of the quality of the clustering procedure : in the majority of the works using clustering methods, the evaluation of the quality of the method used is based on recall and precision parameters. In our case, it is not possible to have an *a priori* reference classification. The reference classification is highly domain- and task-dependent. The only criterion that we have at the present time is a qualitative one : that is the usefulness of the results of the clustering methods for a KE building a conceptual model. We asked the KE to evaluate the quality of the clusters, by scoring each of them, assuming that there are three types of clusters :
  1. Non relevant clusters.
  2. Relevant clusters that cannot be labelled.
  3. Relevant clusters that can be labelled.

  Then an overall clustering score is computed. This elementary qualitative scoring allowed the KE to say that the clustering obtained with the second data set is better than the one obtained with the first.
- LEXICLASS is a generic clustering module, it only needs nominal (or verbal) compounds described by dependancy relationships. It may use the results of any morpho-syntactic analyzer which provides dependancy relations (e.g. verb-object relationship).
- The interactive conceptual analysis : in the present article, we only described the first step of the KA process (the "conceptual fields" construction). Actually, this process continues in an interactive manner : the system uses the conceptual fields defined by the KE to compute new conceptual structures; these are accepted or rejected by the KE and the exploration of both the terminological network and the documentation continues.

References

Bourigault D., Gonzalez-Mullier I., and Gros C. 1996. Lexter, a Natural Language Processing Tool for Terminology Extraction. In *Proceedings of the 7th Euralex International Congress*, Göteborg, Sweden.

Grefenstette G. 1994. Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, Boston.

Harris Z. 1968. Mathematical Structures of Language. Wiley, NY.

Hindle H. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, Pennsylvania. Association for Computational Linguistics, Morristown, New Jersey.

Kotz S., Johnson N. L., and Read C. B. (Eds). 1985. Encyclopedia of Statistical Sciences. Vol.5, Wiley-Interscience, NY.

Rastier F., Cavazza M., and Abeillé A. 1994. Sémantique pour l'analyse. Masson, Paris.

Resnik P. 1993. Selection and Information : A Class-Based Approach to Lexical Relationships. PhD Thesis, University of Pennsylvania.

Zernik U. 1993. Corpus-Based Thematic Analysis. In Jacobs P. S. Ed., *Text-Based Intelligent Systems*. Lawrence Erlbaum, Hillsdale, NJ.