Incorporating Subject Areas into the Apertium Machine Translation System

Jordi Duran, Lluís Villarejo, Mireia Farrús, Sergio Ortiz, Gema Ramírez

Abstract. The Universitat Oberta de Catalunya (Open University of Catalonia, UOC), is a public university based in Barcelona. The UOC is characterised by three main factors: (a) it is a virtual university based in an e-Learning model, (b) it is based in a strongly Spanish-Catalan bilingual region, and (c) students come from around the world, so that linguistic and cultural diversity is a crucial factor.

Within this context, it becomes essential to meet the UOC's linguistic needs taking into account its particular characteristics. One of the tools created to this end is the adaptation of Apertium, a free/open-source rule-based machine translation platform, which can be found under http://apertium.uoc.edu/, customised to the translation needs of the institution in order to offer the best possible service to their user community.

In order to continue adapting and adding value to the existing tool for generalisable large-scale applications, the UOC's translation system has recently implemented a semantic filter based on subject fields aimed at improving the translation quality and at better fitting the university needs. The paper will explain all the steps of this adaptive process, as well as a demonstration of the resulting tool: (a) the choice of the subject fields according to the university studies, (b) the design and implementation of the dictionaries used to extract the required information to filter and disambiguate homonym and polysemous terms, including source code in the dictionaries, and (c) the design and implementation of the corresponding web interface.

Lluís Villarejo – Mireia Farrús – Jordi Duran
Universitat Oberta de Catalunya, Barcelona, Spain
e-mail: lvillarejo@uoc.edu, mfarrusc@uoc.edu, jdurancal@uoc.edu

Sergio Ortiz – Gema Ramírez
Prompsit Language Engineering, Alacant, Spain
e-mail: sortiz@prompsit.com, gramirez@prompsit.com

1 Introduction

During the last decades, machine translation (MT) tools have proved highly useful in human translation work, since they save human translators a great deal of time by quickly providing initial draft translations (Craciunescu et al. 2004; Kirchhoff et al. 2004; Villarejo et al. 2009a).

The Language Technologies group from the Office of Learning Technologies at the Universitat Oberta de Catalunya (Open University of Catalonia, UOC) and the university's Language Service have recently been working on improvements to the UOC's Apertium-based translation (http://apertium.uoc.edu). Apertium is an open-source MT platform that can be easily modified and adapted to users' needs (Forcada et al. 2009).

At the time of writing, the UOC has created an Apertium-based web application incorporating specific terminology used by the university. It has also contributed to the development of the system multiple-format support (Word, PowerPoint, Excel, Writer, Calc, Impress, HTML files, compressed files, etc.) for document translation service, which includes the option of creating and exporting translation memories that allow for the customisation of translations (Villarejo et al. 2009a; Villarejo et al. 2009b; Villarejo et al. 2010). The advantage of working with Apertium is that all linguistic improvements made to the system are saved in a common repository so that its entire community of users can benefit from them.

The Apertium application was placed at the disposal of the UOC's internal community (the university's teaching staff and staff who manage and edit teaching material) in December 2009, and was well received by the users. Soon we realised that a user translating legal texts and another one translating computing texts required from different translations. With a view to obtaining high quality translations fully adapted to the university, the UOC subsequently set about to implement the one sense per discourse assumption to improve the quality of Apertium's translations.

The one sense per discourse assumption states that if a polysemous word is used more than once in a coherent discourse, it is very likely that all its occurrences refer to the same sense. This assumption has been proved to be effective to improve results in tasks like Word Sense Disambiguation (Gale et. al. 1992) and Statistical Machine Translation (Carpuat 2009). If we could detect that some translation pairs (source and target word) are specifically used in some of the university subject areas, we could build up a dictionary with them, ask the user to manually introduce the subject area for the source text which is to be translated and implement this way the one sense per discourse assumption. In this paper we describe how we incorporated subject areas related to the university fields of knowledge and the subjects taught into

Apertium in order to improve the translation quality. The inclusion of subject areas in the MT system, where they act as semantic filters, allows for the disambiguation of homonyms and polysemes. The translation process thus takes place on the basis of the topic of the source text.
This chapter describes the work carried out in relation to incorporating subject areas into Apertium. Sections 1 to 6 of the article deal with the planning of the project, and sections 7 to 9 with its execution and conclusions, plus future lines of work.

