Multilingual Terminology Extraction and Validation

Antonio S. Valderrábanos, Alexander Belskis, Luis Iraola Moreno

SchlumbergerSema Spain
Albarracín 25, 28037 Madrid, Spain
{antonio.valderrabanos, alexander.belskis, luis.iraola}@sema.es

Abstract

This paper presents the automatic terminology extraction approach developed within project LIQUID[1]. This project aims at developing a cost-effective solution for the problem of cross-language access to multilingual text databases in technical and scientific domains. Cross-Language Information Retrieval faces a major challenge: organizing unstructured textual information according to its contents and regardless of its language. Our solution is based on two main components, a terminology extraction tool and a domain-specific ontology. The terminology extraction tool identifies the terminology that describes the contents of a particular document. Then, these terms are linked to a domain-specific ontology. This paper presents the terminology extraction tool and the experimental results obtained in the domain of Gastroenterology.

1. Introduction

LIQUID aims at providing solutions to the task of Cross-Language Information Retrieval (CLIR) from unstructured, multilingual document bases that belong to highly specialized domains such as Gastroenterology (the one chosen in the project) or to one of its sub-domains like, acute abdominal pain.

The approach presented here is based in the following assumptions:

a) The proliferation of multilingual document bases without parallel structure (i.e., document X in language A is not present in language B, and document Y in language B does not have a counterpart in language A). Examples of these databases are the proceedings of an international congress, the resolutions of the European Commission, etc.
b) Most of these databases contain documents belonging to specialised domains (technical, scientific, legal) and contain highly valuable knowledge.
c) The need to retrieve texts in a multilingual context is becoming a common task for individuals across international organisations and companies (such as international civil servants, employees of multinational companies, medical staff, etc.).

There are a number of major obstacles to guarantee universal access to knowledge in the previously defined context:

a) *Lack of structure*. The preferred format to express knowledge is free text in natural language. Having many other advantages, free text lacks structure and this makes difficult the task of finding a particular piece of knowledge in it.
b) *Language barrier*. English is the language of choice for putting knowledge into paper. This fact poses a major barrier for non-native English speakers and machine translation is not satisfying the expectations generated in the last decades. The problem remains the same for any other language: most of the times, knowledge is expressed in a single language and translation is still an expensive process.
c) *Content barrier*. Most text handling programs store and manage text the same way as numbers, being quite different pieces of information. As a result, the vast majority of computer programs designed to handle text are unable to identify “cars” as the plural of “car”, not only in English but in any other language.

In this context, there is a strong need for software systems that are capable of structuring the knowledge contained in free text according to its content and thus overcoming the language barrier. LIQUID aims at providing solutions to these problems by handling text according to its content and linguistic properties, focusing particularly on technical terms as indexing items.

The idea of making terms the main candidates for indexing documents relies on two main facts:

- In technical or scientific texts, terms bear most of the semantic content
- The monosemic nature of terms makes them ideal candidates for indexing, since they will let us avoid ambiguity

According to (Lewis & Croft, 1990) terms represent best quality descriptors for document indexing due to their high informational content.

1.1. LIQUID requirements

The following requirements where defined for the project:

1. *Affordable and feasible*, the development process should be as streamlined as possible.
2. *Domain independent*, i.e. portable to other scientific or technical domains.
3. *Language independent*, i.e. portable to other (initially EU) languages with a reasonable effort.
4. *Complementary with existing monolingual IR systems* in their current state, so it can be seamlessly integrated with them.

In short, the system can be defined as cheap to develop and effective to use.

With these requirements as starting point, the resulting system will behave as follows: given a query in the native language of the user, it will return documents in different languages available in the multilingual document base. This system will help the user in the formulation and translation of his/her query as well.

[1] LIQUID is an RTD project funded by the European Commission under the 5th Framework Programme (IST-2000-25324). LIQUID started on January 1st, 2001. Four languages are considered in the project: French, German, Spanish, and English.

2. The state of the art in CLIR

CLIR systems can be classified in two main groups, depending on the component that gets translated: those that translate the query and those that translate the target document (Yang et al. 1997). In both cases, the chosen component is translated into the other language(s) covered by the system. Besides, there is a third group that aims at translating both components into an interlingual representation; the availability of new large scale resources like EuroWordnet and its interlingual index are essential to this third approach (Gonzalo et al. 1999).

LIQUID uses a query-translation strategy because it's compatible with the first requirement: the final system must be affordable and feasible. The other approaches are incompatible with our requirements for the following reasons. Using a document-translation approach implies either human translation (which is expensive and slow for large document collections) or machine translation (which is not a practical solution yet due to quality limitations (Hovy et al. 2000). As for the interlingual approach, producing the necessary resources, like EuroWordnet, for specialised areas, like Gastroenterology in our case, is expensive and time consuming (Gonzalo et al. 1998).

