

Automatic terminology extraction and validation: the LIQUID approach

*Antonio S. Valderrábanos, **Alexander Belskis, **Luis Iraola Moreno
*Bitext, **Sema Group sae
Madrid, Spain
asv@bitext.com, alexander.belskis@sema.es, luis.iraola@sema.es

Keywords: natural language processing; content-based indexing; content-based retrieval; cross-language information retrieval; terminology extraction; semantic network; thesaurus; ontology; medical texts; gastroenterology

Abstract

This paper presents project LIQUID. This project aims at developing a cost-effective solution for the problem of providing cross-language access to multilingual text databases in technical and scientific domains. This problem is generally known as Cross-Language Information Retrieval (CLIR). In order to provide a solution to this problem it is necessary to develop a system that, given a query in a particular language, will return relevant documents in any of the languages available in the multilingual document base. Currently, available search engines do not go beyond the language barrier: the query language determines the language of the returned documents.

The solution proposed needs to solve one major challenge: organising unstructured textual information according to its contents; and regardless of its language. Our solution is based on two main components:

- A terminology extraction tool
- A domain-specific ontology

The terminology extraction tool will identify those keywords (smallest content units in free text) that describe the contents of a particular document. Then, these keywords will be linked to the domain-specific ontology. As a result we will obtain a multilingual system to classify documents based on their contents. This paper is mainly devoted to the terminology extraction approach implemented in the project.

Note: LIQUID is an RTD project funded by the European Commission under the 5th Framework Programme. LIQUID started on January 1st, 2001. Four languages are considered in the project: French, German, Spanish and English.

1. Introduction: The CLIR problem

The approach proposed is based in the following assumptions:

- 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.
- Most of these databases contain texts that belong to a technical or scientific domain and contain highly valuable knowledge; however these documents are not multilingual
- The need to retrieve texts in this kind of 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.)
- The typical profile for these individuals is as follows: a person with some (basic) knowledge of one or more languages (apart from his/her mother tongue) and involved in a technical or scientific job or research

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

- 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
- 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
- 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
- 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 terminology as indexing items. The idea to make 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 it will let us avoid ambiguity

According to (Lewis and Croft 90) terms represent best quality descriptors for document indexing due to their high informational content.

In this context, the following requirements were defined for the project:

- affordable and feasible, i.e. the development process should be as streamlined as possible (it should be cheaper and faster to deploy than human translation while quality should be better than machine translation)
- domain independent, i.e. portable to other scientific or technical domains
- language neutral, i.e. portable to other (initially EU) languages with a reasonable effort
- complementary with existing IR systems as they are now 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 a multilingual set of texts. This system will help the user in the formulation and translation of his/her query as well.

2. The state of the art

We will briefly describe the state of the art and outline LIQUID position and contributions to this scenario.

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. 97). In both cases, the chosen component is translated to the other language(s) considered in the system. Besides, there is a third group that aims at translating both components to 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. 99).

LIQUID uses a query-translation strategy in order to ensure affordability and feasibility. 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. 00, chapter 4). 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. 98).

A wide typology of resources is used in CLIR (Radwan and Fluhr 95; Oard 97), ranging from multilingual glossaries or dictionaries to multilingual collections of texts and sophisticated taggers and parsers. MT systems would represent the most sophisticated solution from this point of view.

According to the resources used, CLIR systems can be classified in two main groups (Gonzalo et al. 99; Ballesteros and Croft 97; Jacquemin and Bourigault 01):

- 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 and Grefenstette 96) or (Ballesteros and Croft 96); and of the second (Sheridan and Ballerini 96). Currently there is a tendency 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 exhaustivity. For corpus based approaches the problem is exactly the opposite: lack of broadness or 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).

LIQUID aims at solving these demands of exhaustivity and broadness with a two stage approach. 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) and to enrich the initial glossaries. In this way we can ensure that the final glossary will fully cover the domain of application.

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 resources would be a major obstacle for our final goal: building a cost-effective system. Resources of this kind are parsers and broad coverage dictionaries. Other projects have been devoted only to the production of these resources (like ACQUILEX, LE-PAROLE and LS-GRAM). Besides, other projects like ESPRIT-EMIR have already successfully exploited the potential of using this kind of resources for CLIR.

