Antonio García Jiménez[1], Alberto Díaz Esteban[2], Pablo Gervás[3]
[1]Universidad Rey Juan Carlos (Madrid), [2]CES Felipe II (Aranjuez), [3]Universidad Complutense de Madrid

Knowledge Organization in a Multilingual System for the Personalization of Digital News Services: How to Integrate Knowledge

Abstract: In this paper we are concerned with the type of services that send periodic news selections to subscribers of a digital newspaper by means of electronic mail. The aims are to study the influence of categorisation in information retrieval and in digital newspapers, different models to solve problems of bilingualism in digital information services and to analyse the evaluation in information filtering and personalisation in information agents. Hermes* is a multilingual system for the personalisation of news services which allows integration and categorisation of information in two languages. In order to customise information for each user, Hermes provides the means for representing a user interests homogeneously across the operating languages of the system. A simple system is applied to train automatically a dynamic news item classifier for both languages, by taking the Yahoo set of categories as reference framework and using the web pages classified under them as training collection. Traditional evaluation methods have been applied and their shortcomings for the present endeavour have been noted.

1. Introduction

The recent boom in the popularity of the Internet has resulted in a rapid expansion of the range of information services available to the common user. One such service is that of systems offering to send users a selection of the daily news by e-mail. New ways of understanding information services and information systems are arising. In this paper we are concerned with the type of services that send periodic news selections to subscribers of a digital newspaper by means of electronic mail.

The task of managing the volume of information that the advent of Internet has thrust into our hands faces two significant challenges. The first challenge is posed by the ever present globalisation, which demands a capability for dealing with information in several languages in a homogeneous manner. The second challenge is a much older problem but made severe by the sheer volume of material currently in circulation: how to classify documents with a minimum of effort in order to provide subsets of the whole to which a user interested in a particular topic can address himself without having to shift through the complete set. Once a system attempts to face both challenges at the same time, the problem grows. The main question to be faced is how to improve on a rough an ready initial classification of documents under language heading (documents in English and documents in Spanish) to achieve a classification by topic independently of the language employed. This may present additional problems of granularity of the classification, due to the fact that fine grained classifications in different languages soon lose any semblance of similarity that coarse grained classifications may have had. At a certain level Spanish categories for news items will branch off into a bullfighting section, whereas the English equivalent may branch off to cricket or baseball. This is

* This work has been partially supported by Spanish Ministry of Science and Technology (Ref: 2000/020)

not entirely a linguistic problem and is probably more related with cultural issues, but the problem remains and must be addressed.

2. Resources for Multilingual Information Access

It is very important for multilingual search to take into account both the growth of information services and monolingual digital libraries and the need for tools with multilingual capacity for information retrieval and extraction (Abaitua, 2000). An effective global information transfer faces up to the challenge posed by the large number of national languages in use. Language differences may become a barrier to information circulation in the world, among persons and among organizations. The access to foreign-language information can be facilitated by multilingual glossaries, thesauri and classifications (they can provide multilingual pointers to the subject matter of documents), and translations (Lancaster, 1992).

The use of bilingual corpora is very interesting in the development of applications - as in terminology, automatic translation, and information multilingual search -, specially over the Internet. There are different kinds of multilingual corpora: corpora of texts in different languages to implement quantitative or statistical studies; comparable corpora, consisting of texts in a language and translations of similar documents in the same language; and parallel corpora, the same collection of texts in more than a language, - explicit correspondence relationships should be made between segments of each language, by means of grammatical categories.

Asghar and Revie (2000) provide an interesting discussion of the role of thesauri and classifications in Internet: the growth of information in the worldwide Web and the migration of information resources to the new context demand a better and consistent subject identification; thesauri and classifications collaborate on description of information resources, avoiding problems associated with quality of information retrieved in the Web; thesauri and classifications improve the rapid and easy access to the information in the Web.

Approaches to the construction of a new multilingual thesauri are: usual construction of a thesaurus, seeking equivalencies among terms collected (with different results among languages), without direct references to terms or structures of an existing thesaurus; translation of a monolingual thesaurus; conciliation and adaptation of existing thesauri in two or more languages. In truth, multilingual access to document collections is crucial. Besides, the co-operation improves the instruments connected with the information retrieval and the access to the information, in order to facilitate human and automatic indexation and to create links among related institutions (Lancaster, 1992; Clavel-Merrin, 1999).

According to Aitchison and Gilchrist (1990), after verifying the suitability of the project, terms and categories of the thesaurus are translated with their equivalents. Documents in the source language are analyzed to assign them to categories (classification) or assigning different terms to each document in order to represent and to facilitate its retrieval (thesaurus). The last step is the formulation of the query in another language. By means of an automatic system, the user can search for terms with the equivalent terms in the original language as query.

3. Multilingual Information Access in Hermes

Hermes is a system that applies existing techniques from the field of text classification, text categorization (Sebastiani, 1999) and information retrieval (Salton, 1989), besides user modelling (Amato & Straccia, 1999), to the selection of items, from different newspapers in different languages (Spanish and English), relevant for a user. Each user can create a profile in his language with his preferences and receive daily the news items that interest him from the different newspapers (Díaz et al., 2000).

