AUTOMATIC MULTILINGUAL GENERATION OF ON-LINE INFORMATION SITES[1]

ENRIQUE ALFONSECA, DIANA PÉREZ and PILAR RODRÍGUEZ
Computer Science Department,
Universidad Autónoma de Madrid,
28049 Madrid (SPAIN)
{Enrique.Alfonseca, Diana.Perez, Pilar.Rodriguez}@ii.uam.es

The Welkin toolkit [Alfonseca, 2003]. This system was created as an aid to university students from different fields, as it automatically gathers information from several electronic English texts (with no annotations) and from the Internet, and generates, in a fully automatic and unsupervised way, an adaptive on-line information site from the contents of the texts. In this work, we show that it can be easily ported to new languages, as it has been applied to generate hypermedia sites in Spanish, with most of its functionalities intact, in just one week.

1 Introduction

One of the main goals of Natural Language Processing is to allow a computer to interact with people in the same way that people interact each other [Zuckerman, 2001]. It is widely accepted that we maintain a model of our interlocutor and that, while communicating, we adapt our utterances to those models. Thus, for instance, we may change the language style, the presuppositions and the intonation according to the person to whom we are speaking even though we may be conveying the same message. In the case of Adaptive Educational Hypermedia, the user's interaction with the system can be conceived as a dialogue, where the user's utterances are the mouse clicks on the hyperlinks which express petitions for information or for updating the user profile (the system's model of the user). On the other hand, the generated hypertext pages can be considered the system's answers in this dialogue [Oberlander et al., 1998].

In order to guide this interaction, Natural Language Understanding (NLU) techniques can help the system acquire information about a particular domain, or about its users, and Natural Language Generation (NLG) techniques can help generate text that is adequate with respect to the user's needs, providing a higher degree of versatility than a collection of canned text fragments [Zuckerman, 2001]. Such a system needs a domain-dependent Knowledge Base in order to be capable of generating the textual contents of the hyperpages. Several systems require that this base be built by hand [Lester, 1997; Cawsey, 1999] while others attempt to construct it either partially or entirely from relational databases and textual content [Oberlander et al., 1998, Milosavljevic et al., 1998], having in mind that it may require a manual revision or completion.

The Welkin system [Alfonseca, 2003] is a system for automatically generating hypermedia sites from linear texts, and displaying different views of the site to the different users, according to an internal model of interests. We argue that, even though the system was, until recently, only able to process the English language, it can be very easily ported to new languages. For instance, most of the functionalities of the system were recently ported to Spanish in less than two weeks work. The resulting system is not multilingual in the sense that it is not able to process *simultaneously* texts written in different languages, but it is now able to process, *independently*, two different languages. In this paper we discuss the main problems found, and what should be done in order to port the remaining modules into the Spanish language.

The paper is structured as follows: Section 2 describes the original architecture of the systems and the main modules that it contains. Section 3 describes the work performed to port most of the modules into the Spanish language, and the work that is yet to be performed. Finally, Section 4 contains the conclusions and future work.

[1] This work has been partially sponsored by CICYT, project number TIC2001-0685-C02-01.

2 The original Welkin system

This work aims at desigining a system able to generate automatically multilingual web sites for educational purposes, based on the Welkin toolkit [Alfonseca, 2003]. This system was created as an aid to university students from different fields, as it automatically gathers information from several electronic English texts (with no annotations) and from the Internet, and generates, in a fully automatic and unsupervised way, an adaptive on-line information site from the contents of the texts. This tool has been applied for generating information sites about Darwin's *The Voyages of the Beagle* and Hegel's *Lectures on the History of Philosophy*, amongst others. The processing involves two separate steps: firstly, the source texts are processed off-line with some linguistic tools (a tokenizer, a sentence splitter, a part-of-speech tagger, a stemmer and a cascade of shallow parsers) and with some Term Identification [Cabré et al., 2001] and Classification [Alfonseca and Manandhar, 2002] procedures in order to acquire, from those texts, a list of terms that is expected to be of relevance to the *d*

