METHODOLOGY FOR LANGUAGE ANALYSIS AND GENERATION IN CLOSED DOMAINS: PHARMACEUTICAL LEAFLET

Jesús Cardeñosa, Carolina Gallardo, Adriana Toni

Abstract: The best results in the application of computer science systems to automatic translation are obtained in word processing when texts pertain to specific thematic areas, with structures well defined and a concise and limited lexicon. In this article we present a plan of systematic work for the analysis and generation of language applied to the field of pharmaceutical leaflet, a type of document characterized by format rigidity and precision in the use of lexicon. We propose a solution based in the use of one interlingua as language pivot between source and target languages; we are considering Spanish and Arab languages in this case of application.

Keywords: Controlled languages; Interlingua-based Machine Translation

ACM Classification Keywords: I.2.7 Natural Language Processing

Conference: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008

Introduction

Most of translated publications belong to a technical, commercial or management field. This discussion only ratifies the challenge supposed for automating, as far as possible, this kind of document translation, because due to its nature, they only require translations that could be described as mechanic or routine. Literary texts, or generally the natural language, escape to the efforts of computer treatment because it is difficult to integrate the contextual knowledge that the speaker has, which is the only one that in many cases can solve the ambiguity that cannot be solved at a syntactic or semantic level. The quality of automatic translation remarkably improves when the advantages offered by specialty languages can be taken, regarding its precision, possibility of standardization, and vocabulary limitation. The automation of the translation requires, in this case, a laborious initial linguistic analysis of a corpus (texts of the domain that by its number and variety may be representative), but this initial effort is compensated by the time saved if the post-edition of produced documents is not needed.

Within the text translation pertaining to specific thematic fields (technical manuals, weather forecast, reports, legal texts, etc.), we must still distinguish between free or fixed format texts, understanding by fixed format a document structure divided into sections, with specific headings whose type content is known. The biggest structuring of the document is again an advantage when automating the translation, because it reduces the possible ambiguity of the terms to translate since there is information about the context where they appear.

The approach to the Automatic Translation (AT from now on) by means of the use of interlinguas consists of using an intermediate representation of the contents to translate, independent of source and target languages, from which the text is generated. One of the greater difficulties of this process lays in the definition of an interlingua that can work as an intermediate representation between any of the two languages. In fact, it is a new language requiring a definition of all of its components, with the additional challenge that being an artificial language it has to be as expressive as the natural languages. None of these systems have been successfully developed in text translation regarding domains with opened formats due to the difficulty in the interlingua design.

An advantage of these systems is the possibility of incorporating new languages without affecting the modules already developed for the other languages – it is necessary an encode module and a decode module (*to* and *from* the interlingua respectively) for each language. Figure 1 shows all the necessary modules to cover all possible pairs among A, B, and C languages.

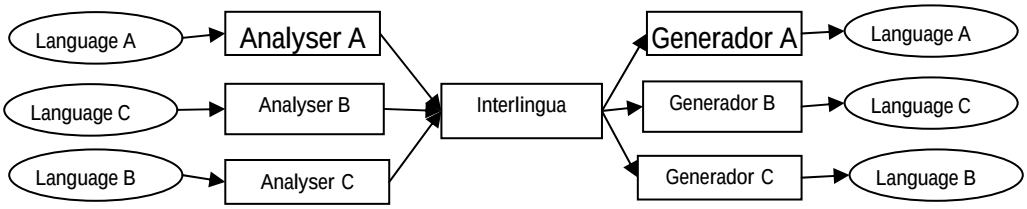

Figure 1. Modules for analysis and generation among 3 languages in an interlingua-base MT system

In this paper we are presenting a methodology devised to approach automation of analysis and language generation in a well defined and limited domain —pharmaceutical leaflet, with Spanish as source language and Arab as target language.

A pharmaceutical leaflet constitutes a text example pertaining to a closed domain with a fixed and standardized format. We are proposing the use of translation based in interlingua, because it is the only one guaranteeing a very precise coherence between the different language versions: the *Universal Networking Language* (UNL) a computing language developed as an essential element of UNL Project, an international project, promoted by the Institute of Advanced Studies of the United Nations (UNU/IAS).