A user accesses the information server and registers for the service. The user selects his language and different data about his preferences (email address, days of the week to receive news, maximum number of items per message) and interests. These interests are: the sections of the newspapers, an alternative system of classification (first level of categories from Yahoo), and terms chosen by the user as interesting.

The system manages two models per user, one per language, and applies each model to the news in the same language. The categories of Yahoo are language independent because there is a hierarchy in each language with the same first level categories. The terms are translated to from one language to the other.

The message received by the user contains: the name of the user, the date, and a list of news items ranked according to the user information interests and respecting the maximum number of items per message defined. Each news item is presented with the source, the author, the title, a short summary adapted to the user (Acero et al., 2001), the relevance, and a link to the news item in the digital newspaper. At the end of the message appear the interests of the user as features in his profile in order to allow the user to check the true relevance of the received news.

Finally, the system allows relevance feedback (Nakashima & Nakamura, 1997). The user can vote about the news in a positive, in a negative or in an indifferent way. This information is captured by the system in another interest for the user, the feedback terms that will be used in the next selection of news item.

4. Multilingual Text Classification in Hermes

Hermes uses three different systems for classifying information: one is the static classification of news items into sections provided by the newspaper domain, a second one is provided by a dynamic classification of the news items carried out automatically in terms of the categories used in the Yahoo directory, and a third one may be provided by the user as a custom-tailored category defined by a set of keywords and which is also automatically applied to the news items. The final classification is obtained by combining these sources through a weighted formula, according to a set of weights specified in the user model during configuration. These systems should ideally be as orthogonal as possible, in order to present truly different classifications of the domain. This is not the case altogether, but the overlap is not excessively significant.

4.1 The Choice of Categories

The categories of Yahoo were chosen as a reference framework in the first approximation for various reasons generally related with the overall efficiency of the process. On one hand, they come associated with distinct sets of classified documents in different languages (those classified under the English and the Spanish versions of Yahoo). These sets of documents were easily accessible in electronic form and could be used to train the automatic classifier to be employed. On the other hand, they are a set of categories specifically designed to facilitate search through a heterogeneous collection of documents, such as is found in the web. It was hoped that the differences between the set of news items in your run-of-the-mill daily edition and the collection of documents available in the web would ensure that this second set of categories add information to the existing one in terms of newspaper sections.

Various problems come associated with this choice. The automatic classifier is trained with documents corresponding to a domain other than the domain of application. The branching structure of the hypertext documents classified under each category implies that it is not always clear what page is an actual good example (possibly only leaves of the resulting hypertext trees should be used, cropping those intermediate pages which simply substructure a given category into subcategories but hold no relevant content themselves), and this introduces a degree of noise in the classification system. The effect of these problems in the evaluated results has been noted, and they are currently being explored in search of an optimized solution.

4.2 Dealing with More than One Language

In Hermes each user builds a model defining his preferences over categories and keywords for a single language, and the system generates a model in the other language automatically. Information about newspaper sections is not generated in this way because it is language dependent. This is a clear instance of equivalence problems between languages, made even more acute by the fact that each newspaper may have its own set of sections, even if working in the same language. The technique employed for generating models in a different language is based on the translation of the keywords defined by the user. The use of Yahoo categories, together with the assumption that Yahoo categories across different languages match, simplifies the process. Once the models for the two languages have been built, the news items for each language are processed with respect to the corresponding version of the model. Each of the language specific classification processes is independent of the other.

The final classification is carried out by combining the three different sources of classification through the weighted formula. Where automatic classification is required, it is achieved by calculating the one-to-one similarity between news items and the representation of the categories using the cosine formula of the Vector Space Model (Salton, 1989).

The representation of each category is obtained by training with different documents associated to that category (Sebastiani, 1999). A possible solution to the problems outlined above concerning the disparity of domains resulting for this particular choice of set of categories would be to train the system with a manually classified set of real news items, but classified under the Yahoo system. This would represent an important volume of work and would lose the advantages of having a dynamically updated set of sample documents for the chosen categories, with matching representation in different languages. Alternative solutions would be to combine both types of documents in training, or to perform co-training (Blum & Mitchell, 1998) on the representation of the categories, using the daily set of correctly classified news items. Either solution would gather together the advantages of both approaches.

5. Evaluation of Multilingual Information Systems

Evaluation of these new instruments requires: a reflection about categorisation, a validation of traditional evaluation measures within the new field of Internet, the consideration of the knowledge acquired during evaluation of search engines, and a close study of the working principles and the required evaluation according to the particular properties and conditions of the service under consideration.

Although there are various procedures for the evaluation of information systems, the emergence of the particular combination of challenges, objectives and techniques involved in personalised news services gives rise to additional issues that need to be addressed during system evaluation. On one hand, these systems have to ensure that the tools they provide for the user to specify his interest in information items of a particular type are sound according to traditional information retrieval measurements. On the other hand, they face a competitive market where different methods of specifying user interest are continuously competing for the user's eye, so any particular technique being employed must prove its worth in terms of user satisfaction. The following aspects must be covered in a thorough evaluation:

a) categorisation, filtering, personalisation.
b) user response
c) the vision that users develop of the system
d) user profiles
e) values of recall and precision for all the users on several specific days

In order to achieve all these aims, explicit evaluations provided by the users are harvested for feedback on system response-time, ease of use, system efficiency, and conceptual and physical presentation. This information is compiled on the basis of a closed questionnaire with specific questions on the relevant main topics. The user is asked to evaluate aspects such as category overlap, category validity, relevance of a document for the assigned category, or quality of the overall category scheme.