3. Architecture of the system

In this section we will describe the main components of the LIQUID system and their interaction to solve the problem of CLIR:

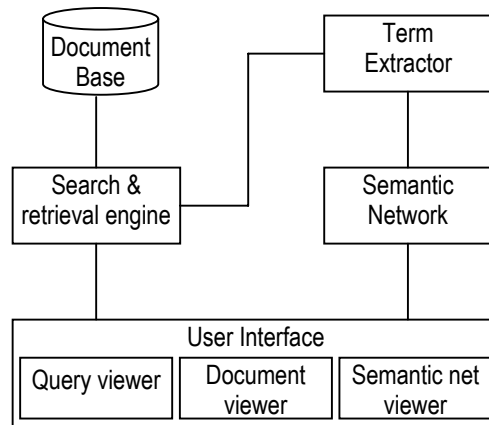


Figure 1. Overall architecture of the LIQUID system

a) The document base

The document base contains the set of documents that will be the target of the CLIR system. This text corpus 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. As a starting point, we are using the document base produced in the ELCANO project¹. It consists of a virtual library of unusual clinical cases in the field of gastroenterology. This preliminary database is being enlarged with new texts in order to obtain a multilingual and non-parallel database.

b) The term sets or terminology

Domain specific terminology (also known as keywords) plays a major role in the system. Terminology provides the link between the text database and the semantic network because it is both present in the texts that will be retrieved and are linked to nodes of the network, according to its content. As a result, it will be possible to link every document in the text database to the semantic network, thus obtaining a conceptual organisation of documents, based on the terminology contained in them. The underlying facts that support this claim are:

- Specialised terminology is monosemic (since its goal is to transmit technical and scientific knowledge); because of this, linking specialised terms (like "squamous carcinoma") to a semantic network poses a much simpler problem than linking general language words (like "house") where polisemy is the rule and not the exception. Several studies reveal polisemy as the most important problem for effective CLIR (Hiemstra 97)
- 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

Different sources have been checked in order to collect the term sets for the languages of the project (French, German, Spanish and English):

- existing term sets or glossaries, such as MeSH or SNOMED, or those provided by associations like ELRA (European Language Resources Association), LDC (Linguistic Data Consortium), etc.
- automatic term identification from the text database

¹ 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. The ELCANO project web site is located at: <http://www.imim.es/elcano>.

The strategy used in LIQUID involves a combination of the two possibilities. Starting from an existing resource has many advantages since it speeds up the development phase and reuses existing resources. Besides, identifying new terms in the text database, whether by hand or automatically, enriches the initial term sets.

c) The semantic network

The semantic network is the component where relevant terminology is structured according to its content. This component reflects the way knowledge is structured in the domain of application. It is characterised by the following features:

- Its structure is language neutral, since it does not follow the linguistic or cultural organisation of any particular language; in other words, it is shared by the scientific community as common background
- Its contents are multilingual, since it contains terms from different languages

Since this conceptual organisation is performed for all languages, terms will be linked across languages.

This network is designed as an abstract structure that reflects the conceptual organisation of the domain of application and the semantic relationships amongst terms. Once this network is designed, the initial term sets will be linked to it. As a result we will have the term sets organised according to their content. Typical relationships amongst terms in this network are hyponymy, hypernymy, synonymy, antonymy, meronymy and holonymy (besides variants, acronyms, abbreviations, etc.) Since terminology from every language is linked to the network, it is possible to translate terms across different languages. The resulting network will be browsable for the user, so he/she can lookup terms, check related terms (hypernyms, synonyms, etc.), learn translations of a term in other languages, etc.

As a result of the combination of the three components we can link every document in the document base (a) to the semantic network (c) through the set of terms (b), thus obtaining a semantic organisation of the documents, based on the terminology contained in them. 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.

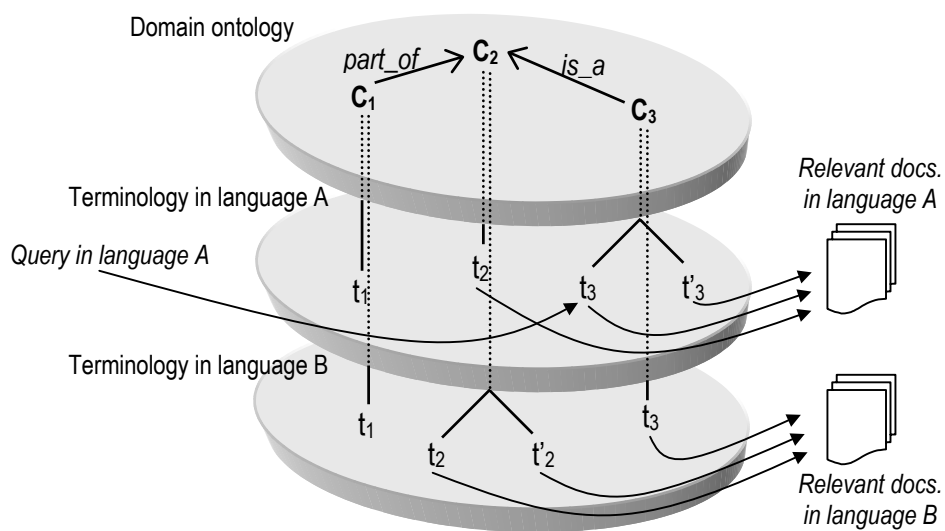


Figure 2. Linking documents and queries through a multilingually mapped ontology

The use of thesauri is closely linked to controlled vocabulary systems, well known for their effectiveness now for over 30 years (Salton 70). 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, which will keep the thesaurus updated with new terms found in the target corpus.