A wide array of resources is used in CLIR (Radwan & Fluhr, 1995; Oard, 1997), ranging from multilingual glossaries or dictionaries to multilingual collections of texts and sophisticated taggers and parsers. Machine translation systems would represent the most sophisticated solution from this point of view.

2.1. Types of CLIR systems

According to the resources used, CLIR systems can be classified in two groups (Gonzalo et al. 1999; Ballesteros & Croft, 1997; Jacquemin & Bourigault, 2001):

- *Knowledge based approaches* that use multilingual glossaries and dictionaries.
- *Corpus based approaches* that use parallel or comparable multilingual corpora.

Examples of the first are (Hull & Grefenstette, 1996) or (Ballesteros & Croft, 1996); and of the second (Sheridan & Ballerini, 1996). Currently there is a trend to combine the two approaches. The major problem for knowledge based approaches is that technical terminology is not normally present in reference works and it grows at a fast pace. Reference works hardly keep up with this new terms and then lack the necessary specificity. For corpus based approaches the problem is exactly the opposite: lack of generality. Since they are based in a particular set of texts, they are very sensitive to domain changes. As we can see, from a terminological point of view, there are two contradictory demands: on the one hand, the need to have a broad coverage (so the system is portable across domains); and, on the other hand, the need to have exhaustive coverage (so no term in the domain is unknown to the system).

Keeping the initial requirements in mind, particularly portability, LIQUID focuses on resources that can be developed or acquired within tight time and money constraints, and avoids the use of resources that are expensive (either in terms of time or money). Using these kind of resources (e.g. full parsers and broad coverage dictionaries) would be a major obstacle for our final goal: building a cost-effective system. Whole projects have been devoted only to the production of these resources (e.g. ACQUILEX I and II, LE-PAROLE and LS-GRAM). Besides, other projects like ESPRIT-EMIR have already exploited the potential of using this kind of resources for CLIR.

LIQUID aims at solving these demands of exhaustiveness and broadness using the following strategy. First, existing glossaries will be used as a starting resource to ensure a reasonable broad coverage of the domain. Then, the corpus that is the target for the CLIR system will be used as a source to extract new terms (strictly speaking, new terms and variant terms too as described below) and to enrich the initial glossaries. In this way we can ensure that the final glossary will fully cover the application domain.

3. Components of the LIQUID CLIR system

The document base. It contains the documents that will be the target of the CLIR system. The text corpus that will be used as target must be multilingual, representative of a specific scientific domain and non-parallel. This corpus is both the problem to be solved (knowledge in different languages) and the starting point to develop other resources, like term sets.

The term sets. They provide the link between the document base and the semantic network, since they are present in documents and linked to concepts of the semantic network. Every document in the base will be linked to the semantic network, thus obtaining a conceptual organization of them based on their terminology. This is possible because:

a) Specialised terminology is monosemous since its goal is to transmit technical and scientific knowledge. Therefore, linking specialised terms (e.g. "squamous carcinoma") to a semantic network poses a much simpler problem than linking general language words (e.g. "house") where polysemy is the rule and not the exception. Several studies reveal that polysemy as the most important problem for effective CLIR (Hiemstra, 1997)
b) In technical or scientific texts, specialised terminology carries most of the relevant information; as a result, classifying terminology present in a document amounts to identifying the conceptual area where the document belongs

The semantic network. It structures terminology according to meaning and reflects the way knowledge is organized in the application domain. This conceptual organisation plays a pivotal role among the terms linked from different languages.

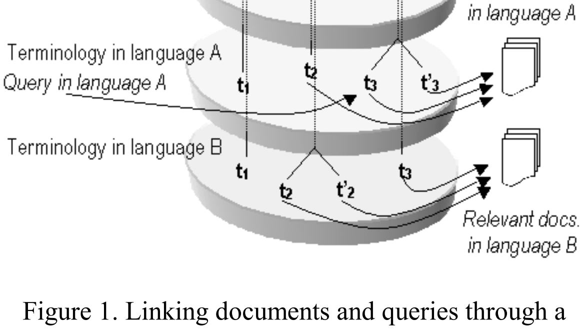
Domain ontology
part_of
C2
is_a
C1
C3
Relevant docs. in language A
Terminology in language A
Query in language A
t1
t2
t3
t'3
Terminology in language B
t1
t2
t'2
t3
Relevant docs. in language B

Figure 1. Linking documents and queries through a multilingually-mapped semantic network.

As a result of the combination of the three components we can link every document in the document base to the semantic network through the set of terms, thus obtaining a semantic organisation of the documents based on the terminology they contain. The linking will be based on the presence of a particular term in both the semantic network and the document. Since the semantic network is multilingual, it is possible to make the text database available across languages.

