Automatic Term Recognition & Extraction Tools: Examining the New Interfaces and their Effective Communication Role in LSP Discourse

Abstract: In this paper we will discuss the possibility of reorienting NLP (Natural Language Processing) systems towards the extraction, not only of terms and their semantic relations, but also towards a variety of other uses; the storage, accessing and retrieving of Language for Special Purposes (LSP) lexical combinations, the provision of contexts and other information on terms through the integration of more interfaces to terminological data-bases, term managing systems and existing NLP systems. The aim of making such interfaces available is to increase the efficiency of the systems and improve the terminology-oriented text analysis. Since automatic term extraction is the backbone of many applications such as machine translation (MT), indexing, technical writing, thesaurus construction and knowledge representation developments in this area will have a significant impact.

1. Introduction

Terminology is concerned with scientific attempts at organizing knowledge. It deals with specialized vocabularies for different fields of knowledge within and across linguistic boundaries. It contributes to LSP, translation, documentation and information science. We are witnessing the growth of new terminologies or technical vocabularies which is caused by the spread of new technologies and the constant expansion of interdisciplinary research. At the same time, LSP communication in any given field of knowledge is less and less confined to specialists of the same domain. LSP use now goes much beyond its former specialized milieus and spreads across the public service, the professions and many authors in mass media communication. Interdisciplinarity and new information technology contribute to an ever faster cross-breeding of knowledge between various subjects and in various languages. This movement is nourished by the practice of an increasing number of people who are not subject specialists but language professionals i.e. technical writers, terminographers, translators, editors, LSP teachers and journalists. From a terminological point of view, effective communication and dissemination of knowledge is now often achieved by the practice of LSP discourse. This will lead us to incorporate phraseology (we will define our use of this term in section 2) in the existing term banks, among the terminology management tools and to broaden the range of NLP recognition and extraction tools. Specialized vocabularies designed for translators consist mostly of nouns and noun phrases from which verbs are absent. This situation is rapidly changing in comparative terminology where significant efforts are being made to integrate LSP phraseologisms into specialized vocabularies as lexical solidarities or polar combinations of collocators and term bases (Heid, 1994; L'Homme, 1996b). Some LSP studies are now focusing on the role of phraseology as terminology-in-the-making, since certain types of phraseologisms can be turned into terminological units (Pavel, 1993).

The paper discusses the advantages of integrating phraseology, taken in a broad sense (lexical combinations in specialized texts, technical verbs, collocations, concordancing) as interfaces to existing term banks, to terminology management tools as well as to NLP term
2. The Role of Phraseology in LSP Discourse

The problem is to define what constitutes effective communication via LSP discourse. Some authors believe that if LSP exists only as a tool of communication it needs a conventional vocabulary and an unequivocal definition of concepts in order to be effective. This is a trend illustrated by most of the concept-oriented term-extraction-and-construction approaches. Other authors believe that effective communication means clarity of the message and its transmission by all means available; terms, phraseologies, idiomatic expressions, figurative language and analogy.

From a terminological point of view, meeting the requirements of effective communication, extending automatic term extraction and building the necessary LSP tools means incorporating a phraseological dimension which will describe the actual functioning of terms in LSP discourse. Beyond terminology as such, phraseology, verbs, prepositional phrases, lexical syntagms, extended terms and so on have to be considered as requirements for specialized communication clarity. All these linguistic units are needed to consider terms in their environment i.e. how they actually operate in discourse. This is essential for technical writing and translation, two major language activities used in the dissemination of knowledge.

One of the problems is the difficulty to delimit the scope of what is called phraseology, since phrasemes, and terminological phrasemes are used as well by some authors, showing a basic uncertainty in the vocabulary itself. Roberts (1993a) defines phraseology in the following: “All usual word combinations which do not belong to a specific grammatical category. It therefore covers the vast majority of what I term fixed expressions (which include idioms in the narrow sense of relatively frozen expressions whose meaning does not reflect the meanings of their components parts as well as more transparent expressions such as similes, proverbs and sayings) and collocations (which are phrases which are only more or less fixed in a given language, both grammatical and lexical”. She eliminates compounds such as aptitude test and jet engine from the field of phraseology.

