Mercedes, A Term-In-Context Highlighter

Raúl Araya, Jorge Vivaldi

Institut Universitari de Lingüística Aplicada
La Rambla, 30-32, 08002 Barcelona
{raul.araya;jorge.vivaldi}@upf.edu

Abstract

It happens very often that researchers in Terminology need to know about the terms included in a given LSP corpus. One possibility is to run a term extractor but in this case such a tool provides just term candidates, but not valid terms. Therefore, it is mandatory a term validation process that is not always easy and affordable. A different option is to take profit of the public lexical resources (dictionaries and glossaries) that are increasingly freely available through the Internet. To accept only such a terms as the valid ones is usually enough for most of the researches.

Mercedes, a term recogniser developed at the IULA, has been designed to fulfil this need. It integrates different public available lexical resources in order to provide researchers with a tool to show and highlight the valid terms (and its contexts) found in a certain text from a given domain. Mercedes consists of two main modules, the recogniser program and the dictionaries module. The recogniser can be run on specialised texts that have been previously tagged and lemmatised as part of the IULA's *Corpus Tècnic*. After running the recogniser, the user can then navigate the output with any web browser.

Introduction

Term extraction is commonly accepted as a hard task, a lot of different strategies have been envisaged by many researchers (see Kageura, et al., 1996 and Bourigault et al., 2001). All the existent systems just provide term candidates; therefore a stage of validation is mandatory. Only a few systems rank their candidates according to their termhood in order to help this validation task. In spite of these difficulties, a common problem for researchers in terminology is to find the valid terms present in their corpus. Frequently, these researchers are not specialists in the domain; hence they require the assistance of specialists in the domain under study to validate a list of term candidates. This is not always possible. Often it happens that a fully complete list of all the terms present in a LSP text is not necessary, at least in a first stage.

At the same time, even more and more institutions and researchers are making available their dictionaries /glossaries through Internet. Although such resources are not always complete, it may be considered that by combining a wide range of resources it is possible to obtain an acceptable list of already validate terms in a given domain. The idea behind Mercedes, the tool introduced in this paper, is to use such resources to find already validate terms in LSP corpora. It has been designed to provide researchers in Terminology with an easy to use tool that will show and mark terms in context.

The use of the Internet as a source for natural language data is not new (see Kilgariff, et al., 2003). It is being used to feed corpora with raw data due to its advantages over the mandatory "scanner-ocr" routine with paper published materials. Also, nowadays it is also fairly easy to find specialised glossaries in the Internet. Terminology glossaries and databases present very different interfaces: from the search engine model featured in multilingual projects such as Eurodicautom[1] or Unescoterm[2] to more flexible and accessible glossaries such as the UNAIDS Terminology Database[3]. We rely on search engines to find specialised glossaries, but we do a careful non-automatic study on the glossaries we find on the Internet, so that we can be sure that they contain validated terms. We cannot trust on search engine results for a given term as other researches have previously done (see Fujii et al., 2000).

Following this introduction, we present an overall description of the recogniser, some applications and processing examples. Finally we present some conclusions on the relevance of this tool as well as some future improvements.

Description of the system

Mercedes is based in two different modules: the term recogniser module and the dictionaries acquisition/management module. Both modules are used in conjunction with the dictionary database that actually contains the data. Figure 1 shows the general diagram about how all modules interact. In the next two subsections we will briefly describe the two main modules.

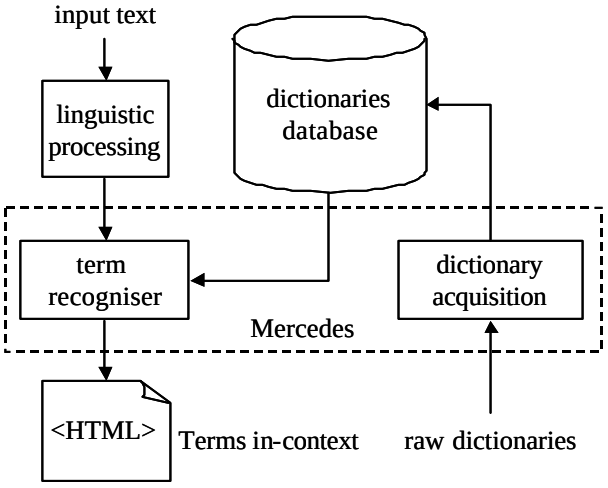

Figure 1. General architecture of Mercedes

[1] http://europa.eu.int/eurodicautom/Controller
[2] http://termweb.unesco.org/
[3] http://www.unaids.org/en/resources/terminology/