2 Relevance of the project

The new Apertium interface enables any user to translate documents on the basis of the subject area corresponding to each source text. The function in question can also be deactivated by selecting the 'General vocabulary' option instead of a specific subject area.

Additionally, the subject area analysis process makes it possible to identify new specific terminology and incorporate it into the MT system, as well as to produce glossaries of terms to complement universities' teaching material and academic and administrative documentation.

As Apertium is an open-source platform, all linguistic improvements made to it are stored in a freely accessible general repository. Any user can thus benefit from the semantic filters implemented in this project. Given that the subject areas have been established in relation to the UOC's fields of knowledge, the improvements made to Apertium in this project will be extremely useful in any sphere related to the university.

3 Goals

The incorporation of subject areas into the Apertium machine translation system within the context of the Universitat Oberta de Catalunya sought to accomplish some objectives. In the first place, a specific analysis and selection of the most useful subject areas for the UOC's community should be performed. As a consequence, the subject areas will serve as filters for the MT system with a view to improving its translated output. Secondly, the creation of a glossary of polysemes and homonyms which can be disambiguated by assigning subject areas to them will be fulfilled. And finally, the implementation of the subject filters in the MT system's interface so that users can choose a subject area and generate a much more accurate translation featuring the terminology used in the field in question will be carried out.

4 Beneficiaries

First of all, it is of great importance for the success of the project to identify those target people that will be potential users of the new machine translation. The main groups set to benefit from the project are the following ones:

(1) Administrative staff looking to circulate versions of documents or communiqués in various languages and who want to use MT to speed up their work.
(2) Linguists who correct and translate universities' teaching, academic and administrative documentation.
(3) External language professionals who work with university language services.
(4) Teaching staff seeking to use MT to prepare the teaching material with which they provide students.
(5) Students who want to use MT to obtain rough translations to enable them to understand material or carry out work in languages in which they lack proficiency.
(6) In general, institutions, organisations, businesses, the media and members of the public interested in using MT in their day-to-day activities.

5 Project stages and groups involved

Language technology projects require collaboration between two types of experts, namely linguists, with their mastery of language, and IT engineers specialising in natural language processing, who provide access to the technology necessary for the implementation of such projects. In that regard, the UOC boasts a multidisciplinary team that combines the linguistic expertise of the staff of its Language Service and the skills of the language technology specialists from its Office of Learning Technologies.

Broadly speaking, the project consisted of three stages. Each stage and the groups it involved are briefly outlined below.

Stage 1 included subject area analysis and description, as well as the compilation of university material from which specific vocabulary could be

extracted. The activities involved in this stage were carried out by the Language Service and the Office of Learning Technologies.

Stage 2 included producing a glossary of semantically ambiguous terms and adding tags to dictionary entries to identify their subject area. The activities involved in this stage were carried out by the Language Service together with the Office of Learning Technologies and Prompsit, a company that had already participated in technical work geared to MT in similar projects. In the case in hand, Prompsit designed the data files for subject areas format and the integration within Apertium and collaborated in the generation of linguistic data.

Stage 3 involved the implementation of subject areas in the new Apertium interface. The Language Service and the Office of Learning Technologies designed the interface, assessed it from a linguistic perspective and disseminated it, while Prompsit carried out the actual implementation work as well as subsequent technical testing.

6 General tool description

At present, the UOC has an internal translation service based on Apertium (available at http://apertium.uoc.edu). Anyone wishing to do so may download the same version of the system from a freely accessible general code repository (http://sourceforge.net/projects/apertium/).

The MT system can be used to translate texts entered directly in the platform's text box, documents in multiple formats, web pages (retaining the full original structure), HTML files and compressed files. It also allows for the creation and use of TMX-format (Translation Memory eXchange) translation memories, which help to generate much more accurate translations with a greater degree of customisation.