Meyer et al. (1996) define phrasemes as a combination of collocations and compounds (in the sense of Sager, 1990). Both of them are considered as terminological word combinations, however a collocation is different from a compound in the sense that a compound stands for a single concept while a collocation does not.

As there is a difficulty in finding a consensus on the definition of phraseology itself, we will use this term here in the sense of lexical combinations as used in LSP discourse. In other words we will consider phraseology as the various linguistic units which can roughly stake out a term’s environment, the term itself standing at the center of the field of forces thus delineated. L’Homme (1996) adopted the same broad position on this topic and we consider her approach to be operationally efficient.

3. Corpus-based term recognition and extraction

Extracting and building terminologies from machine-readable corpora increasingly uses natural language processing (NLP) techniques. As a result of the growing interest in this type
of approach to computational linguistics, a number of products aiming at term recognition and extraction have been developed. Early applications of automatic term extraction and recognition were developed in the context of Information Retrieval (IR). While this research can be traced back to the late 1950s, automatic term extraction aiming at building reference tools for translation, technical writing or knowledge-based systems is a recent trend. IR work shares with terminology the assumption that a certain type of lexical unit can express concepts or topics without ambiguity. If we accept that, the recognition of proper lexical units can contribute to indexing. Term recognition and extraction (TRE) for IR purposes is being well documented in the literature, we will not dwell on it. We will instead focus on TRE as a computational linguistic activity, explaining the growth of MT research, natural-language interface design, lexical units management for specific applications to some sublanguages and to technical writing needs.

3.1. Approaches to Automatic Term Recognition

We will briefly describe the basic idea underlying TRE systems. Most of the extracting tools consider terms as noun phrases. Systems identify terms by using frequency, distribution and categorial pattern matching (Dagan, 1996; Lauriston, 1994). All lexical units contained in a given text are analyzed and matched to patterns (typical forms of terminological units) described in rules. The linguistic approach are heavily language-dependant, hence extension to another language involves a thorough redesigning of programs. More term extractors are accounted for elsewhere (L’Homme, 1996; Kageura et al., 1996; Dagnan et al., 1994). Some of the systems described by these authors are tested in the framework of an ongoing evaluation project (ACABIT, LEXTER, and ANA). Other term extraction tools, classifiers and semantic relation extractors are tested in the same framework.

The main characteristics and purpose of these systems i.e. their description as far as linguistic approaches are concerned, are documented in previous works: Mustafa el Hadi et al., 1996a; 1996b, Jouis et al., 1997; Béguin et al., 1997. We will now give a brief description of the three major categories we identified. The first category is defined as term extractors and consists of syntactically-based systems to which statistical modules are sometimes added. This category comprises a) ACABIT, a prototype geared to automatic term-bank construction. It is designed to help experts in a given field by offering them a set of potential candidate terms in a pre-defined morpho-syntactic format. The statistical module role is used to rank the selected terms according to their relevance to the field. b) ANA, which uses both statistical and linguistic methods to extract terms from corpora in French, according to their frequency. Simple terms so extracted are used in combination with linguistic heuristics and frequency considerations in order to extract further candidate terms. c) LEXTER extracts complex terms, first by selecting all the maximum-length noun phrases on the basis of surface grammatical analysis and parses them to extract suitable units as term.

Term extraction system designers developed strategies designed to achieve the retrieval of lexical combinations with forms comparable to those of complex terms. These systems will focus on the same formal characteristics i.e. frequency of complex forms and the nominal nature of terms. The problem is that terms can be identified by other than formal criteria. Candidate terms provided by extraction systems such as ANA, ACABIT, LEXTER, have then to be validated by subject specialists, since some of them would not have been picked by humans. In their current state these systems can only be used as tools to assist human identification of terms and semantic relations. They are used mostly as a means to speed up tasks which still require human intervention during their last stages (see also L’Homme, 1996b).

The second category of systems we identified is that of classifying tools. They aim to
build classes of terms. Such tools work purely on statistical and/or connexionist models. CONTERM is representative of this category.