It has software programs that allow introducing Spanish contents to UNL and from UNL to Arab. They work in interaction with linguistic resources typical of each one of the languages, stored in an electronic format: Grammar rules for language analysis and Spanish/UNL dictionary in the Spanish case; and Grammar rules for the language generation and UNL/Arab dictionary in the Arab case. The final mission is to adapt rules and dictionaries to the leaflet domain, so that the quality level of the produced translations may be acceptable, not requiring a post-edition process. This project is in phase of accomplishment, and the work plan — that we will expose in section 4 of this article—, has been born from the study of the difficulties posed by the dominion and characteristics of the source and target chosen languages.

Section 2 introduces UNL language and generally the translation process by means of an interlingua. In section 3 we give a brief explanation of software tools referred in the work plan. The most relevant contribution of this article, however, goes beyond the interlingua specifically chosen or the tools. The work plan could be carried on using other interlingua or other tools, because basically their functionality would have been the same one.

UNL Project: Aims and Components. Description of Translation Process

The UNL Project is born with the aspiration of developing an interlingua system to support a multitude of languages without any domain or lexical type restriction. It is an international project developed by the Institute of Advanced Studies of the United Nations together with research groups throughout the world, such as the Spanish Language Centre (CLE, `http://www.unl.fi.upm.es`), to which the authors of this paper belong. Their main objective is the dissemination, promotion, and formation in UNL technology with the aim to eliminate linguistic barriers in the Internet.

The UNL is composed of three main elements: Universal Words, Relations and Attributes. Formally, a UNL expression can be viewed as a semantic net, whose nodes are the Universal Words, linked by arcs labelled with the UNL relations. Universal Words are modified by the so-called attributes. UWs are a key element for UNL. A UW is intended to express a concept found in any natural language. To do that, UNL uses words and phrases taken from English but these English words are modified by semantic restrictions in order to eliminate ambiguity present in the vocabulary of natural languages. Thus, UWs are linked to the vocabulary of natural languages. The reason for choosing English is just practical: the inventory of the English vocabulary is rather well covered by many authoritative dictionaries and there are bilingual dictionaries of English to almost any other natural language.

Next we present an overview of the translation process by means of the UNL. For more detailed information see the references suggested in the bibliography (section 5).

The Analysis Process or Enconversion

This process consists in putting contents from a natural language to the UNL. The essential resources in this phase of the process are:

- A set of Grammar rules linked to the source language to transform the contents written in that language into UNL. Each language requires the development of a specific set of rules.
- A dictionary where each word of the source language having a semantic meaning is linked to a UW — again each language requires the development of such dictionary. The following pairs illustrate how a different UW is linked to each sense of the word "state" in order to disambiguate its meaning:

  Pair 1: <estado, state(icl>administrative_district>thing, equ>country)> denoting the territory occupied by a nation.

  Pair 2: <estado, state(icl>political_unit>thing)> denoting a politically organized body of people under a single government.
- The *Enconverter* is the component – a *parser* - that allows automatically passing the content from a source language to UNL. It interacts with the set of Grammar rules and the dictionary already mentioned in the previous section.

The UNL produced texts come to be part of the "Base Documental UNL", containing documents written in UNL language which are available to the user community of UNL System.

The Generation Process or Deconversion

This process consists in putting contents from UNL to a target language. The elements taking part in the process are:

- A set of Grammar rules for the generation, beginning from the UNL linked to a target language. Each language requires the development of its specific set of rules.
- A dictionary that links each UNL word to a word in the target language (same dictionary used in the process of *Analysis*).
- The *Deconverter* is the "opposite" component to the *Enconverter.* It is responsible from passing the UNL to a natural language. The set of programs forming the *Deconverter* works with the rules and dictionary already mentioned in the previous section.

For the Spanish language there are several thousand rules for the *Generation* process in Spanish language, that provide a rather acceptable cover up for the Spanish generation from contents written in UNL.

The Research Centres and Universities working together with the UNL/IAS Centre in the UNL Project are responsible for the development and permanent updating of the resources own to their respective languages — Grammar rules for the analysis and generation, and dictionaries — besides supervising the correct application of the UNL standards, maintaining the language servers for the testing tools, giving technical support to the users, content providers and builders, and in general to carry on all the necessary activities to promote the UNL system.

Software Tools of the Spanish Language Centre

The Spanish Language Centre represents the UNL Center for the Spanish language, and must help it in the programming, coordination, support, financing, research, and formation of the whole UNL System. It includes all the languages whose roots are related to the Spanish language, namely, all the Spanish-speaking countries and developments affecting the indigenous languages of Latin America. Most of the CLE members are also

researchers of the Grupo de Validación y Aplicaciones Industriales (VAI, http://www.vai.dia.fi.upm.es) of the Facultad Informática, Universidad Politécnica de Madrid.

