Managing a web site in several languages

.

Author: Jacob Palme, Department of Computer and Systems Sciences, KTH Technical University and Stockholm University, Forum 100, Kista, Sweden.
E-mail: jpalme@dsv.su.se
Keywords: Content management, Web site management, Multi-lingual, Natural-Language.
Subject area: FOR EVERYONE.

Abstract

Some web sites provide their information in multiple languages, with mostly the same information in each language. This paper discusses the experience in developing such a web site (http://web4health.info/). Web4Health is a psychology and psychiatry information web site, containing about 800 answers to common questions in this area for non-specialists. Web4Health had in October 2004 more than 140 000 visitors and more than 450 000 page downloads. Its content is available in German, English and Swedish, and partly also in Greek and Italian. The content was developed by medical experts in Germany, Greek, Italy, the Netherlands and Sweden.

Users can access the content either through a taxonomy (hierarchical subject tree) or through a natural-language question-answering system.

Each page has a number of attributes, such as title, author, question, answer, source references, internal-id, templates for the natural-language question-answering system, etc.

The content-management system stores the pages in internal format, and can export them in formats for screen viewing, for printing, source references and entry in the data base used by the natural-language question-answering system. The text of the same page in different languages are coupled, and a developer or translator can easily see the same text in two languages side-by-side, and then write or edit the text in one of the languages based on the content in another language.

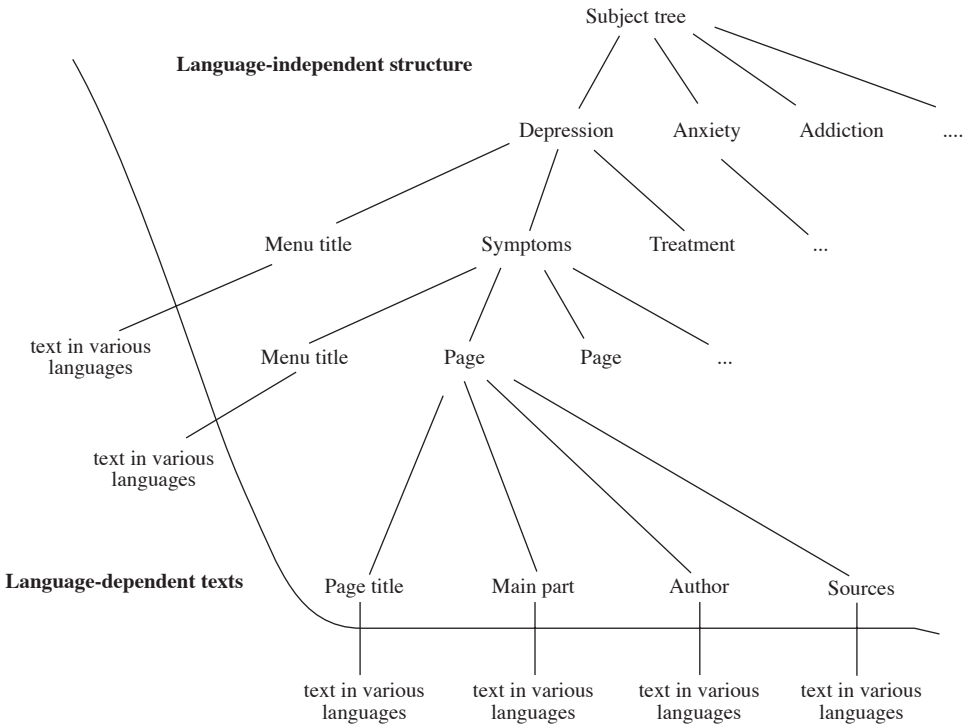

Figure 1: Separation of language-independent structure from language-dependent texts

Our experience is that it is very important to separate language-independent-structural information from texts which are different in each language. Such a separation makes is possible to change the structure in a single operation for all languages (see Figure 1).

The language-dependent text should be in the leaves of the structure. It should be easy to find which texts have not yet been translated to a particular language.

Cross-lingual question-answering

As described above, the natural-language question-answering method we have used means that we have to produce question-matching templates for each page. These templates also often need to be updated, based on entries in the usage logs, which show where the system did not provide the best answer to a certain question. The work of developing and managing these templates

require a special skill. Not even an ordinary professional translator can do it without some days of instruction on how to create such templates.

Because of this, it is an advantage if only some of the people need to have this particular skill. Also, it is very important that a change in these templates can be done in one language, and the result be immediately available for natural-language question-answering also in other languages.