Term Recognition Module

The working procedure of this module is extremely simple. It takes as the input the text of the document to be processed and the dictionary selected by the user. Firstly, as usual in most of NLP tasks, the text must be previously tagged and lemmatised[4]. Secondly, the user must choose the language and the domain of interest. Currently, this module finds just nominal terms, which usually make up around the 90% of the total amount of terms in a specialised text. It means that this module keeps looking for nouns (and its correspondent expansion, if any). that are included in the chosen language and dictionary. This kind of proccessing does not allow to detect (left or right) coordinated terms. This procedure has proven no to be a limitation because this kind of construct are uncommon in our corpus.

Mercedes processes the text and produces a file in standard HTML 4.01 format; so that, later, the user can navigate it with any standards compliant web browser. The user may easily find for each found term all the sentences in which it is included. In such sentences this module highlights both the searched term and the other terms located in the same sentences.

Obviously, the quality in recognition fully depends on the quality of the terms included in the dictionaries. It may happen that not all the specialists in a given domain fully agree in the termhood of a given term. As Mercedes completely relies in the linguistic analysis of the text, any mistake in the tagging procedure may also affect the output quality.

Mercedes has been designed to be used with texts of a given domain with a high specialization level. Terms, as words, are lexical units that are potentially ambiguous. Therefore, if the user applies Mercedes to a non-specialised text it may happen that such lexical unit is used in a non-specialised sense.

Currently, Mercedes can recognize terms from the genomic domain and works with Catalan, English and Spanish written texts, but it has been designed to allow for inclusion of other domains.

Mercedes modular design allows for rapid extendibility. If the user needs to recognise and outline terms in context from a domain that is not currently supported by the system, the dictionary management system has been designed in such a way that the user can update the dictionaries database, incorporating new domain specific dictionaries previously extracted from other sources such as the Internet or translation memory databases.

In addition to the terms contexts, Mercedes provides the user with other useful information:

- Full text with all the terms highlighted
- List of all dictionaries (and a link to it when it is obtained from the web) used during the search
- Distribution of the terms along the text (absolute values and density)

Dictionaries management module

The dictionaries module manages the lexical information the recogniser program needs to fulfil its task. This dictionary management system has been designed focusing in rapid re-use of already available data. Being this one of the priorities, and taking into account the continuous growth of publicly available lexical resources on the Internet, the variety of HTML and XHTML designs implemented on these lexical resources and the lack of standards implementation on them (i.e. making use of RDF or RSS), we decided to develop our own exchange format following XML standards. By doing so it is easier for us to rapidly develop (or adapt) a Perl script for each new resource that we need to incorporate into the system rather than trying to develop a huge application capable of coping with this diversity of formats and designs.

Dictionary acquisition

Dictionary management has been designed in such a way that we can easily insert new dictionaries at any time. This way, researchers may look for new electronic lexicons publicly available to work on a specific domain and in a specific language.

To ease new dictionary insertion, a new DTD has been developed as a definition of an interchange format. We are using XML to ease data exchange, since XML transforms into other data formats should be fairly simple. This way we can reuse electronic lexicons and export them as data sources for other similar software applications currently developed at IULA.

One of our main dictionary sources is the Internet, but we also get access to published dictionaries through collaboration agreements[5]. In these cases, we convert the provided data into our XML exchange format. Whenever we need to update the dictionaries database with an Internet available lexicon, we perform the following steps:

- Analyse the lexicon format to confirm that we will be able to convert it into our own XML exchange DTD.
- Write a Perl script to harvest the lexicon parts we are interested in.
- Write a Perl script to convert it into our own XML exchange format.
- Parse the resulting XML file and incorporate it into the dictionaries database server.