Next, the functionality of the software tools, already mentioned in our work plan, is briefly explained. The tools have been developed in the VAI laboratory.

Generator of Universal Words: it is used in the *Analysis* phase in particular in the identification and extraction of the words from the document in Spanish language that have a semantic content, and the construction of the corresponding universal word.

UNL Editor: it is a platform that integrates the components used in the *Analysis, Verification,* and *Generation* phase, allowing carrying on these processes in a comfortable and unified way from a unique program. Basically, the environment consists of a central module (UNL Editor) that controls and coordinates the rest of the system components; see Enconverter, Deconverter, rules and dictionaries. The functionalities of this editor are:

- Manual and automatic analysis of documents in a given language
- Text edition and graphic structures
- Validation and verification of UNL code
- Generation of a code in any language from UNL
- Access to word dictionaries

Work Plan: Pharmaceutical leaflets

Next, it is presented a list of tasks whose fulfillment will provide us the resources (programs and linguistic resources in an electronic format) needed to automate the translation of the pharmaceutical leaflets from Spanish to Arab.

Among the different tasks needed to approach, we emphasize two as the most important:

- Exhaustive study of linguistic characteristics, characterizing the leaflets. A deeper study will result in a higher lucidity in conclusions, more quality in translations and greater automation of translating process, since we will be able to anticipate every difficulty that may arise. This study will be carried on from a very numerous and varied number of leaflets —that we shall compile— in order to obtain general conclusions
- Adaptation of linguistic resources managed by the programs in charge of translation, to the outcomes of the study. We are considering sets of syntactic and morphologic rules, linked to Spanish and Arab languages, that interacting with the parsers and generators we have, will determine the translation from Spanish to UNL and from UNL to Arab respectively, and also bilingual dictionaries Spanish/UNL, and UNL/Arab.

A first examination of a reduced number of pharmaceutical leaflets in Spanish language shows us that there are several variations in their format, regarding their section division as well as their extension — reduced or extended versions — being the latter the most widely used at a commercial level. Differences are determined by issues as: type of symptom or disease for which medicine is prescribed, administration form — syrup, pills, etc. — and the manufacturing laboratory. Therefore, before compiling the set of leaflets that will constitute the corpus of the work, we will establish the quantitative criteria — how many will be enough? — and the qualitative criteria that they must fulfill in order to make them representative of the different types of leaflets found in the market. We must as far as possible include all the different groups of medicines regarding their pharmaceutical classification, and within each group, cover all administration forms, and also include medicines from the greater possible number of laboratories.

Once compiling and storage is finished, we will proceed to study the corpus. First, we pay attention to the documents structure, and we find a basic core of sections (“Indications”, “Composition”, “Administration”, etc.) common to all of them. Variations come mainly from the manufacturing laboratory and type of product (symptom or fighting disease). The format study concludes relating each epigraph to the type of content it includes. It is important to identify the vocabulary, sentences etc., that appear linked to each epigraph, because it helps us to

disambiguate meanings in the process of translation and to recognize new words (is it a “component”, a “secondary effect”, a medicine name?).

As to the linguistic study, due to the previous knowledge of the difficulties posed in that sense by automatic translation, and by detailed observation of the texts, we are synthesizing the most relevant aspects to be considered:

- Lexical aspects: as the degree of semantic ambiguity of the used words; estimation of specific vocabulary ratio compared to common vocabulary; or the guidelines in the words composition by frequently used suffix and prefix.
- Syntactic aspects: as the existence of syntactic constructions linked to the different sections and contents (type of sentence to express contraindications, or administrative instructions, etc.); and the observation of prepositional ambiguity.

The conclusions obtained in the phase of corpus study will be used for the adaptation of the syntactic and morphologic rules that we have in order to put contents from Spanish to UNL, and from UNL to Arab. Because of the type of vocabulary included in the leaflets, we shall emphasize the following aspects:

- Recognition of unknown words
- Recognition of proper names
- Recognition of morphologic composition processes.

We must also adapt the available Spanish/UNL and UNL/Arab dictionaries, incorporating new vocabulary own of the domain and disambiguating the translations of common words, which appear with certain frequency in the leaflet field. The updating of the dictionaries must be done according to the rules of general dictionaries and UNL specifications.