The use of thesauri is closely linked to controlled vocabulary systems, well known for their effectiveness now for over 30 years (Salton, 1970). However, their limitations are also known. The major drawback that thesauri and controlled vocabulary systems pose affects terminology: terms used in the query must be restricted to the ones present in the thesaurus. In LIQUID, we intend to exploit all the benefits of thesauri and overcome their limitations via the term extractor (TExtractor), which will keep the thesaurus updated with new terms found in the target corpus.

Validating the resulting broad and exhaustive set of terms and linking it to the thesaurus are the most costly development tasks, since they need manual verification. For this reason, automatic thesaurus building will be one of the main focus of the project to achieve the overall goal of cost-effectiveness. The major contribution we expect from this task are new insights on how to bridge the gap between controlled vocabulary systems and free text retrieval.

The methodology can be applied to any domain-specific document base. One of the strongest aspects of the LIQUID system is its portability to other scientific and technical domains.

4. Term detection and extraction

In this section we briefly introduce the state of the art in terminology detection, present the LIQUID approach to terminology extraction, paying special attention to the derivation rules and the validation process, and conclude with some experimental results.

4.1. Short introduction to current approaches

There are two major research trends in the field of terminology detection or extraction: statistical and linguistic. Statistical approaches can cope with high frequency terms but tend to miss low frequency terms (Evans, 1996), generating what's called 'silence'. Conversely, linguistic approaches are more efficient at identifying infrequent terms (what we call 'new terms'), as proven in (Bourigault, 1993; 1996). However, strategies based on linguistic knowledge tend to produce 'noise', i.e., they identify as terms word combinations that are not.

By 'detection' we refer here to two major activities in the field of terminology and Natural Language Processing (Jacquemin & Bourigault, 2000):

- term recognition: identification of known terms.
- term acquisition: automatic discovery of new terms.

Both activities refer to the automatic processing of text corpora as a source of terminology. (Jacquemin & Bourigault, 2000) provides a clear overview of the different existing terminology recognition or acquisition systems. In LIQUID, term recognition refers to both concepts; term extraction is used as a synonym of recognition.

4.2. The LIQUID approach

Our term extraction strategy is based on corpus evidence and driven by linguistic data. Linguistic analysis is based on identifying phrase delimiters and on very shallow parsing. As already stated, expensive resources like general dictionaries or full-fledged parsers, as used by (Arppe, 1995) or (Justeson & Katz, 1995), will not be part of the strategy in order to ensure feasibility and portability.

Whenever possible, we will not start building term sets from scratch but from previously existing glossaries. Reusing previous efforts and ensuring coverage of most common terms are two reasons for doing this. Besides, other researchers, like (Jacquemin et al. 1997, and Jacquemin & Tzoukermann, 1999), have stressed the fact that starting with an initial term set improves the results of automatic term extraction strategies. Since we start with a set of terms, the study of term variation becomes a key component (Daille et al. 00).

4.2.1. Variant and new terms

Term variation negatively affects the performance of information management systems that are unable to identify as synonyms terms that differ in their morpho-syntactic realisation (e.g. "polio vaccine" and "vaccine against polio"). The term extraction tool developed within the LIQUID project helps to solve this problem in an automatic way, providing two main benefits:

1. To increase the quality of the initial term sets (which is particularly necessary when these sets do not have a wide coverage of the domain), and
2. To facilitate the task of keeping the whole system (text databases and semantic networks) synchronised and updated as new documents are added.

Variant terms are terms that express the same concept as the term they derive from. They include the following types of changes or variations:

a) *Morphological variations*, identifying the root and its forms, like in: "X-ray therapy" and "X-ray therapies".
b) *Syntactic variations* in the construction of terms, like in: "HIV vaccine" and "vaccine against HIV".
c) *Formal variations*, like abbreviations or acronyms, like in: "PAHO" and "Pan-American Health Organization".

New terms express a different concept than the one expressed by the term they derive from. Different strategies and linguistic knowledge are employed:

a) Using known terms as source, like when extracting "common bile duct obstruction" based on "common bile duct".
b) Using suffixes, like "-itis": "diverticulitis".
c) Analysing other linguistic phenomena like co-ordination, as in the derivation of: "stomach ulcer" and "duodenal ulcer" from "stomach and duodenal ulcer".

Variant and new term generation patterns have been expressed in derivation rules. These, together with bits of linguistic knowledge, are applied by the extraction tool over an initial set of seed terms in order to produce the variants and the new terms.

