Automatic Construction of the Knowledge Base of an Onomasiological Dictionary

Gerardo Sierra*, Laura Hernández
Language Engineering Group, Engineering Institute
Universidad Nacional Autónoma de México, Ciudad Universitaria, México

ABSTRACT

For almost 14 years in the Language Engineering Group we have worked on a wide variety of Natural Language Processing (NLP) problems, being one of the earliest in the creation and operation of onomasiological dictionaries. During that time we have focused on search engine dictionary improvement, but recently our aim has been a development methodology for creating specialized onomasiological dictionaries in a semi-automatic way.

To automate the creation of onomasiological dictionaries necessarily implies the automatic execution of used processes to populate the dictionaries knowledge base. Due to the nature of these dictionaries, the definitions that must be included in the knowledge base are both normative and colloquial.

In this paper we present a proposal for semi-automatically populating the knowledge base of these dictionaries.

1 INTRODUCTION

An onomasiological dictionary is a dictionary that works in back to front way from "regular" or semasiological dictionaries. In onomasiological dictionaries users already know the definition of a term, but they do not know or have forgotten the name for that concept (this last problem is commonly known as *having a word on the tip of the tongue*) (Zock *et al*, 2011).

Onomasiological dictionaries have been classified into visual dictionaries, reverse dictionaries, thesaurus and synonym dictionaries. These dictionaries were created in order to solve the tip-of-the-tongue problem, but people still have difficulty using them because they require either that the user knows the precise words to describe the term, or its classification (i.e. when using a reverse dictionary to find the word *potato*, you might have to know that a potato is a *tuber*, and that tubers are a kind of *plant*). With visual dictionaries there is also the problem that not every concept has a visual image to represent it. For these reasons it has been suggested that free-text searcher ―also known as Natural Language searching― is a viable option for solving this problem (Lancaster, 1972) since they allow the user to describe their idea of the concept in the way they would use to explain it to another human.

The creation of onomasiological dictionaries that solve inputs written in natural language improves the user experience, but it creates some major challenges that the developers must handle (Dutoit *et al*, 2002 and Bilac *et al*, 2004). The most demanding task might be the one arisen from the different ways in which a person can express the same concept, and also the fact that user definitions might not match the formal definitions found in conventional dictionaries.

In short, natural language onomasiological dictionaries need a rich knowledge base which includes not only formal, but also informal definitions. Knowledge bases can be obtained from ontologies, like in the projects Genoma KB (Cabré *et al*, 2004) and ONTODIC (Alcina, 2009). However, given the main goal of onomasiological dictionaries, for this work we decided to extract their Knowledge Bases from definitions written in texts. These definitions, on the other hand, can be used not only to populate the Knowledge Base, but also to create ontologies (Sierra, 2008).

2 DEBO

DEBO is the first onomasiological dictionary developed in the Language Engineering Group and it works with user queries given in natural language. DEBO is a specialized dictionary and it was originally made as a dictionary of Natural Disasters, but today its structure and search engine has been extrapolated to other areas such as Linguistics, Metrology, Veterinary, and Sexuality.

2.1 The search method

The dictionary works with a search engine developed by Sierra (1999) and improved later by Hernández (2011). This engine is comprised by

- A number of *terms* of an area of specialization, which are the ones that the dictionaries can retrieve as a possible answer to the user's queries.
- A *knowledge base* that includes a variety of both normative and colloquial definitions.
- A set of *key words* extracted from the definitions and associated with the terms.
- A *stop list* that contains a catalog of "empty words", such as prepositions, articles and conjunctions.

* GSierraM@iingen.unam.mx

- A set of groups of words called *paradigms*, which are groups of words with similar meaning either in area of specialization or in regular speech.

The search method follows 5 steps:

(1) The system receives the query of the user as an *input*.

(2) The system analyzes de input and extracts its keywords by filtrering them with the aid of the stop list.

(3) The system searches among the paradigms the ones to which each keyword of the input corresponds.

(4) The system searches for terms that coincide in at least one paradigm with the input's one.

(5) The system retrieves the terms ordered by the number of paradigms that each term has in common with the input —in case of a tie, the system ranks each term according to the order in which the paradigms are presented in the definition against the input—. The terms are divided in "*very probable*", "*probable*" and "*not too probable*" columns.