We have implemented this, using a technique called cross-lingual natural-language question-answering [Peters 2000]. How this works is shown in Figure 2.

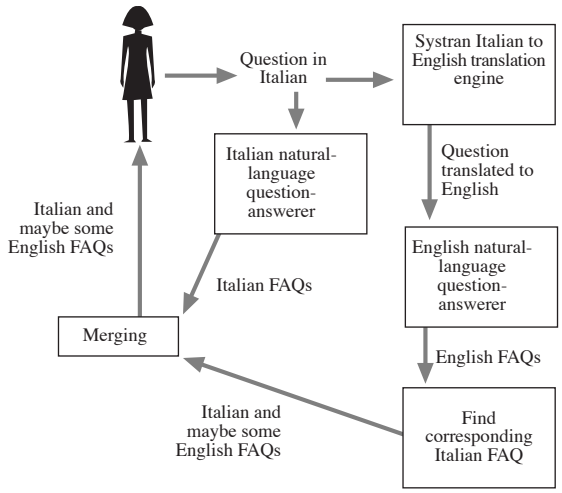

Figure 2: Cross-Lingual Natural-Language Question-Answering with Italian as an example.

Incoming questions are translated by Systran machine-translation to English. The English question is then put to the English-language answering engine. When the results have been found, the corresponding native language objects are shown. This could be implemented so that the user never sees that any other language than his own is involved. We have chosen, however, to show the English answer if the text of the answer has not yet been translated to the user's language. This means that users may sometimes see English answers mixed with the answers in their own language.

We also have some texts which are only available in the native language, since each national editor can add texts which are only available in his/her own language. For these texts, a native-language question-answering system is used to find answers.

We have compared the quality of the answers found in this way to question-answering directly in the language of the questions [Sjödin 2004]. These comparisons indicate that taking the Systran machine-translation engine as is, the quality will be somewhat inferior to that of direct language answering. However, if the dictionary used by the machine-translator is extended with the terminology suitable for our subject area, the quality will be almost as good as with direct language answering.

The reason for this is that the standard Systran dictionaries are designed for office documents, not health. For example, the word “body” is by Systran translated in the meaning “main part”, which is the most common use of this word in office documents, but which, of course, is usually not suitable when talking about health.

One might argue that augmenting the dictionary with new terminology (in our case about 6000 words) is as much work as writing the classification separately in each language. However, this is not true, because the same dictionary entry can be used in the classification of many answers. For example, the dictionary entry for “cause” can be used in many pages discussing causes of various disorders. Another important advantage with cross-lingual question-answering is that development of the dictionary does not need the special skill needed for doing the classification. Thus, cross-lingual question-answering allows a separation of tasks between people with different skills. A large advantage is also that corrections and improvements in the templates can be done only in English, and will automatically be available for all other languages.

Acknowledgements

The work reported here was partially funded by the Commission of the European Communities in the projects Senior Online, SALUT and KOM2002.

The software described was developed by Lars Enderin and Torgny Tholerus.

References

Palme 2004 — Web4Health Complete Final Project Report, by Jacob Palme, July 2004, http://web4health.info/documentation/D-7-4-full-final-rep.pdf

Palme&Sneiders 2003 — Natural Language Question Answering System Classification Manual by Jacob Palme and Eriks Sneiders, http://web4health.info/documentation/D2-2b-classification.pdf

Peters 2000 — CLEF - Cross-Language Evaluation Forum, by Carol Peters, http://www.ercim.org/publication/Ercim_News/enw40/peters.html

Sjödin 2004: — Sjödin, Elin: Preliminary Results (extracts from a forthcoming Master's thesis, Stockholm University).

Sneiders 1999 — Automated FAQ Answering: Continued Experience with Shallow Language Understanding. Question Answering Systems by Erik Sneiders. Papers from the 1999 AAAI Fall Symposium. Technical Report FS-99-02, November 5-7, North Falmouth, Massachusetts, USA, AAAI Press, pp.97-107 at http://www.dsv.su.se/~eriks/Sneiders1999.pdf

Author Biography

Jacob Palme is professor of computer science at Stockholm University. He has authored 4 textbooks in the computer science area, made 29 presentations at international peer-reviewed scientific conferences, published 38 papers in international peer-reviewed scientific journals and 2 papers in anthology books, 4 standards documents, and has been invited speaker at 19 international scientific conferences.