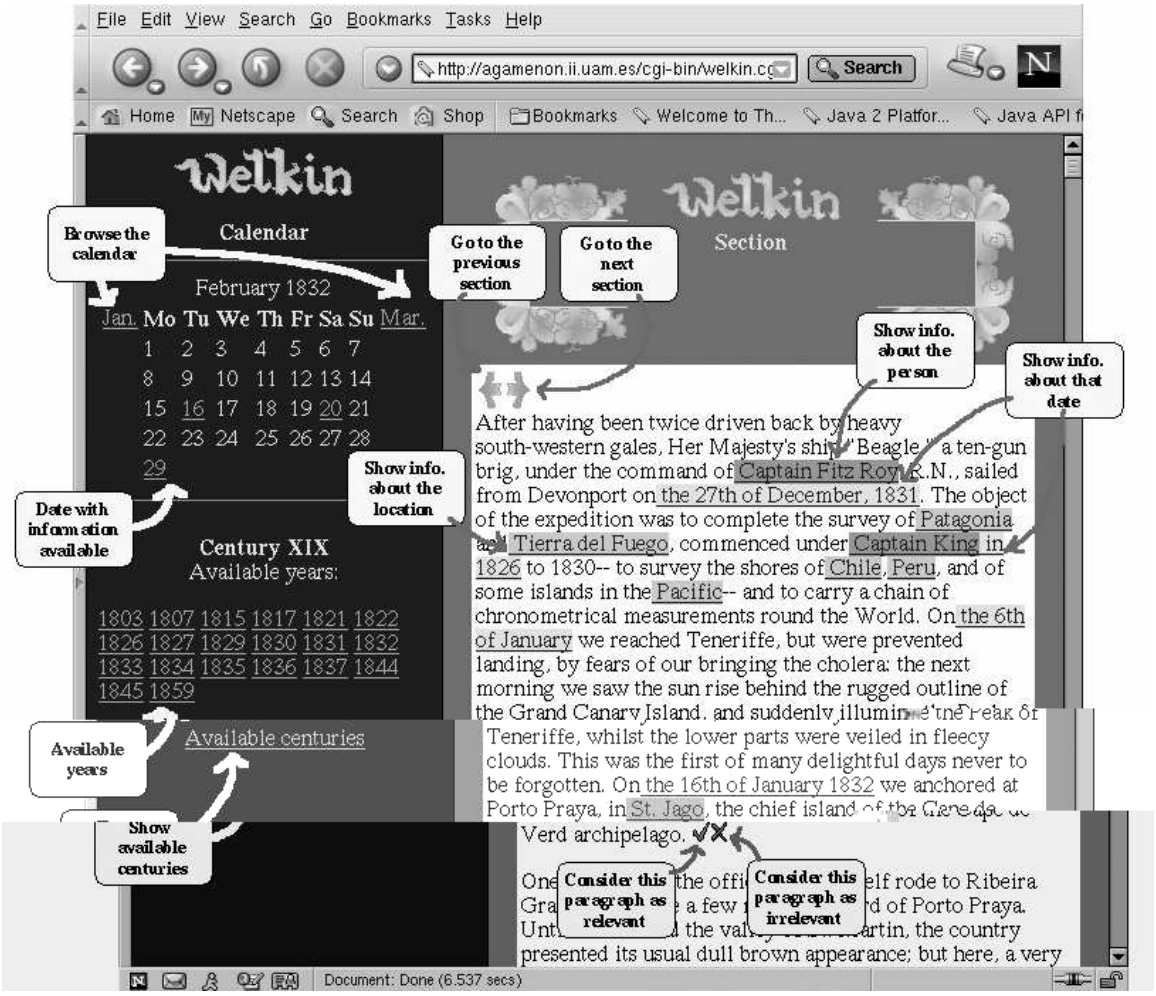

Figure 2. Snapshot of the Welkin interface.

We can cite the following advantages of our approach:

- It does not require any human supervision. It is only necessary to provide the system with the source texts and the web site will be automatically generated and ready to use.
- It provides additional functionalities to the students, such as the summarisation component based on sentence-extraction procedures [Mani, 2001] that can be used to summarise each of the hypermedia pages in the web site; or the topic filtering module, which makes use of the vector-space model IR techniques [Baeza-Yates and Ribeiro-Neto, 2001], and which can be used to select only information about some relevant topic (e.g. animals, plants, geography, biographies, etc.).

3 Porting the system to an additional language

Obviously, the tools that perform the linguistic analyses (the part-of-speech tagger, the morphological analyser, the chunkers and the parser) and those that handle the domain-specific terminology have to be adapted to the new settings. On the other hand, we argue that both the architecture and the remaining modules are general enough to process documents written in different languages:

Task	The original Welkin system	Welkin ported to Spanish
Tokeniser	Transform the text to the wraetlic format in XML. Paragraphs are marked with <p> tags and words with <w>.	The same, unchanged.
Sentence splitter	Sentences are wrapped by <s>.	Spanish list of abbreviations.
Part-of-speech tagger and stemmer	For p-o-s tagging, it uses the algorithm described in Brants [2000], coded in Java. For stemming, it uses a stemmer based on the LasIE stemmer [Gaizauskas et al., 1995]	We have chosen MACO+ [Atserias et al., 1998]. We have also written a wrapper for it, from and into our XML format, so it integrates perfectly with our linguistic tools.
Parsing	Three transformation list chunkers for identification of QPs, NPs and VPs, and several hand-crafted rules for subject-verb and verb-object relations.	We have used TACAT [Atserias et al., 1998] and we have adapted its output to Welkin so that it can be used in the next step without further modification.
Dates identification	A script written in *flex* keeps the regular expression to look for the dates in the text.	In one hour we have changed the script so that it recognizes the Spanish formats for dates and we haved translated the name of days and months to Spanish
Term identification	WordNet 1.7 has been used.	As WordNet contains only English words we have used the Spanish WordNet from EuroWordNet, ported to our format.
Term classification	Using hyponym patterns and word co-occurrences.	Future work.
Summarisation algorithm	Based on sentence extraction and using genetic algorithms.	The same, unchanged.
Vector-space model	Independent of the language.	The same, unchanged.