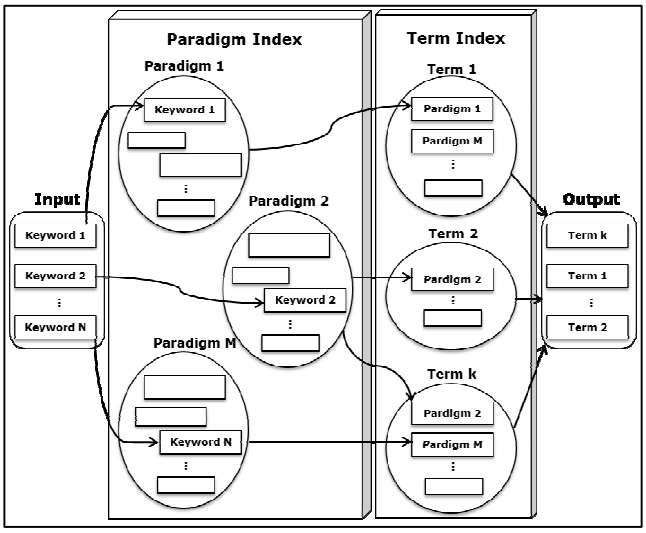

Fig. 1. Diagram of the search method.

For example, suppose someone enters as an input of the dictionary "*someone who hates gays*", and in the knowledge base there is the definition "homophobic: a person who despises homosexuals", and in the knowledge base there are also the following three paradigms.

Paradigm 1	Paradigm 2	Paradigm 3
someone	hate	gay
person	loathe	homosexual
people	contemn	lesbian
Individual	despise	queer
dude	abhor	dyke

Fig. 2. Example of paradigms in the knowledge base of an onomasiological dictionary of sexuality.

The user's definition is related to paradigms 1, 2 and 3, while one of the definitions of the term *homophobic* is related to the same paradigms in exactly the same order, which means that the term *homophobic* will be on the top of the output for this query.

2.2 The search engine performance

Hernández (2011) created the Onomasiological Dictionary of Sexuality for Mexican Spanish (DOS-MX) which used this search method. The knowledge base of this dictionary consisted of 975 both colloquial and normative definitions for 332 terms. All the definitions were found and retrieved manually from the Internet.

This dictionary had an added difficulty since it also had to be able to handle double-meaning words and phrases that are very commonly used in Mexico when talking about sex. In order to cope with this additional component, the dictionary's paradigms were extended to include double meaning words and even pejorative terms (see Paradigm 3 on Fig. 2), taking into account not only formal synonyms but also colloquial equivalents. In total there were over 33,000 different words organized in over 25,000 paradigms.

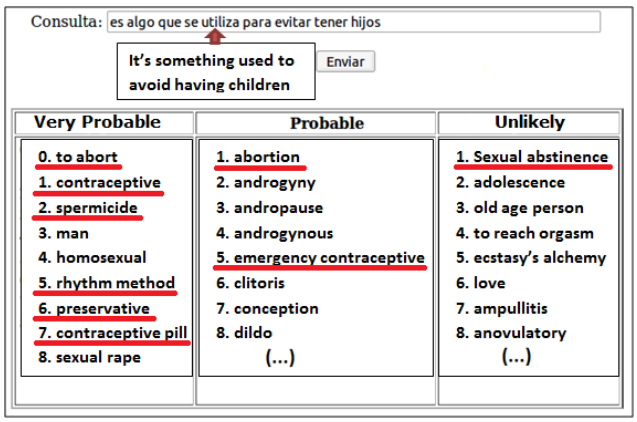

Fig. 3. Example of an output of the Onomasiological Dictionary of Sexuality for Mexican Spanish (DOS-MX)

There was an experiment to test the precision of the DOS-MX. This experiment consisted on making students write definitions of sexuality terms and to give their definitions to another student who wouldn't know which terms were described and would try to guess.

The precision of the dictionary was 71% when tested with natural language entries from users that weren't involved in the development of the dictionary, which is not bad compared to other non-English onomasiological dictionaries such as the one of El-Kahlout *et al* (2004) which has a precision of 66% in similar tests. However, a vast opportunity to improve exists.

2.3 A new improvement proposal

After the experience of the sexuality dictionary, it was concluded that the use of paradigms is not enough to try and cover all the ways in which a person can describe a concept. It was clear that there is a need to obtain many more different definitions in order to have a wide variety of expressions and ideas for every concept.