4.2.2. Input resources

The process of term generation and subsequent validation employs the following resources:

a) *Linguistic knowledge* such as elementary morphological rules (stemming) plus lists of function words for each of the languages covered.
b) *Initial term sets* containing known terms for the domain and languages of choice[2].
c) *Derivation rules*. Applied over known terms, they produce candidate new terms. Because of the approach followed, derivation rules are highly re-usable among languages. French and Spanish rules are almost identical, the same happens with the English and German rule sets.
d) *Validating document base* containing documents for the domain and languages of choice.

All these resources are provided as plain text files to TExtractor, a Java-based application that automatically produces a set of new terms that are valid indexing items for a given domain. The results are generated also as plain text files, though in order to ease their inspection they are loaded into a database in which new terms are inspected and traced back to the input data that generated them.

4.2.3. Derivation rules

Rules for deriving new indexing terms conform to the classical conditional structure:

IF Antecedent Conditions THEN Consequent Actions

Antecedent conditions are checked on a singular term (a member of the initial set of terms) and, if fulfilled, the final result of applying the sequence of consequent actions over it produce a newly generated term. Both conditions and actions apply over the individual tokens that compose a typical multi-word term.

The kind of conditions that can be checked in a derivation rule over any individual token fall under one of the following categories:

a) *Typographical*, such as the presence of a hyphen or an initial capital in the token.
b) *Morpho-syntactic*, such as the property of number for nouns and the part-of-speech of the token.

Morphological properties are determined by means of simple suffix checking and applying highly productive heuristics. Because of their simplicity and the public availability of these kind of morphological resources, these mechanisms are cost-effective and scalable to most European languages. Of course, mistakes are sometimes made, but they are pruned in the subsequent validation phase.

Regarding part-of-speech determination, it has been introduced mainly to improve the comprehensibility of the rules, since no wide coverage mechanism for POS determination has been incorporated to TExtractor. Following the general approach towards cost-effectiveness and scalability, only function words have been fed into the term extractor. Conditions involving an open POS category for a certain token are automatically granted, as in the following rule that derives a new term if the initial one is a singular noun:

tr1[Noun, Singular] > MakePlural(tr1)
lobotomy > lobotomies

Even when the term extractor does not have in its current state any means for tagging "lobotomy" as a noun, the rule fires anyway and the plural form is generated.

POS tagging of open categories has been included in the rules mainly to improve their readability, since no wide coverage mechanism for POS determination has been incorporated. On the other hand, words belonging to closed categories (function words) have been compiled in lists and are available for checking tokens in rules like:

tr1 tr2[Class:ConjunctionCopulative] tr3 tr4 > tr3 tr4
Head and neck neoplasms > neck neoplasms

Consequent actions apply over individual tokens identified in the antecedent, as in the previous example where the action "MakePlural" is applied over token number one. Possible actions may affect to individual or to groups of several tokens:

- Re-ordering the token sequence.
- Joining two tokens in a single one.
- Remove/insert a certain token.
- Modify the typographical, morphological and/or syntactic properties of a token.

In addition to these elements, derivation rules are enriched in their antecedents with operators for bounded and unbounded repetition, thus greatly simplifying the task of writing rules.

As an example of the usage of the unbounded repetition operator (*, meaning zero or more occurrences of the base category), the following rule states that hyphenated tokens occurring in a term may produce un-hyphenated variants, regardless of the presence of previous and/or following tokens in the initial term:

tr1* tr2[Hyphenated] tr3* >
    tr1 MakeUnhyphenated(tr2) tr3
fine-needle aspiration biopsy >
    fine needle aspiration biopsy

4.2.4. Term validation

TExtractor works in two phases: generation and validation.

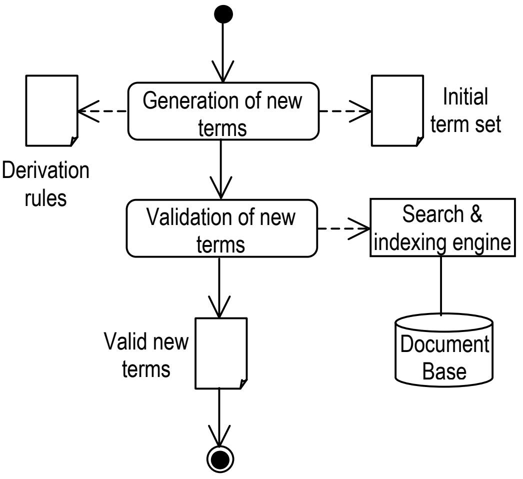

Figure 2. UML activity diagram of the term generation and validation process.

[2] In the medical domain, MeSH (Medical Subject Index), SNOMED (Systematized Nomenclature of Medicine) and ICD (International Classification of Diseases) are very valuable terminological resources.