The methodology can be applied to any domain-specific text databases. One of the strongest aspects of the LIQUID system is its portability to other scientific and technical domains (currently, we are thinking mainly of areas of the medical domain, like gastroenterology). Once we have developed a methodology to link domain specific sets of terms to a semantic network, the creation of new semantic networks (if not existing yet) and the linking to specific terminology can be performed semi-automatically.

The main contribution of the system to actual users is query expansion. This system will perform a query expansion in two steps:

- Expand a given query with linguistic information, such as morphological variants ("Nissen fundoplication" and "Nissen funduplications") and semantically related terms, such as synonyms ("pancreatitis" and "irritation of the pancreas")
- Translate the query to other languages using the semantic network and expand it in these languages

4. Term detection or 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. A short introduction to current approaches

There are two major research trends in the field of terminology 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 93-96). However, strategies based on linguistic knowledge tend to produce "noise", i.e., 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 and Bourigault 00):

- Term recognition: "the identification of known terms"
- Term acquisition: "the automatic discovery of new terms"

Both activities refer to the automatic processing of text corpora as a source of terminology. (Jacquemin and Bourigault 00) provides a clear overview of the different existing terminology recognition or acquisition systems. In LIQUID, term extraction refers to both concepts.

4.2. The LIQUID approach

Our term extraction strategy is based on statistical 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 95) or (Justeson and Katz 95), will not be part of the strategy in order to ensure feasibility and portability. LIQUID will not start building term sets from scratch but from previously existing resources. Reusing previous efforts and ensuring coverage of most common terms are two reasons for doing so. Besides, other researchers, like (Jacquemin et al. 97, and Jacquemin and Tzoukermann 99), stress the fact that starting with a reference set improves the results of automatic term detection strategies. Since we start with a set of terms, the study of term variation becomes a key component (Daille et al. 00). 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 ability to detect variant and new terms will have two main benefits:

- 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
- facilitate the task of keeping the whole system (text databases and semantic networks) synchronised and updated as new documents are added.

There is a bias towards identifying terminological units in the form of phrases, rather than words, since phrases provide valuable context that may help to disambiguate ambiguous words and to cope with polysemy, a major problem in general-purpose CLIR (Hull and Grefenstette 96).

The term extraction tool works with four sets of terms:

1. Initial terms. Already existing terms (reused resources) provided to the extractor as starting point.
2. New terms. Those created automatically by the tool applying a set of derivation rules over the set of initial terms.
3. Valid initial terms. Subset of the initial terms that have been validated against the document base as good subject matter markers. The ratio between valid initial terms and initial terms provides a quantitative measure of the quality or relevancy of the initial term set regarding the document base.
4. Valid new terms. Subset of the newly generated terms that have been validated against the document base as good subject matter markers.

According to our previous definitions (Jacquemin and Bourigault 00): the detection of known terms is equivalent to "term recognition"; while the detection of new terms corresponds to "term acquisition". As we will see, variant terms represent an intermediate status as, on the one hand, they are equivalent to new terms from a statistical point of view but, on the other hand, are equivalent to known terms from a semantic point of view. The following diagram shows the overall workflow of the term extraction and validation process:

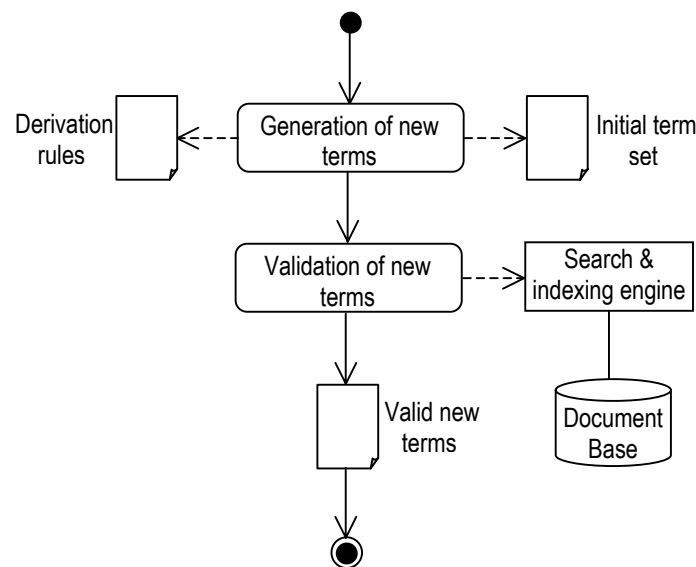


Figure 3. UML activity diagram of the term extraction and validation process

As expressed in the previous figure, the generation activity depends on two resources: a collection of known terms and a set of rules that applied over those terms produce new terms. Newly generated terms fall under two broad categories, each one divided itself into several subcategories:

1. *Variant terms*, i.e. terms that express the same concept as the term they derive from. They include the following types of changes or variations:

- 1.1. Morphological variations, identifying the root and its forms, like in:
 - "X-ray therapy" and "X-ray therapies"
- 1.2. Syntactic variations in the construction of terms, like in:
 - "HIV vaccine" and "vaccine against HIV"
- 1.3. Formal variations, like abbreviations or acronyms, like in:
 - "PAHO" and "Pan-American Health Organization"
2. *New terms*, i.e. terms that express a different concept than the one expressed by the term they derive from. Different strategies and linguistic knowledge are employed:
 - 2.1. Using known terms as a source, like extracting
 - "common bile duct obstruction", based on "common bile duct"
 - 2.2. Using suffixes, like "-itis":
 - "diverticulitis"
 - 2.3. Analysing other linguistic phenomena like co-ordination, as in the derivation of:
 - "stomach ulcer" and "duodenal ulcer" from "stomach and duodenal ulcer"

All these types of derivations have been expressed in derivation rules. In the next section we are going to describe the general framework of the rule-based terminology extraction system implemented in LIQUID.

4.3. 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 set of initial 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:

1. Typographical, such as the presence of a hyphen or an initial capital in the token.
2. Morphological, such as the property of number for nouns.
3. Syntactic, the part-of-speech of some categories.

Morphological properties are determined by means of simple suffix checking and applying highly productive heuristics. Because of their simplicity and the public availability 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. This approach could also be described as a form of stemming (Porter 80).

Regarding part-of-speech determination, and following the general approach towards cost-effectiveness and scalability, only function words (closed categories) are considered. 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 derived.

POS information has been introduced mainly to improve the readability of the rules, since no wide coverage mechanism for POS determination has been incorporated to the term extractor.

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 several tokens of the term, such as:

1. Re-ordering the token sequence
2. Joining two tokens in a single one

Or may affect to an individual token:

3. Remove/insert a certain token
4. Modify the typographical, morphological and/or syntactic properties of a certain 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, disregarding 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”

4.4. Validation of newly generated terms

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 generation process highly error-prone. 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 good 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, the criterion provides us with a 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 00) 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 candidate terms in our document base. The following figure shows the dependencies between these components:

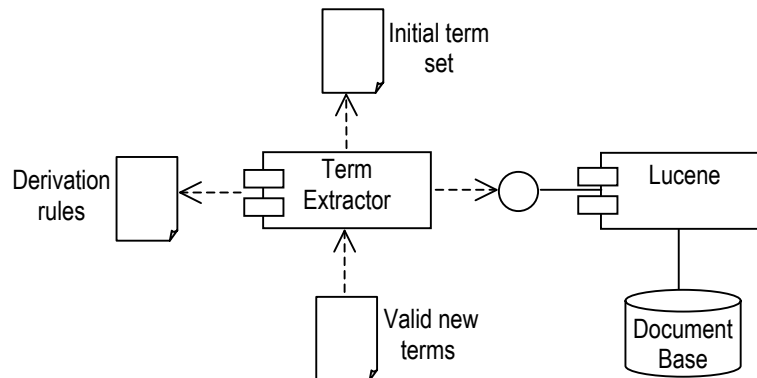


Figure 4. UML diagram of the components involved in the extraction and validation process

4.5. Experimental results

As an initial test of the approach being developed in LIQUID, we have taken the ELCANO document base. This base includes 563 medical articles belonging to the domain of gastroenterology and written in English. Each article is provided with several indexing terms (keywords) also in English. In total, 1222 initial terms are provided. We have written 30 derivation rules attempting to capture useful derivation patterns that will enrich substantially the initial term set. Newly generated terms have been validated against the ELCANO document base.