Passing to expose the work plan with more detail, we distinguished 5 great tasks whose content we will be refining according to the difficulties posed. Basically, these tasks must be successively developed, because the results of each task are used by the following ones. Figure 2 presents the dependencies diagram and priorities of the identified tasks and subtasks. After it, we are including an explanation table of the required objectives and resources in each task.

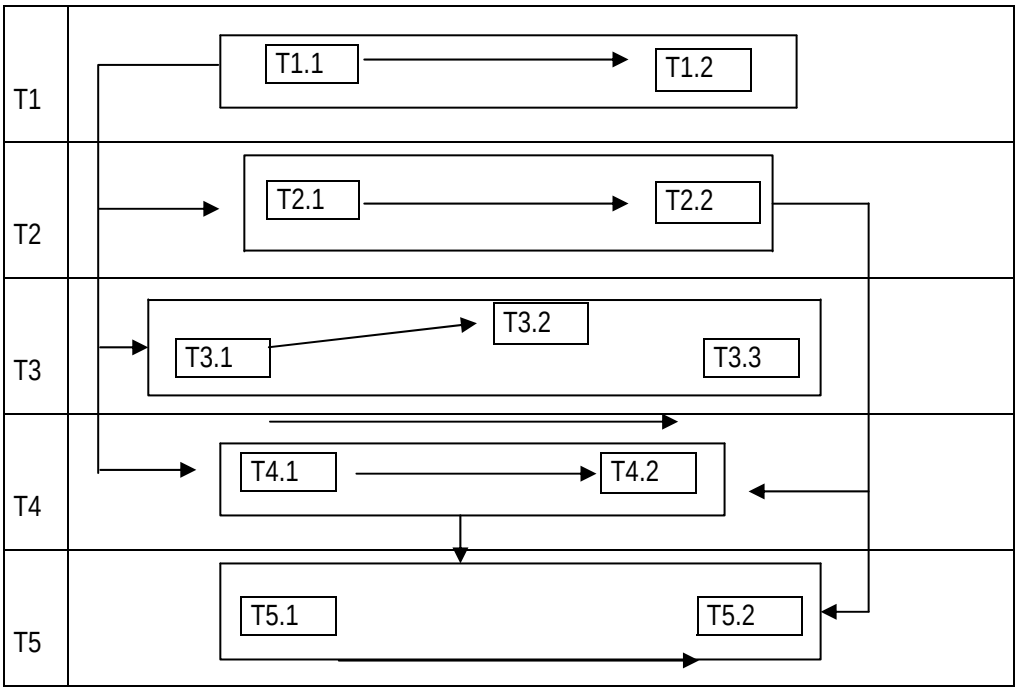

Figure 2. Diagrama de dependencias

Task identifiers stand for:

- T1.1: Criteria for the selection of texts
- T1.2: Obtaining and classification of texts
- T2.1: Analysis of text structure
- T2.2: Linguistic analysis of texts
- T3.1: UWs production
- T3.2: Update of Spanish/UNL dictionary
- T3.3: Update of Arab/UNL dictionary
- T4.1: Review of resources for Spanish language
- T4.2: Codification of texts into UNL
- T5.1: Review of resources for Arabic language
- T5.2: Generation into Arabic language

Detail of the tasks

TASK 1: CORPUS CREATION

Aims: compile a set of leaflets in Spanish language in order to constitute a representative corpus. To do that, it is necessary to establish how many are needed and the qualitative criteria that the chosen leaflets must fulfill (T1.1), and subsequently, proceed to their obtaining, classification, and storage (T2.2).

Resources: Access to leaflets in Spanish language (in electronic format).

TASK 2: CORPUS STUDY

Aims: study of the leaflets structure —section division, their content—- and the international rules and standards that may exist regarding this (T2.1). Also, the linguistic characteristics of the texts as to the lexical employed — polysemy, general or specific vocabulary, suffix, and prefix— also, syntactic construction own to each section, and the possible prepositional ambiguity will be studied (T2.2).

Resources: Access to the corpus of the selected texts.