As shown in Figure 2, the generation phase depends on two resources: an initial term set and a set of derivation rules that, applied on them, generate new terms. The initial term set is provided by reusing existing glossaries; they are used as seeds or examples in order to enrich them. This is one of the most significant differences with respect to existing approaches.

The heuristic nature of the morphological derivations, the limited scope of the syntactic information (reduced to the knowledge of function words) and the absence of any semantic or contextual information, makes the process to over-generate. This is by no means an unexpected consequence, and in fact the whole approach can be viewed as an instance of the generate-and-test paradigm. Although produced according to linguistically motivated rules, many of the newly generated terms are not valid (indexing) terms and should be discarded during the validation process.

Every generated term is validated against a document base containing a substantial amount of domain-related documents. As a first validation criterion, a term is considered valid if it occurs in at least one of the documents of the base. This initial criterion can be modulated afterwards considering the frequency of appearance of the new term in the collection and/or usability constraints. For our current purposes, this initial criterion has provided us with a surprisingly reliable indication of the potential usefulness of the new term.

In order to implement the validation process, we have employed Lucene, an open source tool that provides extensive search and indexing capabilities over text files. Lucene (Goetz, 2000) offers a well-documented interface for accessing its capabilities and we have coupled our term extraction tool (TExtractor) to Lucene for checking the presence of derived terms in our document base. Figure 3 shows the dependencies between components:

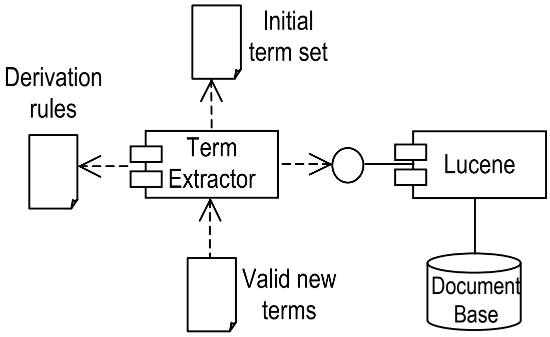

Figure 3. UML diagram showing the components involved in the term validation process.

5. Experimental results

In order to test empirically our approach to automatic term acquisition, we have:

1. Written generation rule sets for the four languages covered in the project (English, Spanish, French, and German).
2. Selected a document base of clinical reports in the domain of Gastroenterology.
3. Defined a quantitative measure of how successfully the tool acquires new valid terms.
4. Selected initial term sets in the chosen domain and for the four languages covered.

5.1. Generation rule sets

One of the key requirements of project LIQUID is to devise a term acquisition technique that is as language independent as possible. This requirement has been met employing as few (and simple) linguistic resources as possible and writing generation rules as general as possible. Most of the linguistic differences are located in the morpho-syntactic knowledge (encoded in separate data files), thus allowing us to write very general generation rules.

We have written 67 derivation rules for English and the same number of rules happened to be required for Spanish. Most of the rules are identical for both languages and the divergences are mostly due to syntactic differences between both languages in the structure of noun phrases. German generation rules have been written taking the English ones as a starting point, and most of the changes involve modifications due to the different morphologies. A total of 68 generation rules have been written for German.

Finally, the French rules are identical to the Spanish ones except for a couple of minor differences.

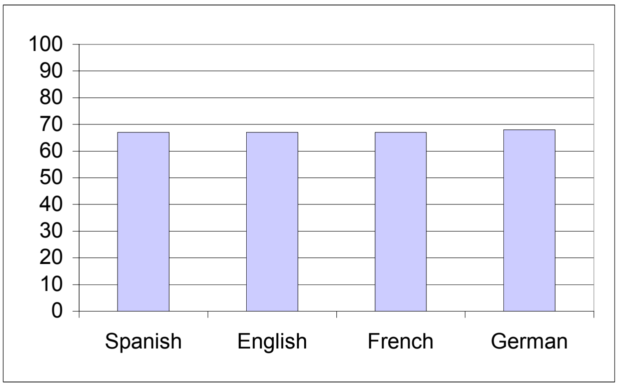

Figure 4. Number of generation rules written for each language covered in the project.

5.2. Validating document base

As previously mentioned, the generated terms are automatically validated against a collection of documents belonging to the application domain. In our case, we have selected the ELCANO document base[3], a publicly available collection of clinical cases in the domain of Gastroenterology (http://www.imim.es/elcano).

The initial base included 563 clinical reports written in English and the same amount written in Spanish. During the course of LIQUID, this base has been considerably enlarged with more English and Spanish clinical reports and with reports written in German and French, thus producing a substantial multilingual corpus in the application domain.

The following figure presents the absolute size of the validating document base for each language in terms of the number of different words.