Our main performance metric is the ratio between valid new terms and (valid) initial terms, since it gives a quantitative measure of how successfully we have enriched the initial term set. Given that 525 newly generated terms have been found valid, the ratio is $525/1222 = 0.43$.

It must be noted that although the initial term set consists in the keywords found in the very same ELCANO articles used for validating the new terms, it is the case that a significant fraction of these keywords do not occur in any of the articles. This is clearly the case when the keyword has been mistyped (e.g. “Polypropylene mes” instead of “Polypropylene mesh”) or simply the keyword is not used as a regular word in any of the 563 articles (e.g. “eventration”).

This fact affects negatively to our performance metric, since derivations from an initial term that does not appear in the validation document base most often do not appear themselves. In order to account for this fact, we have validated the *initial* term set against the document base and have found that 874 of them were valid (conversely, that 29% of the initial term set do not occur in any article of our document base). The ratio between valid new terms and valid initial terms is then: $525/874 = 0.60$. This means that the initial set of *valid* terms has been enriched by 60%.

5. Conclusions and further work

We have presented project LIQUID, a solution to the problem of cross-lingual access to multilingual document bases in specific domains. The solution involves 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 on 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:

- Very low dependency on linguistic resources.
- Small set of linguistically motivated derivation rules.
- Incorporation of publicly available software tools.
- Exhaustive validation of the newly generated terms against a domain document base.

This approach has been tested in the domain of gastroenterology with a collection of documents and an initial set of indexing terms, both in English. In further work, we will pay 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 proving 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, such as LIQUID, that combine both approaches are a promising path that deserve to be tested in a wide variety of languages and domains.

6. References

- Arppe 95 - 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 and Croft 96 - 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 and Croft 97 - L. Ballesteros and W.B. Croft (1997). "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval", in Proceedings of ACM SIGIR Conference, 20: 84-91.
- Bourigault 93 - 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 95 - 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 et al. 96 - 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 et al. 2000) 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 96 - 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 00 – Goetz, Brian (2000) "The Lucene search engine", Java World. The official Lucene web site is located at: <http://jakarta.apache.org/lucene>.
- Gonzalo et al. 99 - Gonzalo, J., F. Verdejo and I. Chugur. (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 et al. 98 - Gonzalo, J., F. Verdejo, C. Peters and N. Calzolari (1998) "Applying EuroWordNet to Cross-Language Text Retrieval". *Computers and the Humanities, Special Issue on EuroWordNet*.
- Hiemstra 97 - Hiemstra, D., F.M.G. de Jong, and W. Kraaij. "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 et al. 00 - Hovy, E.H., N. Ide, R.E. Frederking, J. Mariani, and A. Zampolli (editors) . 2000. *Multilingual Information Management*. In press. Also available at <http://www.cs.cmu.edu/~ref/mlim> - 20-12-00.
- Hull and Grefenstette 96 - Hull, D., and Grefenstette G. (1996) "Experiments in Multilingual Information Retrieval". Proceedings of ACM, SIGIR'96. Zurich.
- Jacquemin et al. 97 - 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 and Tzoukermann 99 - 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 and Bourigault 00 - Jacquemin, C., Bourigault, D. (2000) "Term Extraction and Automatic Indexing". In R. Mitkov, editor, *Handbook of Computational Linguistics*. Oxford University Press, Oxford.
- (Jacquemin and Bourigault 2001) Jacquemin, C., and Bourigault, D. (2001). *Term Extraction and Automatic Indexing*. In R. Mitkov, editor, *Handbook of Computational Linguistics*. Oxford University Press, Oxford.
- Justeson and Katz 95 - 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 and Croft 90 - Lewis, D. and Croft, W. (1990) "Term clustering of syntactic phrases". In *ACM SIGIR-90*, 385-404.
- Oard 97 - Oard, D. (1997). "Alternative Approaches for Cross-Language Text Retrieval". *AAAI Spring 1997 Symposium on Cross-Language Text and Speech Retrieval*.
- Porter 80 - M.F.Porter (1980) "An algorithm for suffix stripping", *Program*, 14(3):130--137.
- Radwan and Fluhr 95 - 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 90 - Sager, J. C. (1990) *A Practical Course in Terminology Processing*. John Benjamins. Amsterdam.
- Sheridan and Ballerini 96 - Sheridan, P. and Ballerini, J. P. (1996). "Experiments in multilingual information retrieval using the spider system". In *Proceedings of ACM/SIGIR*.
- Yang et al. 97 - Yang, Y., J. Carbonell, R. Brown, R. Frederking (1998). *Translingual Information Retrieval: Learning from Bilingual Corpora*. *AI Journal Special Issue: Best of IJCAI'97*.