With this approach, we can reuse the code generated for each script. Since the dictionaries we find on the Internet are on HTML format, it is easy for us to find how it is linked and modify the base harvester script to adapt it for a specific lexicon. Also, it is fairly easy to modify the base HTML parser script to convert from the original HTML format into our own XML exchange format.

In the final step, a parser checks first for formal XML integrity and then the script updates the dictionary database with the new entries.

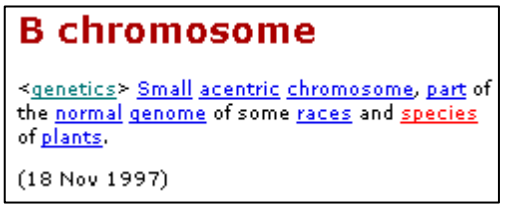

Figure 2. Entry for *B chromosome* on the CancerWEB Project as rendered by the browser.

[4] The texts should be processed as to those that take part of the IULA's LSP Corpus. See (Badia et al., 1998) for details.

[5] It has been the case, with the (Kaufmann et al., 1998).

Figure 2 illustrates the rendering of the B chromosome entry as shown on the CancerWEB Project web page[6]. We are interested in saving the term itself, the domain name and the definition. So after capturing the web page with the harvester script, we convert it into the XML exchange format. Figure 3 exemplifies the XML representation of the *B chromosome entry* in XML exchange format e have prepared.

```
- <entry>
    <term> B chromosome </term>
    <term-lemma tokens-cg="NULL"> B
    chromosome </term-lemma>
  - <definitions>
   - <def lang="en">
     - <p>
         <s> [genetics]</s>
       - <s>
           Small acentric chromosome, part of the
           normal genome of some races and species
           of plants.
         </s>
       </p>
     </def>
   </definitions>
 </entry>
```

Figure 3. Entry for *B chromosome* in our XML exchange format.

The XML exchange format.
We have defined our own DTD to store the following information:

- Term form,
- Term lemma, together with its grammatical information (if available),
- Definitions (as many as we find),
- Translations (as many as we find).

By resorting to XML, we can reuse lexical data in other similar projects. Either with XSLT transforms or implementing our own XML parsers, we can reformat this data conveniently.

Dictionary Database
Each entry is stored in the dictionaries database. We are using a MySQL database server running on a Fedora Core 1 Linux server. We store the entries on a per-domain-and-language table basis. That is, for each domain and language pair, we have defined a set of data tables to store term forms, term lemmas, their definitions, translations and source-dictionary related information. It is the task or our XML parser to create all the necessary table sets for new domain and language pairs.
The database schema has been designed so that we can keep track all term forms associated with a term lemma.

Results

The primary task of Mercedes is to provide Terminology researchers with easily readable raw data. Mercedes' output consists of a set of web pages that the user can navigate with any standards compliant web browser.
The user may query for contexts of a given term using a scrollable list. Mercedes will show all the contexts (at sentence level) that include such term. For each context, Mercedes will highlight not only the requested term but also all other terms (that is, terms it has in the database) that appear in the context). Figure 4 shows a typical result obtained by searching the term *genoma* (genome) found in a text of our corpus.
Currently this tool has succesfully contributed to several researches in the terminology field and has been adapted for a research in the conceptual relations area.