The third category is that of semantic relations extractors. It includes systems such as IOTA, SEEK and SPIGRAPHE. These systems focus particularly on semantic relations. IOTA is a system based on a Sowa conceptual graph model. Its purpose is to identify noun phrases in order to index documents from textual. SPIGRAPHE is an extension of SPIRIT (Syntactic and Probabilistic System for Indexing and Retrieving Textual Information). The aim of this supplementary module is to improve document retrieval capabilities through a conceptual network which describes the contents of the base which is being queried. The user may select from this representation the information he/she needs. 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 concept relations (Jouis, 1995). SEEK could be considered as an atypical system compared to the systems we are testing because it does not need a parser nor a general language dictionary.

4. The Advantages of Integrating Phraseology Recognition and Extraction Strategies with Existing Automatic Term Recognition Tools

There are two advantages in integrating phraseology recognition and extraction strategies with existing automatic term recognition tools.

4.1. Advantages from the Point of View of Translation and Technical Writing

Integrating context provision facilities (concordances, Key-words in context (KWIC), etc.) with existing tools could be extremely useful for translators. Translators need not only foreign language equivalents, but also information on how to use a term properly. Even if definitions can help in understanding the meaning of a term, they give no information on how to use it in the target language (Bowker, 1996, 33).

The role of translation is becoming increasingly important in information retrieval, especially with the advent of the INTERNET. NLP systems can be developed as multilingual access tools to textual databases, because the problem is in fact an extension of general information retrieval. To implement multilingual querying using this approach, it is necessary to provide the corresponding translations for each of the languages identified. Multilingual interrogation tools can use existing machine translation systems to automatically translate the queries or even the entire textual data-base. Thus terminology is seen again as an essential component of information retrieval through translation. An efficient translation needs not only well-organized terms but also all the relevant information about them, such as definitions phraseology, verbs (L'Homme, 1996b) and so forth.

4.2. Advantages from the Point of View of Extracting Relevant Information on Concepts

Meyer et al. (1996) consider terminological phrases as potential indicators of conceptual meaning as well as indicators of particular meaning problems. These linguistic elements provide important conceptual footholds for terminologists working in phraseologically rich domains and ideally with machine-readable texts.

The role of terminological phrasemes as indicators of concept information for terminographers is outlined by Meyer et al., 1996, 4-5). They show how a domain can be influenced by concepts in other related domains. They then help identify the specialized concepts and hence the terms. Finally they can assist in organizing domain concepts into systems.

Integrating lexical combination with recognition and extraction strategies to corpus-
based NLP terminology processing techniques can help define conceptual attributes or relations. This idea is discussed in Meyer et al. (1996). In a different perspective, Bowker (1996, 34-35) observes that knowledge-rich contexts can be described as free-language combinations that frequently identify a particular conceptual attribute or relation. It is generally recognized that certain types of lexical phrase can be used to denote particular relations between concepts. These particular lexical phrases are assumed to express relations such as synonymy, hyponymy, myronymy, causality, material (see Bowker, 1996, 34-35 for a detailed description). Ahmad et al. (1992) define these lexical phrases as knowledge probes whose function is to help the terminographer in bringing together the various aspects of a concept, and in trying to find the meaning of a given term. Knowledge probes can also be compared to the linguistic markers used in Contextual Exploration Strategy (see Louis, 1995) which consists of rules of contextual exploration, detecting the presence of contextually relevant linguistic markers in texts so as to progressively build semantic representations. These rules are the expression of a linguistic knowledge which does not require expertise in any specific domain.

Furthermore terminological phrasemes serve as potential indicators of particular meaning problems such as polysemy or synonymy (see Meyer et al., 1996, 4-5).

Semantic relations normally existing between terms, and towards which most of the systems strive, are essential aspects of knowledge representation and retrieval. These relations they can be identified by the linguistic strategies we described.

5. Proposals for Future work

Readapting existing term extraction and recognition tools follows in the path of comparable work in the field of computational linguistics such as TERMIGHT Dagan et al. (1996) a tool which uses a part-of-speech tagging and word-alignment technology in order to extract candidate terms (noun phrases) and translations. It then sorts the extracted candidate terms and presents them to the user together with reference concordance lines to ensure the efficient construction of a glossary. The advantage of this tool lies in its capacity to be used in contexts other than human-based translation. The monolingual component of TERMIGHT can be used to construct term lists for a variety of other applications: technical writing, hypertext linking, natural language interfaces, text categorisation and indexing for digital libraries and information retrieval. The bilingual component can also be useful for information retrieval in multilingual text collections.