[3] The ELCANO document base is one of the results of the ELCANO project (European and Latin-American Countries Associated in a Networked database of Outstanding Guidelines in unusual clinical cases), INCO/DC Programme, DG XIII.

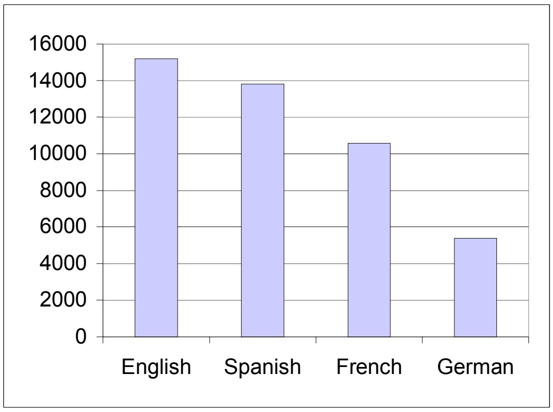

Figure 5. Size of the validating document base (in number of different words).

5.3. A quantitative measure of term enrichment

In order to give an objective performance measure, we have focussed on the number of valid newly generated terms that are acquired automatically by the system, compared to the number of valid initial terms. The ratio between both figures gives a quantitative measure of how successfully we have enriched the initial term set with newly generated terms.

It is important to note that this measure involves *valid* terms, i.e. terms that do occur at least once in the validating document base. This means that not only the newly generated terms but also those coming from the initial term set are validated in order to compute the measure. Validating the initial term set allows us to employ as initial term set any domain glossary, even one whose quality is not assessed or is known to be poor.

5.4. Results using the ELCANO initial term set

A first test has been performed using as initial terms the keywords employed as document descriptors in the original ELCANO document base. Given that the original document base only contained Spanish and English documents, this first test only involves the generation of new terms in these two languages.

Language	initial terms	initial valid terms	new valid terms	new / initial valid terms
English	1222	916	666	72,7%
Spanish	1226	916	888	96,94%

Table 1. Results using the ELCANO term set.

The English and the Spanish initial term sets are almost identical in size and exactly the same number of initial terms have found valid (916). However, the Spanish generation rules have produced more new valid terms than their English counterparts (888 against 666 new valid terms). Consequently, the enrichment ratio is higher for the Spanish term set than for the English one.

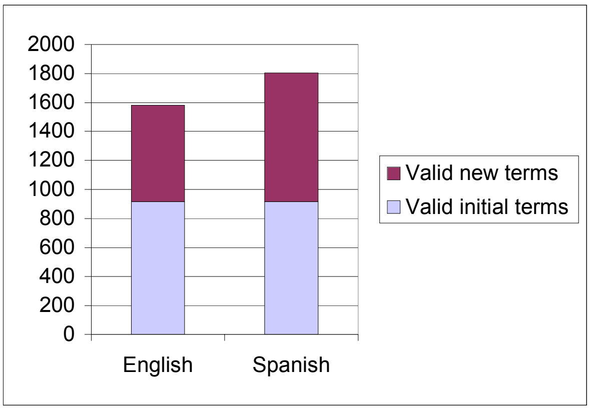

Figure 6. ELCANO valid terms.

5.4.1. Manual validation

In order to verify the quality of the automatic validation process, the set of valid new terms (whether coming from the initial term set or automatically generated) has been manually checked. Valid terms in both languages have been revised, looking for:

1. *Incorrect terms*: syntactically ill-formed terms, such as “infection in surgical”. These text fragments are considered new valid terms because they appear as part of larger phrases in the validation corpus. However they make no sense as isolated terms.
2. *Irrelevant terms*: generic, domain unspecific terms such as “History” or “expert” that are poor content descriptors for clinical reports.

Type of error	English	Spanish
Number of incorrect terms	13 (3%)	8 (0,94%)
Number of irrelevant terms	8 (1,14%)	27 (3,19%)
Incorrect plus irrelevant	21 (4,14%)	35 (4,13%)

Table 2. Errors found in the ELCANO term set.

5.5. Results using the Radcliffe initial term set

A second extraction experiment has been performed employing one of the most prestigious multilingual terminological resources in Gastroenterology: *The International Wordbook of Gastroenterology*, by R. Pounder & M. Hudson, published by Radcliffe Medical Press, 1994.

The Radcliffe term set allows us to check our generation rules in the four languages contemplated, given that all the medical terms are translated in all of them. A second advantage of this term set is its independence with respect to the validating document base. The ELCANO initial term set was directly related with the ELCANO document base: the terms were the keywords employed for describing the contents of documents included in the base. The rather good enrichment figures obtained may be partly due to this close relation between the seed terms and the validating corpus. This second experiment removes that influence; now the seed terms and the validating corpus come from different sources.