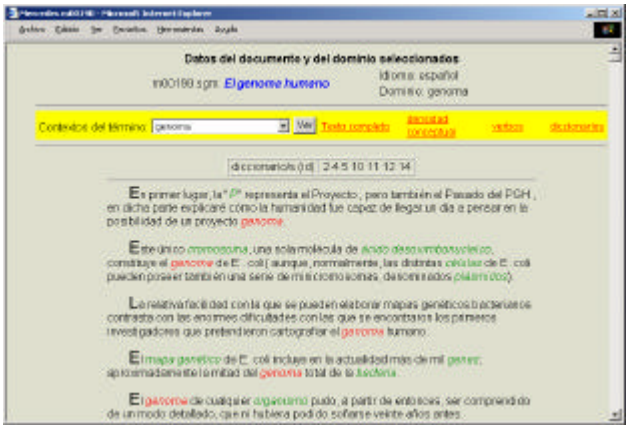

Figure 4. Term in context as shown by Mercedes

Behind the design of Mercedes there is the assumption that all the occurrences of all the terms in the text are in one of their specialised sense. This hypothesis may be tested using the precision and recall measurements. Unfortunately for doing such calculations it would be necessary that a specialist fully evaluates the whole document to find all the terms included in it. Due to the manual nature of this task, which makes it hard and time consuming, it has not been possible to complete it untill now.

Conclusions And Future Research

This paper shows Mercedes, a term-in-context highlighter. Its main characteristics are the following:

a) It is a system that makes use of lexical material publicly available, which ensures the user that recognised terms are already validated terms.
b) It has a modular architecture.
c) It is easy to extend since we are making use of standard technologies (XML for the exchange format).
d) The exchange format allows for rapid data sharing with projects and applications having similar needs.
e) It is not restricted to a domain specific input. Since it can be fed with domain specific dictionaries publicly available on the Internet, it can be put to work on different domains.
f) Its capability to re-use already available lexical data. Using XML for the exchange format, it would be relatively easy to transform, for example, TMX data into Mercedes own DTD.

Whenever a new Terminology researcher needs to investigate new domains, we will extend its range of

[6] http://cancerweb.ncl.ac.uk/

application by incorporating them. This is possible because Mercedes can be easily extended for use in other domains, since our parser already creates the necessary table sets to store new domain and language entries found in the input files.

Since we store as much lexical information as we can find (entry form, lemma, definitions and translations), we may reuse this information on other Terminology oriented projects and researches.

The data saved in the dictionary may also have a number of other purposes. For example, a term extractor based in the context needs to know which are the valid terms that are in the context of a given candidate. Usually, systems that take profit of this idea (see Maynard, 1999 and Vivaldi, 2001 for actual examples) use some kind of simplification based in their own resources and calculation. The terms of a given domain included in the dictionary may give a better support to this task.

Another example is given by those NLP tasks that use an ontology; in such cases, it may be useful to have some domain indication attached to the entries of such resource. (Vivaldi et al., 2004) proposes to use a domain vocabulary in order to add automatically domain information to EWN.

In the near future we will develop a web interface that will allow performing queries over Mercedes dictionaries database. This web application will let the users query the dictionaries entries, definitions and translations filtering by domain and language.

References

Badia, T.; Pujol, M.; Tuells, A.; Vivaldi, J.; de Yzaguirre, L. and M. T. Cabré, 1998. IULA's LSP Multilingual Corpus: compilation and processing. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98). Granada, Spain. Available at: http://www.iula.upf.es/corpus/corpubca.htm

Bourigault, D., C. Jacquemin and MC. L'Homme (eds), (2001). Recent Advances in Computational Terminology. Amsterdam: John Benjamins

Kageura, K. and B. Umino (1996) Methods of automatic term recognition: A review. Terminology. 3:2. (pp 259-289).

Kafmann U. and Bergenholts (1998). "Diccionario enciclopédico de ingenieria genética". Ontario: Lugus Libros Latin America Inc.

Kilgariff A. and Grefenstette (eds.),(2003).Computational Linguistics. Vol. 29 Num. 3.

Atsushi Fujii and Tetsuya Ishikawa (2000). Utilizing the World Wide Web as an Encyclopedia: Extracting Term Descriptions from Semi-Structured Texts. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), (pp. 488-495).

Maynard, D. (1999). Term recognition using combined knowledge sources. PhD thesis. Manchester Metropolitan University. Manchester.

Vivaldi, J., (2001). Extracción de Candidatos a Término mediante combinación de estrategias heterogéneas. PhD thesis. Universitat Politècnica de Catalunya.

Vivaldi, J. and Rodríguez H. (2004). Automatically selecting domain markers for terminology extraction. Proceedings of the Language and Resources Evaluation Conference (LREC2004). Lisbon. Portugal.