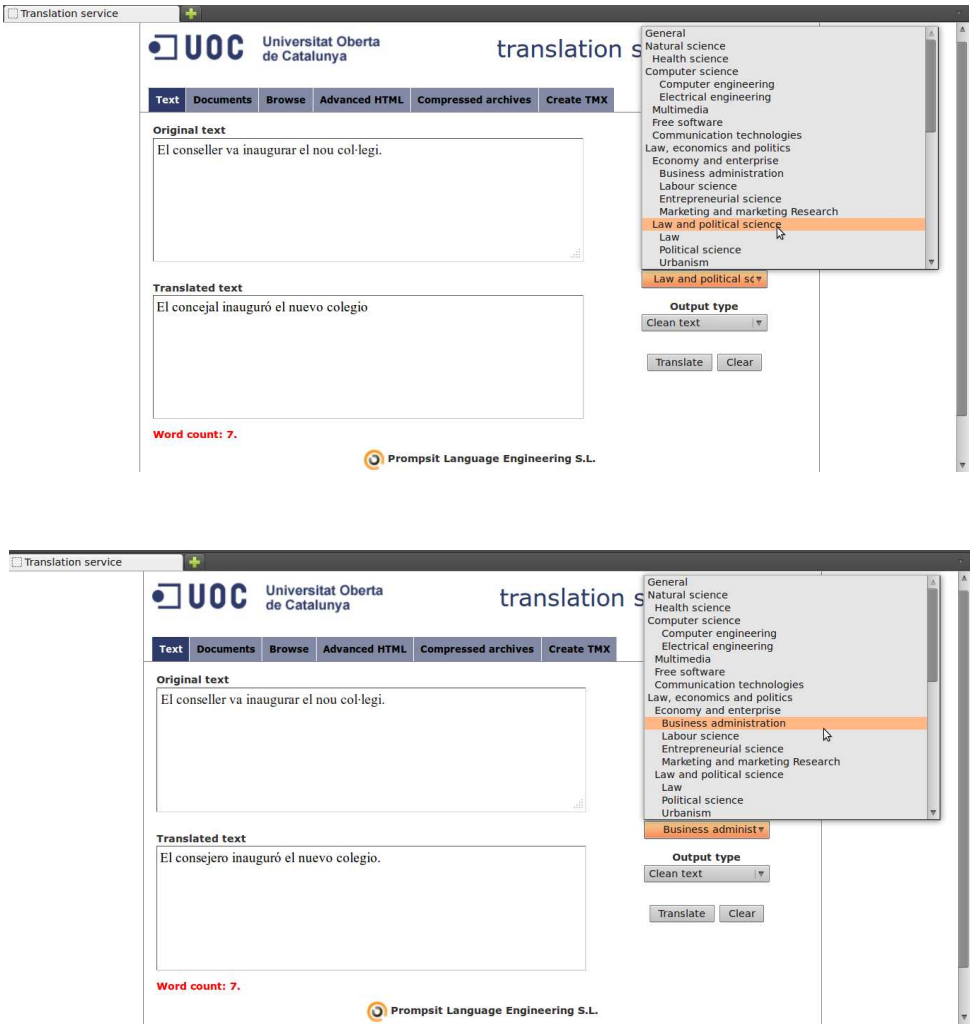

Figure 1. Proposed new interface. In the example, word "conseller" in Catalan is translated into "concejal" (councillor) in Spanish in the Law and Political Science domain but into "consejero" (advisor) in the Bussiness Administration domain.

The proposed new interface developed in this project features a drop-down list of subject areas from which users can choose in line with the topic of the translation to be performed (See Figure 1). The list of subject areas is based on a general basic classification of eight areas established by the Catalan university system's language services for previous language technology projects. The eight areas in question are:

(1) Pure sciences
(2) Natural sciences
(3) Computer sciences
(4) Law, economics and politics
(5) Social sciences
(6) Art and humanities
(7) Construction and architecture
(8) Industry

The above areas served as a starting point for establishing sub-areas on the basis of the UOC's different study programmes and requirements.

7 Goal fulfilment

The degree to which the project fulfilled the goals established in its application for funding is described below.