Language	initial terms	initial valid terms	new valid terms	new / initial valid terms
English	4253	2142	971	45,33%
Spanish	4306	1585	1053	66,43%
French	4310	1260	1034	82,06%
German	5243	525	201	38,28%

Table 3. Results using the Radcliffe term set.

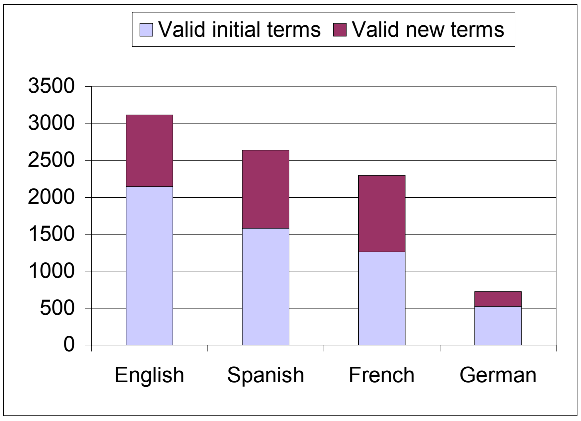

Figure 7. Radcliffe valid terms.

5.6. Discussion of the results

The comparison of the results obtained with the ELCANO and the Radcliffe term sets, taking into account the corpus sizes employed for validating each language, allows us to draw the following conclusions:

1. It is safe to assume that the performance differences observed between the first and second experiments (average enrichment ratio of 85% against 58%) are due to the close relation between the seed terms and the validating corpus.
2. There is a clear correlation between the validating corpus size and the number of terms (whether coming from the initial set or newly generated) that are accepted as valid: the larger the corpus, the larger the number of terms found valid. This explains the comparatively poorer figures obtained for the German language: its validating corpus is significantly smaller than the remaining ones.
3. The significantly larger absolute figures of valid terms obtained in English may be due to two facts: the larger size of the English initial term set, and the English origin of the Radcliffe terminology set, whose terms are firstly compiled in English and then translated to other languages. Differences between the proposed translations and the technical terminology found in actual clinical reports may have hindered the results in languages other than English.
4. Focussing now on the enrichment percentages rather than in the absolute number of terms found valid, and leaving aside the German case (bogged down by a comparatively smaller validating corpus), the ratios show a remarkably consistent performance. Given that the generation rules are essentially identical for all the languages (except minor differences), we interpret this result as a strong evidence in favour of the language independent nature of the approach taken.

6. Conclusions and future work

We have presented a solution to the problem of cross-lingual information access to multilingual document bases in specific domains. The solution involves as central components a language independent domain ontology and a terminology extraction tool that provides to the ontology linguistic realisations of domain concepts in four languages: English, French, German and Spanish.

In this paper we have focussed in the second element of the proposed solution, the terminology extraction tool. We have shown that it is possible to enrich substantially an initial set of indexing terms applying a generate-and-test framework. Our approach to term extraction can be characterised by:

a) Very low dependency on linguistic resources. Only basic morphology, lists of function words, prefixes and suffixes are employed.
b) Small set of linguistically motivated derivation rules. Less than one hundred rules per language have been written, most of them common to all languages.
c) Incorporation of publicly available software tools. We have integrated Lucene, a publicly available IR engine.
d) Exhaustive validation of the newly generated terms against a domain document base.
e) Low error rates of incorrect and irrelevant terms. On average, less than 5% error rates have been observed.

We have tested our approach in the domain of Gastroenterology with a collection of medical documents as validating corpus and two initial sets of indexing terms: ELCANO and Radcliffe. The first one allowed us to test the system in English and Spanish while the second added to these languages French and German. The results encourage us to pursue this technique, paying attention to issues such as:

- Re-use of derivation rules, attempting to capture language independent derivation phenomena.
- Incorporation of publicly available, wide coverage linguistic resources that will enhance the derivation capabilities while maintaining the overall cost-effectiveness and scalability.
- Incorporation of publicly available terminological resources in the medical domain and for the languages considered in the project.

Terminology is bound to be a major source of knowledge in different areas of text analysis. However, its identification in unstructured text involves the use of large text collections (statistical methods) and/or costly linguistic resources (lexicons and grammars). Knowledge-light strategies combining both approaches, as the one presented here, are a promising path that deserve to be tested in a wider variety of languages and domains.

7. References

Arppe, Antti, 1995. “Term Extraction from Unrestricted Text”. Paper presented at NODALIDA-95, Helsinki (Available at http://www.lingsoft.fi/doc/nptool/term-extraction.html - 20-12-00).

