PILLS: Multilingual generation of medical information documents with overlapping content

Nadjet Bouayad-Agha, Richard Power, Donia Scott and Anja Belz

ITRI
University of Brighton
Lewes Road
Brighton BN2 4GJ, UK
{Firstname.Lastname}@itri.brighton.ac.uk

Abstract

In the pharmaceutical industry, products have to be described by a range of document types with overlapping content. Moreover, much of this documentation has to be produced in many languages. This situation is commonplace in many commercial domains, and leads to well-known problems in maintaining a set of related documents and their translations. We describe a potential solution explored in the PILLS project. All relevant knowledge about a product is entered only once, through a natural-language interface to a knowledge base. From this 'master model', specialised models for a range of document types are derived automatically; from each specialised model, documents are generated automatically in all supported languages. As an illustration of this approach, the PILLS demonstrator generates three medical document types in English, German and French.

1. Introduction

The research reported here was carried out within PILLS (Patient Information Language Localisation System), a one-year project in the European Commission's eContent programme (Scott et al., 2001). The objective was to develop a prototype tool that supports the creation of different types of pharmaceutical documentation in multiple languages. The project partners were Berlitz GlobalNET, who performed market investigation and analysis, the Information Technology Research Institute (ITRI), who developed the editor and document generator, and the Medical Informatics group at Freiburg University, who created an ontology and other domain-specific resources.

For pharmaceutical companies, the publishing of information (package inserts, documents required for authorisation purposes, web content etc.) presents a major challenge, in particular for companies that market their products worldwide. Similar but not identical information must be provided to consumer, physician, pharmacist, and regulatory body, in different languages, adding up to a large number of documents per product. These documents must conform to strict regulations, which differ from country to country and which are constantly undergoing revision. At present, localisation of content requires significant investment in translation, as well as solutions for the complex task of maintaining multiple language versions of product information and other relevant content. Similar problems obviously arise in other domains, such as the production of technical manuals and other documentation for software or mechanical equipment.

PILLS explores the possibility that a computational tool, based on language generation, might allow a company to produce technical documentation more cheaply and more quickly; the production of medical information documents by the pharamceutical industry serves as a test domain.

2. Architecture

PILLS produces medical documentation in multiple languages by using Natural Language Generation (NLG); Machine Translation plays no role at all. A standard architecture for multilingual NLG is shown in Figure 1: the content of the document is first encoded in some kind of language-neutral predicate logic, which we have named the *content model*; this is passed to a generator, with access to linguistic resources (lexicons, grammars) for the supported natural languages; the generator produces output documents which express the desired content in each of these languages.

The main problem in applying NLG commercially has been to find a convenient way of creating and maintaining the content model. Graphical tools have been proposed as a way of editing the knowledge, but a diagrammatic presentation of any reasonably complex model quickly becomes such a tangle of arcs and nodes that it perplexes even an experienced knowledge engineer. PILLS uses an alternative method, developed at the ITRI, known as 'WYSIWYM editing': instead of presenting the knowledge by means of a diagram, the system generates a *feedback text* through which editing operations can be performed by opening pop-up menus on mouse-sensitive phrases (Power and Scott, 1998; Power et al., 1998). The architecture for a basic WYSIWYM system is shown in Figure 2; compared with Figure 1 the difference is that the generator now performs a dual function — as well as the output documents, it generates the feedback text for use during knowledge editing. Like the output documents, the feedback text can be expressed in any of the supported languages, so that people speaking different languages are provided with a window on exactly the same formal encoding of the content.

An innovation in PILLS, compared with earlier WYSIWYM systems, is that the program can generate documents of several different types, the overlapping content being defined only once in a *master model*. This addition to the architecture is shown in Figure 3. As before, the generator

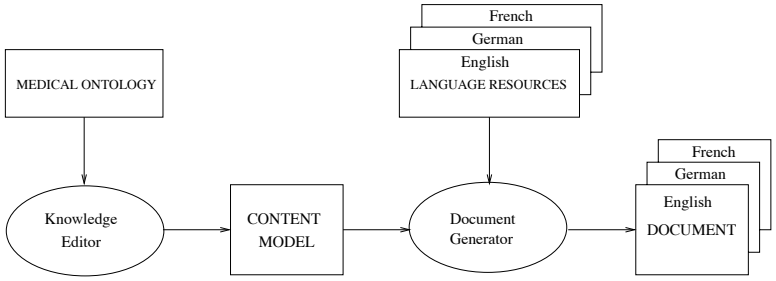

Figure 1: Natural Language Generation

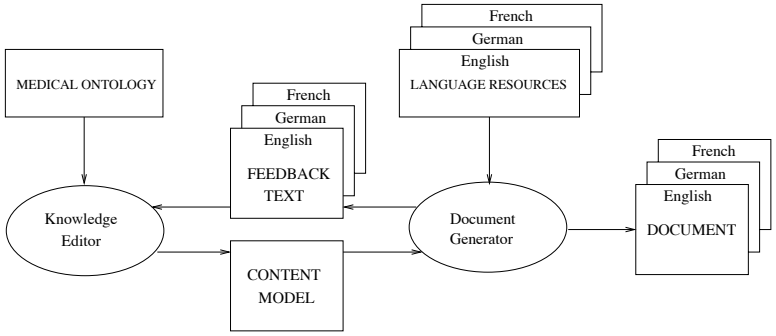

Figure 2: WYSIWYM editing

can produce feedback texts as well as output documents, but the feedback texts allow only the editing of the master model, which encodes all the information needed in order to produce several types of *specialised document* — the Patient Information Leaflet (PIL, see (ABPI Compendium, 1997), the European Pharmacopia (EP), and the Summary of Product Characteristics (SPC).