The purpose of the project was to improve the UOC's Apertium existing interface and adapt it to the university requirements. The main aspects of the project were as follows:

(1) Incorporation into the MT system of subject areas related to the UOC's fields of knowledge and the subjects taught at the university.
(2) Configuration of the subject areas to act as semantic filters for the disambiguation of homonyms and polysemes.
(3) Enhancement of translation quality through the application of the semantic filters.
(4) Open access for the system's community of users, who can apply and benefit from the improvements made.
(5) Ease of use, even for those unfamiliar with MT.

8 Stages and execution of the project

As stated in section 5, the work was carried out in three different stages, which are described below in detail.

Stage 1

a) Selecting subject areas

As mentioned previously, the classification established by the Catalan university system's language services (see Table 1) was used as a starting point for selecting subject areas to act as semantic filters. The classification was further developed to include the subject areas contained in the UOC's institutional repository and analysed in this project (see Table 2).

	Area	Sub-areas
1	Pure sciences	Chemistry, physics, mathematics, statistics
2	Natural sciences	Earth science, life sciences, health sciences, veterinary science, agriculture, stockbreeding, fishing, environmental science
3	Computer science	Computer science, telecommunications, electronics, electrical engineering
4	Law, economics and politics	Law, public administration, defence, economics, home economics, administration, work and employment, politics
5	Social sciences	Sociology, anthropology, psychology, pedagogy, communication, information, sports, games, tourism, hotel industry, food
6	Art and humanities	Geography, history, philology, philosophy, religion, art, music
7	Construction and architecture	Construction, architecture
8	Industry	Industrial engineering, various industries

Table 1. Subject areas established by the Catalan university system's language services (2001)

b) Compiling material

Parallel documents (i.e. documents in Catalan with translations into Spanish or English) corresponding to different subject areas (see Table 1) were used for the purpose of identifying ambiguous terms. The documents in question were taken from the volume of texts corrected and translated by the UOC's Language Service from July to December 2009.

0.	**General**
2.	**Natural sciences**
2.1.	Health sciences
3.	**Computer science**
3.1.1.	Computer science engineering
3.1.3.	Technical engineering in computer systems
3.2.	Multimedia
3.3.	Free software
3.4.	Communication technologies
4.	**Law, economics and politics**
4.1.	Economics and business
4.1.1.	Business administration and management
4.1.2.	Labour sciences
4.1.3.	Business studies
4.1.5.	Marketing and market research
4.2.1.	Law
4.2.2.	Political sciences and public management
4.2.3.	City management and urban planning
5.	**Social sciences**
5.1.	Information and knowledge society
5.2.2.	Teaching
5.2.3.	Psychology
5.2.4.	Educational psychology
5.2.	Psychology and educational sciences
5.3.	E-learning
5.4.1.	Communication
5.4.2.	Audiovisual communication
5.4.3.	Information and documentation
5.5.1.	Tourism
5.5.2.	Travel programme
5.6.	Food systems, society and culture
5.7.	Humanitarian cooperation, peace and sustainability

6.	**Art and humanities**
6.1.	Humanities
6.1.1.	Philosophy
6.2.	Languages and cultures
6.3.1.	Catalan philology
6.3.2.	Translation and technologies
6.3.4.	Arab and Islamic studies
6.3.5.	Cultural management
8.	**Industry**

Table 2. Subject areas analysed in this project

Stage 2

a) Designing dictionaries

Apertium has a monolingual dictionary for each of its languages and a bilingual dictionary for each of its language pairs. Each dictionary, be it monolingual or bilingual, is a separate file that can be modified and used on the basis of the language pair involved in a translation.

