Evaluating Natural Language Processing Systems as a Tool for Building Terminological Databases

Abstract: Natural language processing systems use various modules in order to identify terms or concept names and the logico-semantic relations they entertain. The approaches involved in corpus analysis are either based on morpho-syntactic analysis, statistical analysis, semantic analysis, recent connexionist models or any combination of two or more of these approaches. This paper will examine the capacity of natural language processing systems to create databases from extensive textual data. We are endeavouring to evaluate the contribution of these systems, their advantages and their shortcomings.

1. Introduction

Setting up the conceptual design of a database (object-oriented or relational) is not a simple task for the designer. He or she has to examine a new field of knowledge i.e. analyzing an extensive textual data (e.g. analyzing users’ needs and profiles, interviews with field specialists, term glossaries, technical documentation, etc.). The designer may not be acquainted with the representation of the field, its structures and the articulations between its objects.

To make the designer’s task easier, natural language processing systems can be of help particularly those dedicated to the identification of terms or concepts names related to a specific field of knowledge (construction of a reference terminology) and the logico-semantic relations they entertain. These systems can be applied to the modelling and designing of the following types of systems:

- The modelling of an object-oriented database design (static aspects: i.e. describing the structure),
- Knowledge-based systems: modelling the hierarchies between classes and the relations between the objects concerned by a set of rules,
- Modelling the conceptual design of a relational database (domains, relations, coherence maintenance);
- Thesaurus construction (documentary databases, ...),
- Terminological database construction, etc.

This paper will examine the capacity of natural language processing (NLP) systems to help the designers in creating databases from extensive textual data. The scope of our study will be limited to terminological databases.

NLP systems we are evaluating use various modules in order to identify terms or concept names and the logico-semantic relations they entertain. Most of these systems need, in addition, a general language dictionary, a glossary of technical terms covering the relevant field, etc. The identification of terms is in fact an extraction of noun phrases corresponding to

the concepts representing the field of knowledge. In their current state, these systems are mostly semi-automatic processing tools.

We are presenting in part two a state of the art of terminology building tools and a brief account of the model we are proposing. Then we review the systems proposed for the evaluation from a linguistic point of view i.e. the underlying linguistic models. The comparison is continued with the typology of systems from the output point of view. At this point we discuss the possibility of integrating different models in order to have better systems output i.e. ideally terms and semantic relations they entertain with each other. We compare the capacities of the tested systems and SEEK and we show in what respect SEEK can be a complementary module to syntactically and statistically-based systems. In section three we propose a paradigm of systems evaluation based upon the following two criteria: digital evaluation and qualitative evaluation.

2. Scientific Background

2.1 State of the Art

We can distinguish several areas in which terminology develops. The most traditional field of application is that of translation. The second area is that of monolingual activities which are generally prescriptive. (Wijnands, 1993). The third field of terminology is documentation and information science (respectively thesaurus descriptors whose role is to describe documents, knowledge acquisition tools i.e. terms representing a conceptual field, etc.). This field of application reveals the importance of terminology as a means for information retrieval i.e. domain knowledge can be acquired by repeated information retrieval (Kukulska, 1993). Ideally a terminological database or terms bank should be a multi-functional tool answering the needs of a wide range of users: information systems designers, technical writers, knowledge-based systems designers, specialized translators technical and scientific dictionaries authors. The question of potential users of these tools is discussed in (Sager, 1990, 169).

If we focus on the tools devoted to terminology building we can distinguish two types: conventional Term Banks and Terminological Database Management Systems. The questions a specialist can address to this type of data model (conventional term bank) can be on the gender, the spelling, the equivalent, definition, subject label, synonym, contextual usage of a term, etc. They are in principle the same ones that a specialist would address to a general language dictionary. In addition to the type of the questions we listed above, some retrieval capabilities are now available in some term banks: compilation of all terms used in a domain, definitions and conceptual links, subordinate/partitive terms, the name of the terminologist who entered a group of terms over a period of time (Sager, 1990, 168).

In addition to providing terms to describe concepts in specialized domains terminology is a means of representing all the knowledge relating to a particular term (linguistic and conceptual information). Moreover, computing techniques such as concordancing can be used to explicit information from texts such as word frequencies, collocations, etc. thereby enabling terminologists to assess the extent to which different levels of language are needed. Collocations in terminology leads to phrases and phraseology: it is only recently that most advanced banks such as TEAM and DANTERM (see Thomas, 1993) have begun to include these elements.