Additionally, a manual analysis of news items and user models logged by the system for a set of chosen days is carried out in terms of classic information retrieval measurements, which provide quantitative values for system efficiency .

The experience of evaluating system performance and user satisfaction for different personalised news services (Díaz et al., 2000) has proven the importance of the nature of the information in this tasks, the relative merits of the three most popular methods of specifying information interests (sections, categories, and key words) with respect to this particular set of tasks, and the risks of careless application of recall and precision measures in systems such as these where different methods of specifying interests are combined (Díaz et al., 2001).

An initial evaluation of a prototype of our system has given good feelings about the performance. This evaluation has been developed using a working pattern adapted to a monolingual version of the system used in previous experiments. This pattern includes several aspects as interface evaluation, newspaper sections, categories, summaries, bilingual capacity and user estimated recall and precision.

In general, users found the system suitable. They are satisfied with the different aspects of the user model, they estimate that the translation of the keywords is sometimes less than adequate but they value in a positive way the possibility to receive news in different languages.

We have yet to perform a more complete evaluation with a larger number of users and the relations between the different features that appear in our system must be studied in greater detail. For instance, how the multilinguality and the user modeling affect the traditional way of evaluating information retrieval systems, i.e. recall and precision measures.

6. Conclusions

This system can be a powerful tool in a multilingual context. In a globalized environment information services may take a principal role in overcoming linguistic and knowledge barriers, and contributing to the interrelation and even integration of cultures, economies and societies In truth, this integration depends on the efficiency of the system. The construction of this crucial instrument for the Information Society requires an evaluation that takes into account the user, the impact of automatic categorisation and user modelling, as well as the problems derived from the use of more than one language. Nonetheless, this tool will work in an integrating manner, from a cultural and knowledge perspective, whenever the contents that it helps to retrieve are specifically structured for this purpose - for instance, by respecting the differences between the different cultures, and supporting the common ground.

References

Abaitua, J. (2000). Tratamiento de corpora bilingües. In *La ingeniería lingüística en la sociedad de la información*. Held at Fundación Duques de Soria, Soria, 17-21 July 2000. (Provisional version). [http://www.servinf.deusto.es/abaitua/ konzeptu/ta/soria00.htm]

Acero, I., Alcojor, M., Díaz A. and Gómez J.M. (2001), Generación automática de resúmenes personalizados. *Procesamiento del Lenguaje Natural,*27, 281-188.

Aitchison, J., Gilchrist, A. (1990). *Thesaurus construction. A practical manual*. 2º ed., London: Aslib.

Amato, G. and Straccia, U. (1999). User Profile Modeling and Applications to Digital Libraries. In S. Abiteboul and A.M. Vercoustre (eds.), *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science*, Springer-Verlag, vol. 1696, 184-197.

Asghar Shiri, A., Revie, C. (2000). Thesauri on the Web: current developments and trends. *Online Information Review*, 24(4), 273-279.

Blum, A., and Mitchell, T. (1998). Combining labelled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory,* 92-100.

Cacho, I., Latorre, A. (2001). Tesaurus multilingüe europeu sobre la sidas i la infecció pel VIH. In Cabré, M. T., Codina, Ll. I Estopá, R. (ed.), *Terminología i Documentació*. I Jornada de Terminología i Documentació, 24 May 2000. Barcelona: Institut Universitari de Lingüística Aplicada, U.P.F. p. 61-70.

Clavel-Merrin, G. (1999). La necesidad de cooperación en la creación y mantenimiento de archivos temáticos multilingües de autoridades. In *65th IFLA Council and General Conference*. Held at Bangkok, Thailand, 20-28 August, 1999. [http://www.ifla.org/IV/ifla65/papers/080-155s.htm]

Díaz, A., Gervás, P. and García A. (2000). Evaluating a User-Model Based Personalisation Architecture for Digital News Services. In *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries, Lectures Notes in Computer Science*, Springer Verlag, 259-268

Díaz, A., Gervás, P., García, A., Chacón, I. (2001). Sections, categories and keywords as interest specification tools for personalised news services. *Online Information Review*, 25(3), 149-159.

Lancaster, F. W. (1992). *Vocabulary Control for Information Retrieval*, 2ª ed. Arlington: Information Resources Press.

Nakashima, T., and Nakamura, R. (1997). Information filtering for the Newspaper. In *1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*. Held at Victoria, B.C., Canada, 20-22 August 1997.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Reading, Massachusets: Addison-Wesley

Sebastiani, F. (1999). A Tutorial on Automated Text Categorization. In *Proceedings of the First Argentinean Symposium on Artificial Intelligence,* 7-35.