A second innovation is that the ontology employed during knowledge editing has been derived, in part, through an automatic extraction from a large medical database, the Unified Medical Language System or UMLS (Schultz and Hahn, 2001). As a result, the program begins to address the problems of scale that would arise in a commercial application, with thousands of lexical entries for some common medical categories like diseases and ingredients.

When using the system to produce documentation for a product, the author creates a single model encoding all the desired content (the 'master model'). This is done by editing a feedback text called the 'master document'; for an example, see appendix A. From the master model, the system can automatically derive models for three specialised documents (PIL, EP, SPC); examples are shown in appendices B-D. Both the master document and the specialised documents can be generated in three languages — English, German and French.

3. Model management

The idea of drawing information from a 'master document' has been employed before, in the HealthDoc project (DiMarco et al., 1995), as a means of producing brochures customized to the needs of particular patients. The master document in HealthDoc seeks to describe 'all the information, including illustrations, that might possibly be included in any individual brochure', together with 'annotations as to the conditions under which each piece of information is relevant'. However, the information encoded in this master document is linguistic, and thus language-dependent: it cannot be used as a source for generating documents in languages other than English.

Although PILLS also uses the concept of a 'master document', its significance is very different. Instead of serving directly as a resource for generating texts, it is merely a presentational device used by the knowledge editor as a means of creating a master *model*, encoded in language-neutral predicate logic. This brings two advantages. First, we can generate the specialised documents in several languages, since we are starting from a precise language-neutral source. Secondly, we can define mappings between *models*, not texts, as a first step in producing the specialised documents. Thus the process of selecting and organising information for the specialised documents can be performed only once, using the logical representation; multilingual versions for each document type can then be generated in the usual way.

PILLS contains rules for performing three mappings: Master $\rightarrow$ PIL, Master $\rightarrow$ EP, and Master $\rightarrow$ SPC. Formally, the objects that are mapped are directed acyclic graphs in which each node represents an instance, node labels represent the type of the instance, and outgoing arcs represent properties of the instance. The mapping process has three stages:

1. Build fixed structure

Each specialised model has some fixed structure that corresponds to the overall design of the corresponding document. For example, any patient information

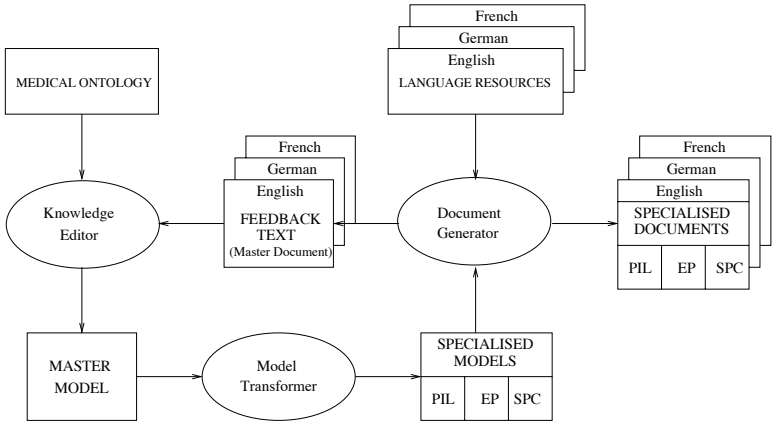

Figure 3: PILLS architecture

leaflet will have sections for ingredients, manufacturer, and purpose, so the corresponding instances and relations in the PIL model can be created without even consulting the master model. This stage will yield the skeleton of a specialised model, in which the external nodes remain to be developed.

2. Perform path mappings

Each external node in the incomplete specialised model can be mapped, by means of a path-to-path rule, to a subgraph in the master model where the relevant information will be found.

3. Perform subgraph mappings

In some cases, the constituent in the master model found through a path-to-path mapping can simply be copied to the corresponding location in the specialised model. In other cases, it will have to be mapped to a different object which organises a similar meaning in another way. This is done through another kind of mapping rule which maps subgraphs from the master model into subgraphs in the specialised model.

4. Conclusions

The PILLS system provides a demonstration, in miniature, of a potential documentation localisation tool for the pharmaceutical industry. It is based on the principle that *information should be entered only once*. The master model, presented to users through the master document, serves as a repository of all the information relating to a product; from this model, documents of various types are automatically generated in multiple natural languages.