The new trends are represented by term banks such as COGNITERM (University of Ottawa) a hybrid between a conventional and a knowledge-base system (in the sense known in
artificial intelligence. COGNITERM is a concept-oriented and multidimensional handling tool (see Meyer, 1993; Meyer and Bowker, 1992).

2.2. Terminological Database Design Proposed

Assuming the theoretical bases of terminology (see Dahlberg, 1978, 1981; Jouis and Mustafa, 1995) our approach to terminology is a concept-oriented one and it will be taking into consideration the state-of-the-art in terminological databases. The database structure we are proposing consists of records. In each record we describe a term related to a certain field of knowledge. In each description hypertext links point towards textual areas in the text as well as towards labelled relations. The database layout is more precisely shown in fig. 2. It will consist of basic entries: term, definition(s), contextual definition(s), conceptual relations underlying terms. The latter will be presented in a semantic-network fashion (see below), hypertext links pointing to contextual definitions, a set of pointers oriented towards other records of the database pinpointing to semantic relations (see fig. 2).

![Diagram](image_url)

**Fig.1:** Terminological database, hypertext pointers and labelled links

The typology of the semantic relations we identified in our model are based on the Applicative and Cognitive Grammar (AGC³). ACG articulates several levels of representations especially a cognitive level in which the meanings of linguistic units may be analyzed under the form of layouts (semantic-cognitive representations) in order to constitute the knowledge representations associated to a given text. ACG proposes a set of semantic concepts which defines an organized system of meanings. We distinguish the semantic types of linguistic units, fundamental static relations and dynamic relations (movement, change of state, conservation of a movement, iteration, intensity, variation, constraints, causes . . .).

![Diagram](image_url)

**Fig.2:** A record layout

2.3 Evaluated Systems Typology within the Context of Linguistic Perspectives

We can distinguish three categories:

(1) The first category consists of syntactically-based systems, to which statistical modules are sometimes added. These systems use a set of grammatical words and punctuation symbols in order to isolate terms in the textual data. The following systems belong to this category:
(a) **ACABIT** is a prototype geared to automatic terminological bank construction. Its task is to help an expert in a certain field of knowledge by proposing a set of potential candidate terms in a pre-defined morpho-syntactic format. The statistical module role is to rank the produced terms according to their relevance to the field (see Daille 1995).

(b) **KES** (Knowledge Extracting and Structuring) is a set of tools developed within the framework of a European research program (GRAAL) a consortium including three shareholders: GSI-ERLI (a French language engineering firm), Aérospatiale (French Aircraft Company) and EDF-DER (French Electric Company). KES's job is to extract and structure knowledge of a given corpus of text thereby providing a structured data from a non-structured one (see Ogononowski 1994).

(c) **LEXTER** is a system used in extracting potential candidate terms related to a field of knowledge. The system is based on an endogenous learning procedure enabling it to automatically acquire syntactic information of sub-categorization. Each term is grammatically analyzed into head and modifier in order to build a lexico-syntactic network. A further step is knowledge modelling which can be undertaken by a terminologist or a knowledge engineer. The latter can exploit this network in order to reconstruct the underlying conceptual network (see Bourigault 1992).

(2) The second category consists of pure statistical and/or connexionist models:

(a) **ANA** identifies single terms and characteristic expressions of a field of knowledge and their organization in a non-labelled network. The system is based on the following postulate "Frequent co-occurring events are significant" (Enguehard 1993).

(b) **CONTERM** is based on a connexionist model which applies a neuronal network to textual chunks in order to build classes of networks of terms linked to a major one. By so doing CONTERM tries to identify distinct semantic networks for each identified term.

(3) The third category is the one including semantically based systems such as **SEEK**, **IOTA** and **NOEMIC**. These systems focus particularly on semantic relation.

(a) **SEEK**, for instance, uses a Contextual Exploration Approach, a method based on the linguistic knowledge of a reader when he explores a text in a given language. The system is capable of identifying concepts and concept relations (see Jouisi 1994; 1995). SEEK can be considered as the most atypical system for the following reasons: (i) Most natural language systems use parsers, general language dictionaries, etc.; (ii) Most of the systems we are testing use intensive resources compared to a contextual exploration-based system like SEEK which needs only a limited number of linguistic markers to do the same job.

If we focus on these systems in order to see how they exploit the linguistic resources we can say that syntactically and statistically based systems use a set of grammatical words and punctuation symbols in order to isolate terms in the textual data. Contextual exploration approach, unlike these systems, considers the contents of this set (i.e. grammatical words and punctuation) as meaningful elements hence their correspondence to the linguistic knowledge independently of a specific subject-field. Their role is to convey meaning and allow the designer of the database to locate concepts and the relations they hold between them in an interactive fashion. The system does not need either a general language dictionary, a parser or a glossary of terms. This is why we consider SEEK as a fundamentally atypical system. This system is capable of identifying concepts and concept relations.