But increasing the number of definitions will also tend to increase the number of options from which the dictionary will have to choose, which is why there is also a need for organizing the definitions and terms into some sort of categories that will facilitate the selection of the correct terms.

The main problem then is to find a way to obtain a large number of definitions for the terms and classify them. This should be done in an automatized way, because by doing it manually will take too long and imply high resource usage.

3 ECODE

ECODE is a program that was developed in the Language Engineering Group with the objective of automatically detecting definitional contexts from specialized texts (Alarcón, *et al* 2008).

According to Alarcón, *et al* (2007), a definitional context is a textual fragment in which the definition of a term occurs. It is structured by a term and its definition, both being connected typographically by means of syntactic or typographic patterns.

These patterns in Spanish can be punctuation marks, such as comas, colons and parenthesis; verbs like *definir* (to define) or *significar* (to mean); discourse markers similar to *en otras palabras* (in other words), *o sea* (that is); and even pragmatic patterns like *en este conexto* (in this context) or *en términos generales* (in general terms). For example:

> *Desde el punto de vista de la sexología, se puede definir una relación sexual como el acto en el que dos personas mantienen contacto físico con el objeto de dar y/o recibir placer sexual, o con fines reproductivos.*
>
> (*From a sexology point of view, a sexual intercourse can be defined as the act in which two people have physical contact with the objective of giving and/or getting sexual pleasure or with reproductive purposes*)

The following features can be obtained from this example:

Term: "relación sexual" (*sexual intercourse*).

Definition: "acto en el que dos personas mantienen contacto físico con el objeto de dar y/o recibir placer sexual, o con fines reproductivos" (act in which two people have physical contact with the objective of giving and/or getting sexual pleasure, or with reproductive purposes).

Connecting verbal pattern: "se puede definir […] como" (can be defined as).

Pragmatic pattern as context modifier: "Desde el punto de vista de la sexología" (From a sexology point of view)

In order to automatically detect the features or components of a definitional context, Alarcón *et al* (2007) propose fifteen definitional verbal patterns divided into simple and compound ones (see Table 1).

Simple verbal definitional patterns	Compound verbal definitional patterns
• concebir (to conceive)	• consistir de (to consist of)
• definir (to define)	• consistir en (to consist in)
• entender (to understand)	• constar de (to comprise)
• identificar (to identify)	• denominar también (also denominated)
• significar (to signify)	• llamar también (also called)
	• servir para (to serve for)
	• usar como (to use as)
	• usar para (to use for)
	• utilizar como (to utilise as)
	• utilizar para (to utilise for)

Table 1. Definitional verbal patterns used by ECODE

The program processes the specialized texts and searches for definitional contexts. However, not every verbal definitional pattern that is found truly corresponds to a definition. There are some other expressions that use the same patterns with objectives other than give a definition. For this reason, Alarcón et al (2006) analyzed the use of these patterns in non-definitional contexts and found some sequences of words that are often used near a definitional verbal pattern.

Those sequences were found in some specific positions. For instance, some negation words like *no* (not) or *tampoco* (either) were found in the first position before or after the definitional verbal pattern; also adverbs like *tan* (so) as well as sequences like *no más de* (not more than) were found between the definitional verb and the nexus *como* (like); finally, syntactic sequences like adjective + verb were found in the first position after the definitional verb.

Once the system has eliminated non-definitional contexts, it proceeds to identify the features that form the definitions. For this, it uses a decision tree based on regular expressions which allows the system to identify and tag the position of every feature. These regular expressions are:

Term = BORDER (Determinant) + Noun + Adjective. {0,2} .* BORDER

Pragmatic Pattern = BORDER (sign) (Preposition | Adverb) .* (sign) BORDER

Definition = BORDER Determinant + Noun .* BORDER

The whole process of definitional contexts detection is shortened in the following diagram.