TASK 3: PRODUCTION OF DICTIONARIES

Aims: To produce dictionaries adapted to the vocabulary employed in the leaflets. The first step (T3.1) is to automatically obtain the pair list [Spanish word/UW] for all those words having a semantic content that appear in the leaflets. The next step (T3.2) is to add the list of pairs to the general Spanish/UNL dictionary, adapting, if needed, the list of attributes describing each word to the domain. Finally (T3.3) a dictionary [UNL/target language] must be produced, from the list of the obtained UWs in T3.1. Again, any existing dictionary of general purpose, if there is one, may be reused, adapting the attributes to the leaflet field.

Resources: Generation Tool of UWs (T3.1), text editors to adapt the text format of the corpus to the one required by the Tool, dictionaries of general purpose pairing UNL with Spanish and Arab languages.

TASK 4: SPANISH ANALYSIS AND UNL ENCODING

Aims: to generate a UNL version of the leaflets in Spanish. In the first place, the base of Spanish analysis rules will be adapted according to the lexical studies and the syntactic phenomena, identified when making the corpus study (T4.1). The enhancements will be focused in the rules adaptation to the unknown words, proper names, processes of morphologic composition, and treatment of syntactic structures identified when making the linguistic study. If it was necessary, ad hoc rules and specific attributes of the domain will be incorporated (the latter will also require the modification of the attributes of the domain dictionary). After reviewing the rules and dictionary, the UNL encoding (T4.2) of the leaflets will be automatically obtained.

Resources: Enconverter, UNL Editor, texts' editors to adapt the texts' format to the corpus required by the Enconverter.

TASK 5: GENERATION OF LEAFLETS IN ARAB

Aims: to generate the Arab version. In the first place (T5.1) the Grammar rules must be adapted for the generation of Arab language and the [UNL/Arab] dictionary with the aim of allowing, as far as possible, producing understandable and correct texts. The last phase of the task and the global process – excluding final tests and evaluations – is to generate the Arab version in the selected leaflets (T5.2). This is a totally automated process, and there will not be any type of post-edition in the generators' outcome.

Resources: Deconverter, UNL Editor, texts' editors to adapt the texts format to the one required by the Deconverter.

We identified an additional task of tests that would consist in assessing all the generated resources along the process. We described in the summary diagram the type of tests that must be carried out and the software tools in order to accomplish them (see section 3 for a summary description of tools functionality).

TASK 6: TESTS AND ASSESSMENT OF RESULTS

Aims: To test and asses the adaptation of all the resources to the awaited results. The UNL Editor will help us to analyze the UNL generated code (T6.1). We will use the inference engine to verify that the corresponding set of rules and dictionaries allow adequately putting of contents (T6.2). Finally it will have to be verified the legibility and correctness of the translated version to the target language from the corpus texts (T6.3).

Resources: Editor UNL, Grammar rules of Spanish analysis, rules of Arab generation and dictionary, set of translated texts to Arab.

Conclusions

This paper describes a methodology of work to approach the translation based in interlingua among different pairs of languages in closed domains. The use of an interlingua allows us, on one hand, the total reuse of the analysis modules and the generation of language already existing before the inclusion of new pairs. On the other hand, to narrow down the problem to closed domains simplifies the analysis and generation tasks, producing results of better quality in comparison with those obtained from opened domains.

This methodology has been applied to the treatment of pharmaceutical leaflets; however, it would be equally valid for the treatment of any type of texts within any domain with a controlled language.

Bibliography

[Boguslavsky et al, 2005]. Boguslavsky, I., Cardeñosa J., Gallardo, C., and Iraola, L. The UNL Initiative: An Overview. Lecture Notes in Computer Science. Volume 3406/2005, pp 377-387. Springer Berlin / Heidelberg: 2005. ISBN 978-3-540-24523-0

[Fellbaum, 1998]. Fellbaum, C., (ed): WordNet: An Electronic Lexical Database. Language, Speech, and Communication Series, MIT Press (1998)

[Uchida et al, 2005] Universal Networking Language (UNL). Specifications Version 2005. Edition 2006. 30 August 2006. http://www.undl.org/unlsys/unl/unl2005-e2006/

Authors' Information

Jesús Cardeñosa – Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: carde@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es

Carolina Gallardo – Group of Validation and Industrial Applications. Escuela Universitaria de Informática. Universidad Politécnica de Madrid. Cta de Valencia Km.7. 28041 Madrid; email: cgallardo@eui.upm.es. http://www.vai.dia.fi.upm.es.

Adriana Toni – Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: atoni@fi.upm.es. http://www.vai.dia.fi.upm.es