The same system of separate files was used for the purpose of assigning subject areas to ambiguous terms. More specifically, two different files were used, one for specifying subject areas and the other for indicating the grammatical and semantic information required for disambiguating terms:

- *Subject area file.* This file contains a list of subject areas, each with an identifying code. For instance, the e-learning subject area, which has been assigned the code 5.3, is represented as follows:

  ```
  <d n="5.3">e-learning</d>
  ```

- *Disambiguation file.* This file indicates the different subject areas that can be assigned to a particular term. The Spanish term *mesa*, for example, can be assigned to the general corpus area (code 0), the law, economics and politics area (code 4) or the humanities area (code 6.1). The corresponding Spanish-Catalan dictionary entries are as follows:

  ```
  <e d="0"><p><l>mesa<s n="n"/></l><r>taula<s n="n"/>.
  <e d="4"><p><l>mesa<s n="n"/></l><r>mesa<s n="n"/>...
  <e d="6.1"><p><l>mesa<s n="n"/></l><r>tabula<s n="n"/>...
  ```

  The following rules must be observed where the disambiguation file is concerned:

  (1) At least the grammatical category must be indicated to take advantage of the Apertium part-of-speech disambiguation module based on hidden Markov models (Rabiner 1989), which automatically selects the most probable part-of-speech for homographs.

(2) The specific subject area must always be indicated.
(3) The terms that appear in the file must be incorporated into the corresponding monolingual and bilingual dictionaries.
(4) If a single term has two translations within a given subject area, an exclamation mark must be used to denote the translation to be prioritised. The criterion used to prioritise a translation over another is explained in the following section.

b) Extracting ambiguous terms

The first step in analysing ambiguous terms on the basis of the parallel corpus consisted of extracting relevant terms for each subject area, including the general area. Term relevance within an area was determined on the basis of its keyness, i.e. its quality of being key in its context (Scott and Tribble 2006; Bondi & Scott 2010). If, proportionately speaking, a term appears with significantly greater frequency in a text corresponding to a specific area than in an extensive general corpus (a reference corpus), it is said to have keyness in the area in question. The higher a term's keyness value in an area, the more relevant it is within that area.

The extraction of ambiguous terms was carried out in six steps, applying the concept of keyness:

Step 1. Calculating keyness values for each area

The keyness values of the terms appearing in each specific area were calculated, giving rise to a list (let's name it List 1), which is an ordered list showing each term's relevance in the area in question.

Step 2. Looking up terms in different areas (I)

In each subject area, a search was performed for terms also appearing in at least one other area and which thus had a certain probability of being ambiguous. This gave rise to another list (let's name it List 2) consisting of pairs of words and its keyness value. (Examples of the eight words with highest keyness values in the e-learning area in Catalan are: "*ser* (to be) 1513"; "*haver* (to have) 1437"; "*aprenentatge* (learning) 1199"; "*uoc*: 918";

“*anar* (to go) 713”; “*xarxa* (net) 670”; *curs* (course) 659; *ensenyament* (teaching) 659”; etc..

Step 3. Filtering using the reference corpus

The thousand most relevant terms, according to the keyness value, in the general reference corpus were extracted. Given that terms that are highly relevant in the reference corpus are held to be of little relevance in specific areas, those extracted from the reference corpus were deleted from List 2. Generic verbs (e.g. to be, to have, to make, to do and to go) were also deleted. The result was List 3, consisting of pairs of words and its keyness value. (Examples of the nine words with highest keyness values in the e-learning area in Catalan are: *“aprenentatge* (learning) 1199”; *“uoc* 918”; “*xarxa* (net) 670”; “*curs* (course) 659”; “*ensenyament* (teaching) 659”; “*nou* (new) 594”; “*social* (social) 572”; “*obert* (open) 561”; “*estudiant* (student) 561”; etc.

Step 4. Looking up terms in different areas (II)

With the initial lists having been altered, a new search for terms appearing in two or more areas was performed, giving rise to List 4.

Step 5. Filtering terms

Terms from languages other than that under analysis were deleted from List 4. Terms other than verbs, nouns or adjectives were subsequently deleted. The result was List 5.

Step 6. Checking for ambiguity

Having obtained a filtered list of terms and the subject areas in which they appear, a search was performed for terms with more than one translation in other languages and in other subject areas. Firstly, to that end, the texts of the Catalan-Spanish and Catalan-English documents corresponding to each area were aligned by using the Hunaling alignment tool (version 1.0). Next, the aligned sentences containing the relevant terms were extracted, classified by area. That information was used to manually determine which terms have different translations in different subject areas. A final list of

relevant terms by subject area having different translations in different subject areas was obtained.