Ballesteros, L. and Croft, W.B., 1996. “Dictionary Methods for Cross-Lingual Information Retrieval”. In Proceedings of the 7th International DEXA Conference on Database and Expert System, 791-801.

Ballesteros, L. and Croft, W.B., 1997. “Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval”, in Proceedings of ACM SIGIR Conference, 20: 84-91.

Bourigault D., 1993. “An endogenous corpus-based method for structural noun phrase disambiguation”. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics* (EACL'93). Utrecht, The Neederlands.

Bourigault D., 1995. “LEXTER, a Terminology Extraction Software for Knowledge Acquisition from Texts”. In *Proceedings of the 9th Knowledge Acquisition for Knowledge Based System Workshop* (KAW'95). Banff, Canada.

Bourigault D., Gonzalez-Mullier I. and Gros C., 1996 “LEXTER, a Natural Language Tool for Terminology Extraction”. In *Proceedings of the seventh EURALEX International Congress*, Göteborg, Sweden. 771-779.

Daille, B., Habert, B., Jacquemin, C., and Royauté, J., 2000. “Empirical observation of term variations and principles for their description”. *Terminology*, 3(2), 197-258.

Evans, D. and Chengxiang Zhai, 1996. “Noun-Phrase Analysis in Unrestricted Text for Information Retrieval”. *Proceedings, 34th Annual Meeting of the Association for Computational Linguistics*, 17-24.

Goetz, Brian, 2000. “The Lucene search engine”, *Java World*. The official Lucene web site is located at: http://jakarta.apache.org/lucene.

Gonzalo, J., Verdejo, F. and Chugur, I., 1999. “Using EuroWordNet in a Concept-Based Approach to Cross-Language Text Retrieval”. *Applied Artificial Intelligence* Special Issue on Multilinguality in the Software Industry: the AI contribution.

Gonzalo, J., Verdejo, F., Peters, C. and Calzolari, N., 1998. “Applying EuroWordNet to Cross-Language Text Retrieval”. *Computers and the Humanities*, Special Issue on EuroWordNet.

Hiemstra, D., F.M.G. de Jong, and Kraaij, W., 1997. “A domain specific lexicon acquisition tool for cross-language information retrieval”. In *Proceedings of RIAO'97 Conference on Computer-Assisted Searching on the Internet*, 255-266, 1997.

Hovy, E.H., Ide, N. Frederking, R.E., Mariani, J. and Zampolli, A. (editors) . 2000. *Multilingual Information Management*. In press. Also available at http://www.cs.cmu.edu/~ref/mlim - 20-12-00.

Hull, D., and Grefenstette G., 1996. “Experiments in Multilingual Information Retrieval”. *Proceedings of ACM, SIGIR'96*. Zurich.

Jacquemin, C., Klavans, J., Tzoukermann, E., 1997 “Expansion of multi-word terms for indexing and retrieval using morphology and syntax”. Proceedings, *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (ACL-EACL 97), 24-31. Madrid.

Jacquemin, C., and Tzoukermann, E., 1999. “NLP for Term Variation Extraction: a Synergy of Morphology, Lexicon and Syntax”. In T. Strzalkowsky, editor, *Natural Language Information Retrieval*, 25-74. Kluwer. Boston, MA.

Jacquemin, C., Bourigault, D., 2000. “Term Extraction and Automatic Indexing”. In R. Mitkov, editor, *Handbook of Computational Linguistics*. Oxford University Press, Oxford.

Jacquemin, C., and Bourigault, D., 2001. “Term Extraction and Automatic Indexing”. In R. Mitkov, editor, *Handbook of Computational Linguistics*. Oxford University Press, Oxford.

Justeson, J. and Katz, S., 1995 “Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text”, in *Natural Language Engineering*, Vol. 1, No 1: 9-27

Lewis, D. and Croft, W., 1990 “Term clustering of syntactic phrases”. In *ACM SIGIR-90*, 385-404.

Oard, D., 1997. “Alternative Approaches for Cross-Language Text Retrieval”. *Proceedings of the AAAI Spring 1997 Symposium on Cross-Language Text and Speech Retrieval*.

Radwan, K. and Fluhr, C., 1995. “Textual database lexicon used as a filter to resolve semantic ambiguity”. *Application on Multilingual Information Retrieval*. SDAIR'95. Las Vegas.

Sager, J. C., 1990. *A Practical Course in Terminology Processing*. John Benjamins. Amsterdam.

Sheridan, P. and Ballerini, J. P., 1996. “Experiments in multilingual information retrival using the spider system”. In *Proceedings of ACM/SIGIR*.

Yang, Y., Carbonell, J., Brown, R., Frederking,R., 1998. “Translingual Information Retrieval: Learning from Bilingual Corpora”. *AI Journal* Special Issue: Best of IJCAI'97.