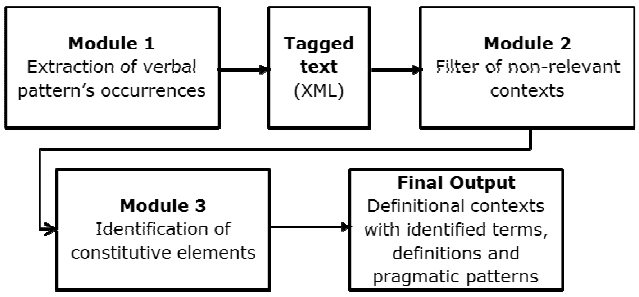

Fig. 4. ECODE architecture (taken from Alarcón, 2006)

4 DESCRIBE

ECODE was originally developed as a definitions extractor from specialized texts. However, the same definitional verbal patterns that are used in formal documents are also used in informal ones.

With this in mind, the Language Engineering Group has been working on the development of DESCRIBE, an extended scope of ECODE which extracts definitions from texts written in colloquial language.

This new adaptation consists on a module that automatically extracts search results from the Internet about a particular term, and then retrieves the contents of the web pages that match that search and analyses them looking for new definitions.

This tool removes all definitions that are repeated, and looks not only in formal websites, but also in open forums, personal webpages, blogs and chats, which provides a rich variety of definitions.

In the end, DESCRIBE retrieves a list of definition candidates that still have to be depurated, since some of the candidates might not really be definitions.

5 DEFINITION CLASSIFICATION

In order to give the dictionary search engine another feature to help the correct identification and ranking of output terms, it has been considered classifying the definitions to match not only the words and the order in which they appear, but also the type of definition given by the user and the ones in the knowledge base.

There are four kinds of definitions based on the Aristotelic definitional model: analytic, extensional, synonymic and functional (Sierra 2008). According to the LingualLinks Libray, the first one refers to "a description of the range of reference of a lexical unit" that allows readers to distinguish the term from similar words; the second kind refers to those definitions that list the objects that fall under the definition or its parts; the third kind uses synonyms or generic terms to describe the term; and finally, the fourth kind of definitions describes the term by providing its uses.

In order to automatically provide a category for each definition obtained through DESCRIBE, the verbal patterns have been divided accordingly to the kind of definition in which they usually appear.

Definition type	**Verbal definitional patterns**
Analytic	• concebir (to conceive) • definir (to define) • entender (to understand) • identificar (to identify) • significar (to signify)
Extensional	• consistir de (to consist of) • consistir en (to consist in) • constar de (to comprise)
Synonymic	• denominar también (also denominated) • llamar también (also called)
Functional	• servir para (to serve for) • usar como (to use as) • usar para (to use for) • utilizar como (to utilise as) • utilizar para (to utilise for)

Table 2. Definition types and their definitional verbal patterns

This definition classification is the first step in the ontology creation since, for instance, analytical definitions allow us to obtain hyponym and hypernym relations, while from extensional definitions meronymy and holonymy relations can be recovered (Soler *et al*, 2008).

6 DEFINITION CANDIDATES' DEPURATION SYSTEM

As most systems in Natural Language Processing, DESCRIBE is not perfect and sometimes the definition candidates turn out to be wrong, or the definition might be misplaced in a particular category.

The Language Engineering group has developed a tool to help the manual revision of the definition candidates' validity and their categorization correctness. This tool presents a series of definition candidates to dictionary developers. Every candidate shown to the user has also the category in which DESCRIBE placed it.

The system allows the developers to easily accept or reject a candidate as a definition and it also allows them to change the category into which the definition was originally placed.

This system helps in the task of polishing the definitions that will be part of the knowledge base of the dictionary, but it also keeps a record of the definition candidates that have been rejected. This record is intended to be used as a corpus that will serve as training data for a machine learning system that will be used to improve the precision of ECODE and, in consequence, of DESCRIBE itself.

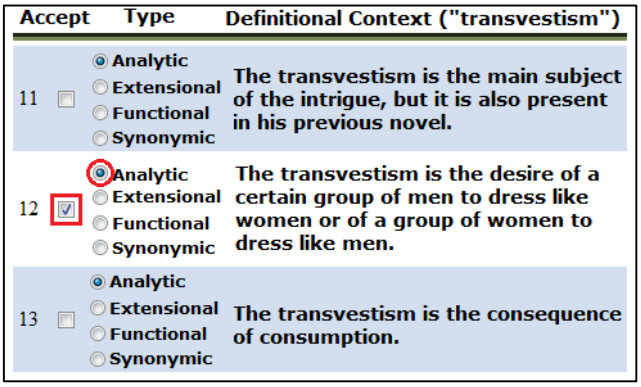

Fig. 5. Example of the use of the Definition Candidates' Depuration System.