Table 1. Summary of changes when porting the system from English into Spanish.

- The linguistic processing stem was designed as a cascade of modules which added XML annotations to the original texts [Alfonseca, 2003]. Given that the format of the annotations is kept consistent, it is applicable regardless of the language chosen. We have kept our tools for tokenisation and sentence splitting (extended with additional support for UTF-8 characters) and we have written a wrapper for the MACO+ morphological analyser and the tacat parser [Atserias et al., 1998], so the output of these tools is converted into our own XML format.

- The term classification component is based on contextual information (co-occurrence of words and syntactic dependencies between heads of constituents). We believe that this kind of information can be collected with the same tools across languages to some extent (and, in particular, it should work amongst Indo-European languages, which share some common syntactic structures). In order for the module to work, it is necessary to collect from the Internet contextual information for each of the concepts in the original ontology. As we are using WordNet as the initial general-purpose ontology, and it contains several thousand nodes, this process takes several months and we have not been able to complete it for the Spanish language, so the classification module has not been ported yet.

- None of the well-known heuristics used for the summarisation algorithm are language-dependent, so it has been possible to apply it without any modification.

- Finally, the vector-space model for topic filtering and text classification have already proven to be applicable across languages [Baeza-Yates and Ribeiro-Neto, 1999].

4 Conclusions and future work

As hypothesised, it has been very easy to port Welkin to another language. We have proven that by porting it to Spanish language in less than a week and giving a practical example text from which a web site has been built automatically. In the table below we compare the original system and the modified one. We can see how the architecture and several modules have remained intact as they are language-independent. Furthermore, most of the modifications are very simple, some of them affecting only a few lines of code, or extending the DTD with special aliases for UTF8 characters.

As already mentioned, we are currently collecting the information from the Internet necessary to apply the Term Classification procedure. Although we believe that our classification algorithms can be ported into Spanish, it still remains to be proved empirically. In the future, we plan to be able to combine the modules for English and Spanish, so texts written in different languages can be combined into a single hypermedia multilingual site.

References

1. E. Alfonseca and S. Manandhar, *Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures*, EKAW-2002, Siguenza, Spain, 2002. Published in Lecture Notes in Artificial Intelligence 2473 (Springer Verlag).
2. E. Alfonseca, *An Approach for Automatic Generation of on-line Information Systems based on the Integration of Natural Language Processing and Adaptive Hypermedia Techniques*. Ph.D. Thesis, Computer Science Department, Universidad Autónoma de Madrid. 2003.
3. E. Alfonseca and P. Rodríguez, *Generating extracts with genetic algorithms*. In Proceedings of the ECIR-03 Conference. Also published in Lecture Notes in Computer Science 2633, pp. 511-519, Springer-Verlag, 2003.
4. J. Atserias, J. Carmona, I. Castellón, S. Cervell, M. Civit, Ll. Màrquez, M.A. Martí, Ll. Padró, R. Placer, H. Rodríguez, M. Taulé and J. Turmo, *Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text*. First International Conference on Language Resources and Evaluation (LREC'98). Granada, Spain, 1998.
5. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley-Longman. UK, 1999.
6. Thorsten Brants, 2000. *TnT - A Statistical Part-of-Speech Tagger*. In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.
7. M. T. Cabré and R. Estopá and J. Vivaldi, *Automatic term detection: a review of current systems*. In Recent advances in computational terminology, volume 2 of Natural Language Processing, pp. 53-87. John Benjamins Publishing Company, 2001.
8. J. Cawsey, 1999. *Patient information systems that tailor to the individual*. Journal of Patient Education and Counselling, 36: 171-180, 1999.
9. R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham and Y. Wilks, University of Sheffield: *Description of the LaSIE System as used for MUC*-6, Proceedings of the Sixth Message Understanding Conference (MUC-6), pp. 207-220. Morgan Kauffmann, 1995.
10. J. C. Lester, Develop*ing and empyrically evaluating robust explanation generators: the KNIGHT experiments*. Computational Linguistics 23 (1): 65-101, 1997.
11. Mani. *Automatic Summarisation*. John Benjamins Publishing Company, 2001.
12. M. Milosavljevic, R. Dale, S. J. Green, C. Paris and S. Williams, Virtual *Museums on the Information Superhighway: Prospects and Potholes*. Proceedings of CIDOC'98, the Annual Conference of the International Committee for Documentation of the International Council of Museums, Melbourne, Australia, 1998.
13. J. Oberlander and M. O'Donell and C. Mellish and A. Knott, *Conversation in the museum: experiments in dynamic hypermedia with the intelligent labeling explorer*. The new review of multimedia and hypermedia, 4, pp. 11-32, 1998.
14. Zukerman and D. Litman, *Natural language processing and user modeling: synergies and limitations*, User Modeling and User Adapted Interaction 11, number 1-2, pp. 129-158, 2001.