Scaling up to a full-scale commercial application would require the following steps:

- Extend the use of medical ontologies so that all necessary concepts are covered. (Obviously maintenance would be needed, since new concepts are often introduced.)
- Extend the grammatical coverage so that more descriptive passages can be generated. (The current range of prose patterns is rather limited.)
- Extend the supported document types so that all documents using the information in the master model can be produced.
- Add more languages; the number of languages required for medical leaflets runs into hundreds.

Obviously these extensions are a huge undertaking, but some benefit would be obtained even through a system that only covered a few languages or a few document types.

An interesting expansion of the current architecture would be to allow for different hierarchical levels of master models in order to further reduce the amount of redundancy. There is often a single SPC for different forms or dosages of the medicine and as many PILs as there are forms and dosages. This could be described in different stages of increasing specificity, allowing for the generation of more specialised documents from a single source. As well as document type specificities, any particular national preferences and legislations should be taken into account in the interface.

Future plans also include the integration of stylistic variation (e.g., house-style of pharmaceutical companies) following the approach used in the ICONOCLAST project (Bouayad-Agha et al., 2000).

5. References

ABPI Compendium. 1997. *Compendium of Patient Information Leaflets*. Association of the British Pharmaceutical Industry.

N. Bouayad-Agha, D. Scott, and R. Power. 2000. Integrating content and style in documents: a case study of patient information leaflets. *Information Design Journal*, 9/2-3:161–176.

C. DiMarco, G. Hirst, L. Wanner, and J. Wilkinson. 1995. Healthdoc: Customizing patient information and health education by medical condition and personal characteristics. In *First International Workshop on Artificial Intelligence in Patient Education*, Glasgow, UK.

R. Power and D. Scott. 1998. Multilingual authoring using feedback texts. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th An-*

nual Meeting of the Association for Computational Linguistics, pages 1053–1059, Montreal, Canada.

R. Power, D. Scott, and R. Evans. 1998. What you see is what you meant: Direct knowledge editing with natural language feedback. In *Proceedings of the 13th Biennial European Conference on Artificial Intelligence*, pages 675–681, Brighton, UK.

S. Schultz and U. Hahn. 2001. Medical knowledge reengineering: Converting major portions of the UMLS into a terminological knowledge base. *International Journal of Medical Informatics*, 64/2-3:207–221.

D. Scott, N. Bouayad-Agha, R. Power, S.Schulz, R. Beck, D. Murphy, and R. Lockwood. 2001. Pills: A multilingual authoring system for patient information. In S. Bakken, editor, *Visions of the Future and Lessons from the Past. Proceedings of the 2001 AMIA Annual Symposium*, Washington DC.

A Master document

Legal information

- Tradename: Elixir.
- Legal Category: POM
- Proprietary: yes
- Product manufacturer: Super Pharma Limited, Super Estate, Brighton BN2 4GJ.

Ingredients

Active ingredients

- 500mg of Paracetamol (an analgesic).
- 100mg of Methionine (an amino-acid).

Indications

Paracetamol is used for the treatment of headache, toothache, backache and the symptoms of cold. Methionine is used for the treatment of overdose.

Inactive ingredients

- Magnesium stearate
- Polyethylene glycol
- Stearic acid

Absent ingredients

- Sugar

Presentation

Elixir is a tablet with the following appearance:

- Colour: white.
- Coat: 1m-coated.
- Shape: capsuloid.
- Mark: 'CM'.

Packages

1. Carton of 24 tablets, price £1.05
2. Carton of 96 tablets, price £2.77.

Shelf life

24 months.

Indications

Elixir is used for the treatment of headache, toothache and backache.
It is used for the treatment of the symptoms of cold.

B Patient Information Leaflet

What is in your tablets

The name of this medicine is Elixir. It contains 500mg of Paracetamol and 100mg of Methionine per tablet. Paracetamol is an analgesic. Methionine is an amino-acid used for the treatment of a potential overdose.

The other ingredients in Elixir tablets are magnesium stearate, polyethylene glycol and stearic acid.

Elixir does not contain sugar.

This pack contains 24 or 96 tablets.

Who makes your tablets

The manufacturer is:

Super Pharma Limited
Super Estate
Brighton BN2 4GJ

What your tablets are for

Elixir is used for the treatment of headache, toothache and backache. Elixir also relieves the symptoms of cold.

C European Pharmacopia

Paracetamol with methionine

Paracetamol is indicated for headache, toothache, backache and cold. Methionine is used for the treatment of potential overdose.

Elixir (Super Pharma)

Tablets. Methionine 100 mg, Paracetamol 500 mg.

Net price 24-tab pack = £1.05, 96-tab pack = £2.77.

D Summary of Product Characteristics

Qualitative and quantitative composition

Paracetamol 500 mg and Methionine 100 mg.

Pharmaceutical form

White, capsuloid, lm-coated tablets, marked 'CM'.

Clinical particulars

Therapeutic indications
Elixir tablets are indicated for headache, toothache, backache and the symptoms of cold. Methionine is used in the treatment of potential overdose.

List of excipients
Magnesium stearate, polyethylene glycol and stearic acid.

Shelf life
24 months.

Nature and contents of container
The product is presented in cartons of 24 and 96 tablets.