Concordancing techniques incorporated in these systems can be useful for several tasks. As far as term recognition and extraction is concerned it can help identify multi-words missing from the candidate lists by reviewing relevant lines of the document (see Dagan et al., 1996, 36).

What we propose is to adapt term extraction systems, by adding, in some cases, part-of-speech taggers, sorting the extracted candidate terms and presenting them to the user along with reference concordance lines. The purpose is to support the efficient construction of glossaries, linking terms and providing various contexts for a given term. A technical verb data-base could then be linked to an existing term extraction tool if the field has many technical verbs. Similar work was carried out by L'Homme (1997).

6. Concluding Remarks

LSP communication is used by people who are not subject specialists but rather language professionals. From a terminological point of view, effective communication and
dissemination of knowledge can be achieved by the use of terms in LSP discourse. This means incorporating phraseology in term processing. It is true that these strategies can help end the out-of-context separateness in which terms have been considered. But these lexicographic and lexicological approaches should never replace the conceptual-oriented approaches to term extraction and recognition tools. They can be useful, especially for writing and translating technical texts because information concerning term use in context is especially relevant for such texts. These interfaces should be focused on the conceptual organization of a given subject. Without going too far and straying from our fundamental conceptual approach (i.e. the term being a concept name) we could reach a compromise and balance the two views. Systems can be designed to enable access to specialized lexical combinations but the focal point of this representation should be the hierarchical organization of conceptual classes (cf. L'Homme, 1997).

Acknowledgements
The author gratefully acknowledges the financial assistance provided by the Association des Universités Entièrement ou Partiellement de Langue Française (AUPELF) - in funding the research project from which this paper derives.

References


Notes

20 Terms used to describe a subject domain form part of Language for Special Purposes (LSP), in contrast to Language for General Purposes (LGP) which does not relate to any particular domain.
Roberts (1993) gave a detailed account of the emergence of the term phraseology and it current use in terminology circles. She has pointed out the difficulties encountered in finding a satisfactory definition of the term. Some linguists see phraseology as limited to fixed expressions. Others consider that phraseology includes proverbs and sayings. Collocations in the sense of word combinations that are usual in language but not fixed, are also considered as phraseology by several linguists (Roberts, 1993, 4). Others include in the category of collocations noun + noun combinations with other term compounds as well as phrases consisting of a dominant word and preposition or grammatical structure. (Meyer et al, 1996). Roberts (1993, 4) concludes that phraseology covers in fact all or part of what has been termed by compounds, collocations, idioms, set/fixed /frozen/phrases/expressions and multi-word units.

Kageura et al. (1996) has given a State of the Art overview of automatic term recognition and retrieval. See also Lewis and Sparck Jones (1997).

It is a research project conducted 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 corpora in French. Systems 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. These systems have been described in our previous works (see Mustafa el Hadi and Jouis, 1996; Béguin, Jouis and Mustafa el Hadi, 1997).

The role of definitions in term banks and other retrieval tools such as thesaurii is examined respectively by Sager et al. (1994); Michèle Hudon (1996).

Contextual exploration aims at simulating the behaviour of a fast reading reader who is instinctively looking for linguistic markers allowing him to frame a network of concepts/relations. Such a reader who has to understand a text in a superficial way picks relevant markers. A system of contextual exploration is made up of declarative rules conveying a textually-based knowledge. Rules (of contextual exploration) are represented in the form of IF <conditions> THEN <actions> or <conclusions>. The existence of these rules enables the system to spot the presence of the relevant linguistic markers in the context. The markers are simultaneously related to several components: morphological, syntactical, lexical. The globalization of these rules enables the progressive building of semantic representations. This process requires the compilation of descriptive analyses in the form of finite and structured lists of linguistic markers aiming for exhaustivity. We are giving examples of linguistic markers in the following: verbs (distinguer / reconnaître / différencier / discriminer / isoler / séparer / distribuer / discerner / singulariser / particulariser / remarquer,...); markers related to the description of superordinate class / subordinate class or instance in a domain: (sous-classe / classe / sous-caste / caste / sous-catégorie / catégorie / sous-groupe / groupe / sous-division /...)

The tool is currently being used by the translators at AT & T Business Translation Services (formerly AT & T Language Line Services).