We assume that contextual exploration is a strategy that an analyst uses while examining textual data: he or she builds up a data model by observing certain linguistic indicators or markers. A semantic analysis using contextual exploration is presented as a knowledge-based system. The model based on linguistic knowledge is chosen for its capabilities in analyzing concepts and/or terms.
(b) IOTA is a system based on a Sowa conceptual graph model. Its purpose is identifying noun phrases in order to index documents from a textual data. It also dynamically builds an automatic thesaurus used for information retrieval (see Bruandet 1980).

(c) As for NOEMIC, it is based on a componential analysis model which accounts for a variety of typed semantic relations (thesaurus relations). NOEMIC does not need a 'vertical' linguistic analysis (i.e. morphological, syntactic). For further information (see Penel 1994).

2.4. Systems Typology from the Output Point of View

The evaluation of these systems raises a series of questions that we should answer: The first question we will try to answer is: to what extent do natural language systems manage to adequately and efficiently identify automatically or semi-automatically terms/concepts and concept relations that can be used to build information retrieval tools? As input the systems will get textual data from specific fields of knowledge (manuals, specialized articles, newspaper articles, dispatches, etc.). The output produced by these systems can be divided into two categories: (i) lists of candidate terms, eventually ranked by frequency of occurrence, (ii) semantic relations. We can already observe that morpho-syntactic methods combined with statistical modules produce lists of terms without significant semantic relations while semantically-based systems tend to focus on semantic relations resulting in rather limited terms productivity. The idea is to tend towards the integration of systems (a hybridization of models or cross-breeding) in order to enhance the capability of these systems. We are putting forward the idea of integrating two or more systems within the framework of this research: integrating SEEK, for instance, to a syntactically and statistically-based system.

2.5. Corpus

The second question deals with the criteria of choosing adequate corpuses. The elaboration of terminology includes the analysis of documents of different types and varying sizes. (textbooks, technical manuals and specialized dictionaries, transcripts of interviews with field experts, etc.). The "ideal" corpuses to efficiently identify terms and relations which can be considered as good descriptors providing efficient access to specialized information: automatic thesaurus construction, broadening of existing thesaurus, etc. should be representative of the field of knowledge considered. Three types of texts are needed (see Ahmed, 1993): instructional (textbooks, technical manuals, encyclopedic texts); informative (learned papers, advanced treatises, interview transcripts of experts, patent documents); and imaginative (popular science material, public information material i.e. advertisement about the goods and services of the subject domain). This variety is needed because the terms in each type of corpus serve different purposes. This typology is used by a team at the University of Surrey to build term banks in about 10 different subject fields (see Ahmed 1993).

In addition to the mentioned elements we are trying to define a set of criteria for evaluating these various systems according to the following areas of performance: (i) information retrieval (recall and precision); (ii) information display (output: lists of terms, semantic networks) or any combination of two or more of these paradigms, etc.).

3. Evaluation Problem

Automatic term construction raises a series of questions: what type of terminology to create? Applied terminology (used for a certain purpose such as translation, indexing, etc.) or a
'general' terminology (used for all possible applications). The major difference between applied and general terminology is that a general terminology tends rather towards the complete coverage of a field. In other words, the whole concept structuring of a certain field of knowledge is supposed to be provided by such a terminology. It is a sort of «scholarly terminology » placing fidelity to the coherence of the relevant field above all else.

But the end-user is an important element to take into consideration when conceiving a terminology. It is possible to create a tool for specialists of a certain field (knowledge representation device), a general purpose tool, or a terminological database specially tailored for translation and technical writing purposes. In both cases the users have to master the terms used in the relevant fields and the relations they entertain. The following diagram sums up the question of systems comparison:

Two different systems S1 and S2 applied to the same corpus C1 can yield two comparable results. But if we take into consideration the purposes of S1 & S2, the nature of linguistic elements (E1 and E2) they use and the conditions of their use (end users) the comparison between the two systems would be rather difficult. For this reason it is important to organize the system in such a way that their output can be comparable. Another problem has to do with the nature of corpus provided as input (type of documents, fields of knowledge). These elements are not necessarily the same for S1 and S2. Given that, it is important to provide the systems with different types of corpus (see above).