CONCLUSIONS

The definitions included in the knowledge base of specialized onomasiological dictionaries must cover both formal and informal concepts, and they also must cover as many forms of expressing them as possible in order to procure more accurate solutions for its users.

It is also convenient to classify the definitions in the knowledge base and the ones given by the user according to their type, so as to provide the search engine with more features to compare and match the user definitions with its own, hence improving its precision. Definition classification is the first step in the creation of ontologies.

In this paper we presented a methodology to automatically obtain definition candidates to fill the knowledge base of onomasiological dictionaries and also classify these definitions according to the Aristotelic definitional model. The source of these definitions is the Internet, which allows us to a very wide variety of speakers and, for that reason, a means of expressing concepts. This methodology has been used and tested in the creation of onomasiological dictionaries of Sexuality and Linguistics, among others, and can be applied to other subject areas, such as Biomedicine, Epidemiology, Veterinary, Laws, etc..

We also presented a tool which will make possible the creation of a corpus with both good and bad definition candidates marked as such. The purpose of creating this corpus is to obtain training data for a machine learning system directed to improve the automatic detection of definitional contexts.

ACKNOWLEDGEMENTS

We would like to acknowledge DGAPA for the sponsorship of the project "Análisis estilométrico para la detección de similitud textual". We also thank the CONACYT Thematic Network "Tecnologías de la Información y la Comunicación".

REFERENCES

Alarcón, R. (2006). *Extracción automática de contextos definitorios en corpus especializados. Propuesta para el desarrollo de un ECCODE (extractor de candidatos a contextos definitorios)*. Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra, Barcelona (Doctoral thesis):

Alarcón, R., Bach, C., and Sierra, G. (2008). *Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica*. Revista Española de Lingüística **37**, 247-278.

Alarcón, R., and Sierra, G. (2006). *Reglas léxico-metalingüísticas para la extracción automática de contextos definitorios*. Avances en la Ciencia de la Computación, VII Encuentro Nacional de Ciencias de la Computación, 242-247.

Alarcón, R., Sierra, G., and Bach, C. (2007). *Developing a Definitional Knowledge Extraction System*. Proc. 3rd Language and Technology Conference (L&TC'07), Adam Mickiewicz University, Poznan, Polonia.

Alcina, A. (2009). *Metodología y tecnologías para la elaboración de diccionarios terminológicos onomasiológicos*. Terminología y sociedad del conocimiento. Bern: Peter Lang, 33-58.

Bilac, S., Watanabe, W., Hashimoto, T., Tokunaga, T. y Tanaka, H. (2004). *Dictionary search based on the target word description*. Proceedings of the Tenth Annual Meeting of The Association for Natural Language Processing (NLP2004), 556-559.

Cabré, M. T., Bach, C., Estopà, R., Feliu, J., Martínez, G., and Vivaldi, J. (2004). *The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities*. 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisboa, European Languages Resources Association, 87-90.

Dutoit, D. y Nugues, P. (2002). *A Lexical Database and an Algorithm to Find Words from Definitions*. Proceedings of the 15th European Conference on Artificial Intelligence, Lyon, 450-454.

El-Kahlout, I., and Oflazer, K. (2004). *Use of Wordnet for Retrieving Words from Their Meanings*. 2nd Global WordNet Conference, Brno, Czech Republic.

Hernández, L. (2011). *Creación semi-automática de la base de datos y mejora del motor de búsqueda de un diccionario onomasiológico*. Universidad Autónoma de México (Master thesis).

Lanacaster, F (1972). *Vocabulary control for information retrieval*. Washington: Information Resources Press.

Sierra, G. (1999). *Design of a concept-oriented tool for terminology*. UMIST, Manchester (Doctoral thesis).

Sierra G., Alarcón R., Aguilar C., and Bach C. (2008). *Definitional verbal patterns for semantic relation extraction*. Terminology 14(1), pp. 74-98.

Soler, V., and Alcina, A. (2008). *Patrones léxicos para la extracción de conceptos vinculados por la relación parte-todo en español.* Terminology 14(1).

Zock, M., and Rapp Reinhard (2011). *Introduction to this special issue on Cognitive Aspects of Natural Language Processing*. Journal of Cognitive Science 12(3).