Stage 3

The list of terms with different translation by subject areas was incorporated into the translation engine's disambiguation files. Apertium's workflow was modified to cope with this information during translation and, finally, the UOC's Apertium user interface was modified to allow subject area selection. At the time of writing (April, 2012), the *apertium.uoc.edu* translation service is awaiting the corresponding update.

9 Conclusions and future lines of work

By way of conclusion, it must be emphasised that the project's original goals were fulfilled. The UOC's Office of Learning Technologies worked with the university's Language Service and Prompsit to ensure that the incorporation of subject fields into the Apertium MT system was carried out with the utmost linguistic and technological rigour.

Nevertheless, the UOC aims at constantly improve the needs of the academic community, and to this end, new lines of work are planned as future developments related to the Apertium MT system.

First of all, one the future objectives is to extend the range of potentially ambiguous UOC terminology incorporated into the new interface featuring subject areas. Second and related to the first objective, another goal is to extend the range of subject areas that serve as semantic filters as the UOC adds further study programmes to the courses it offers. Third, the UOC aims at systematising the ambiguous terminology extraction process as more parallel documents become available from the Language Service and to automatise polysemy disambiguation by exploring the most recent work inside Apertium related with automatic lexical selection (Tyers 2012)Finally, an evaluation of the usefulness of each term identified in this project is also in the scope ot the future work and will be done through the analysis of new documents corresponding to each subject area covered.

Acknowledgements

The work in question was made possible by funding that the article's authors obtained from the Office of the Vice President for Research and Innovation of the UOC, through the APLICA call for innovation projects geared to management made by the university's Open Innovation Office.

10 References

Bondi, M. & Scott, M. (2010). Keyness in Texts. Amsterdam: Benjamins.

Carpuat, M. (2009). “One sense per discourse” In Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 19–27, Boulder, Colorado, June 2009.

Craciunescu, O.; Gerding-Salas, C.; Stringer-O'Keeffe, S. (2005). “Machine Translation and Computer-Assisted Translation: a New Way of Translating?”. The Translation Journal, vol. 8 (3).

Forcada, M.L., Tyers, F.M., Ramírez-Sánchez, G. (2009). “The Free/Open-Source Machine Translation Platform Apertium: Five Years on”. Proceedings of the First International Workshop on Free/Open-Source Rule-based Machine Translation FreeRBMT, Alacant, Spain.

Gale, W, Church, K., Yarowski, D.(1992). “One sense per discourse”. In HLT ’91: Proceedings of the workshop on Speech and Natural Language, pp. 233–237, Morristown, USA. Association for Computational Linguistics.

Kirchhoff, K., Turner, A.M.; Axelrod, A.; Saavedra, F. (2011). “Application of Statistical Machine Translation to Public Health Information: a Feasibility Study”. Journal of the American Medical Informatics Association, vol. 18, pp. 473-478. (doi:10.1136/amiajnl-2011-000176).

Rabiner., L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286

Scott, M.; Tribble, C. (2006). Textual Patterns: keyword and corpus analysis in language education. Amsterdam: Benjamins.

F. M. Tyers, F. Sánchez-Martínez, M. L. Forcada (2012). *Flexible finite-state lexical selection for rule-based machine translation*. Research paper accepted at the EAMT 2012 (to be released).

Villarejo, L.; Cullen, D.; Corral, A. (2009a). “La integració de les tecnologies de la llengua en el flux de treball del Servei Lingüístic de la UOC”. Llengua i ús, revista tècnica de política lingüística 46.

Villarejo, L.; Ortiz, S.; Ginestí, M. (2009b). “Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya”. Proceedings of the First International Workshop on Free/Open-Source Rule-based Machine Translation.

Villarejo, L.; Farrús, M.; Ortiz, S.; Ramírez, G. (2010). “A web-based translation service at the UOC based on Apertium”. In Proc. of the International Multiconference on Computer Science and Information Technology, pp. 525-520. Wisla, Poland.