4. Evaluation Paradigms

Comparing terms extraction tools and the logico-semantic relations they entertain with each other is not as easy task if we compare it to the work done within the framework of conventional algorithms comparisons (such as sorting algorithms, etc.). For those we normally have as input a set of totally formalized and structured data (normally digital) at our disposal. Furthermore the expected results are usually defined beforehand (sorted value lists, etc.). The purpose of 'conventional' evaluation process is not to compare results but rather to rank algorithms according to quantifiable criteria : processing time, required processing resources (usually measurement in required memory space).

In our approach the concept of evaluation is different even if we take into consideration required processing time, required systems resources (i.e. required linguistic tools, electronic dictionaries, etc.). The only really relevant evaluation criteria is the quality of the resulting terminology. So we are going to be focusing mainly on the relevance of the terminology obtained. Within this perspective we are going to put forward two different evaluation methods.
4.1 Digital Evaluation

The first way is to use digital quantification based on methods of statistical correlation i.e. correlating terms and the semantic relations they hold with each other. But this raises other theoretical questions linked with the interpretation of the resulting measurements:

(a) What is the relevance of terms and semantic relations provided by the systems being tested? Does the terminology satisfy minimum requirements? Do we need to define a minimum level of terms production?

(b) Are discrepancies meaningful? For example, it could be that most of the systems being tested are having qualitatively poor outputs, while only one or two produce worthwhile results.

(c) Finally, to undertake statistical analysis of the results it is necessary to have access to a reference terminology dealing with the relevant field of knowledge. But such a terminology has to be evaluated beforehand and only a specialist of that field can assess its relevance and exhaustivity of coverage. We have then to quantify the variations between the various results and the terminology of reference to define the degree of relevance enabling us to rank the systems being evaluated.

4.2 Qualitative Evaluation

We are putting forward a strategy based on qualitative evaluation. In other words we would rather submit the results to specialist (we are distinguishing here between specialists, terminologists and knowledge engineers). This approach supposes that manual results obtained by field specialists can represent a reference terminology. We have to admit that for the same corpus we can get different results (i.e. two different terminologies produced by two different teams of specialists). We are assuming that manual results represent a coherent class. To undertake this evaluation we distinguish two types of human expertise (a) Those who undertake evaluation, i.e. specialists who will rank systems according to the relevance of terms to a certain field;

(b) Those who will establish a reference terminology from the same corpus processed by the tested systems.

To guarantee the accuracy of the test and its neutrality the external aspect of the results (those obtained by automatic processing and the ones obtained by manual) processing should be the same.

4. Concluding Remarks

Morpho-syntactic methods combined with statistical modules produce lists of terms without significant semantic relations while semantically-based systems tend to focus on semantic relations resulting in rather limited terms productivity. The idea is to tend towards the integration of systems in order to enhance their capacity.

As far as evaluation is concerned the most relevant criteria to take into consideration is the quality of the resulting terminology. Within this perspective we are putting forward a strategy based on qualitative evaluation. The idea is to submit the results to specialist (i.e. field specialists, terminologists and/or knowledge engineers). This approach supposes that manual results obtained by field specialists can represent a reference terminology.
Notes

1. We are conducting this research within the framework of a program sponsored by the "Association des Universités Entièrement ou Partiellement de Langue Française" (AUPELF) - an international association whose mission is to promote the dissemination of French as a scientific medium. This research aims to evaluate software capabilities in automatic terminology building from corpuses in French. Softwares submitted to this evaluation are conceived by French and Canadian research institutions (National Scientific Research Centre and Universities) and/or companies: EDF (French Electric Company) among others.

2. The question we are raising which terms are best to build a terminology of reference (absolutely all the terms representing the field of knowledge) and which ones are best used for information retrieval.

3. For a complete and thorough description of ACG the reader may refer to (Descles, 1990 or Jouis, 1994) In this model we distinguish a certain number of elementary types which enable us to categorise the entities of a certain field: individualizable (i.e. separately accountable) entities, Boolean entities (expressions denoting a value of truth), massive entities, collective classes, distributive classes, places, etc.

4. Static relations permit the description of some states (static situations) related to an area of knowledge. Static situations remain stable during a certain temporal interval where neither the beginning nor the end are taken into account. We distinguish more than twenty relations, particularly: identifications (or equivalence), incomparability between two entities, dimensions (measures, etc.), cardinality, comparisons between two values, inclusion among distributive classes, belonging of one individualizable entity to another distributive class (broader/narrower terms), -relations part/whole among collective classes, localizations of one entity in one place.

5. Formerly TAIGA (see Penel 1994).

References


