Annotating logical structure in a corpus of information leaflets

Nadjet Bouayad-Agha
Information Technology Research Institute
University of Brighton (U.K.)
nadjet@itri.brighton.ac.uk

Abstract

In this paper, we report on the task of annotating the logical structure in a corpus of information leaflets. We discuss some problems for identifying this structure in that particular genre. We argue that there is a need to understand better the communicative functions of layout in that genre.

Introduction

This paper discusses the issues raised by the annotation of the logical structure of information leaflets. The annotated corpus is to be used for the investigation of the role of layout in the text production process, from content selection and organisation to wording, the results of which is to be integrated in a generation system producing such documents. This corpus-based approach to natural language processing applications is facilitated by the advent in recent years of markup languages such as SGML and XML, which facilitate the delivery and presentation of documents.[1,2] These also provide a framework of reflection on the nature of document structure(s) which is undertook by projects such as the Text Encoding Initiative (TEI-P3 1997).

The logical structure is understood to be the visually observable elements of discourse structure such as sections and subsections, headings, paragraphs, footnotes and lists (Summers 1995). Thus, the attributes of layout are mapped onto meaningful communicative units. As we will see later, this mapping is not so straightforward in information leaflets. There are two main reasons for this. First, the lack of accepted conventions in this genre means that a greater range of layout attributes are used in these documents. This implies that we cannot often assert with confidence that a particular leaflet is badly designed and should therefore be ignored. Secondly, the logical class of some visual units can be ambiguous or difficult to ascertain when the properties of a particular visual unit do not fully correspond to the properties implicit to the logical class.

In the following section, the automated corpus annotation process is presented. Next, the issues just introduced are exemplified and discussed in more details. These issues require interpretation of content and therefore, are not taken into account by the automated annotation system.

The Corpus Annotation Process

The source of the corpus is the ABPI[3] (1997) Compendium containing all Patient Information Leaflets (PILS)[4] produced in the UK. It consists of 546 PILS (650,000 words). These display a rich variety of layout and linguistic styles.

The two excerpts below from the *ingredients* section of leaflets for patches illustrate the use of different layout and wording to express the same type of propositional content. The first excerpt expresses the information in one paragraph, each group of information being presented in one clause. The second excerpt expresses the information in a list, distributing it amongst the different sizes the medicine can be presented in.

> Evorel patches contain a natural oestrogen called oestradiol. The patches come in 4 different sizes: Evorel 25, Evorel 50, Evorel 75 and Evorel 100. The patches contain 1.6 mg, 3.2 mg, 4.8 mg, 6.4 mg of oestradiol and deliver 25, 50, 75 and 100mcg of oestradiol respectively per 24 hours.
>
> *(Evorel, Janssen-Cilag)*

[1]See the SGML/XML web page at http://www.oasis-open.org/cover/.

[2]In this paper, the term *layout* refers to all graphical aspects "which mark, organise or modify text" (Gilreath 1993), from the low level space features, typographical features (font-and-face alternation, capitalisation, etc.) and mark features (lines, bullets, etc) to the products of these features (lists, items, paragraphs, headings, sections, emphasised text, etc). Following the SGML terminology, in the rest of this paper, we call the latter *layout elements* and the former *layout attributes*.

[3]Association of British Pharmaceutical Industry.

[4]Patient Information Leaflets are the inserts that accompany medicines.

Estraderm TTS patches contain a substance called oestradiol. They come in three sizes:-

- Estraderm TTS 25 containing 2mg of oestradiol. Your body will absorb 25 micrograms of oestradiol each day whilst you are wearing Estraderm TTS 25 patch.
- Estraderm TTS 50 containing 4mg of oestradiol. Your body will absorb about 50 micrograms of oestradiol each day whilst you are wearing an Estraderm TTS 50 patch.
- Estraderm TTS 100 containing 8mg of oestradiol. Your body will absorb about 100 micrograms of oestradiol each day whilst you are wearing an Estraderm TTS 100 patch.

(Estraderm TTS, Ciba)

Figure 1 illustrates the process we went through for converting the PILS Compendium into an electronic corpus marked-up with layout. Each page of the compendium was scanned, OCRed, edited and saved in Microsoft Word 97. These files were then automatically converted to HTML. In order for this automatic conversion to be successful, the Word editor had to respect whenever possible some constraints, for instance, that lists were marked as Word lists, and that multicolumn texts were displayed in tables. The resulting HTML files can be viewed on any browser. However, their markup is only presentational: it does not reflect the logical structure of the document. This is partly due to the Word and HTML representations of a document and partly due to limitations of the conversion facility. For example, page and column breaks are presented with table rows and columns, lists and paragraphs are broken at page and column breaks, headings and emphasised text are not distinguished, list items cannot consist of more than one paragraph, headings marked with a bullet are tagged as (one-item) lists, the structural hierarchy between sections is lost, etc.

In order to remedy these problems, a program was written that derives the logical structure of these HTML documents. The outputs are SGML files that conform to a Document Type Definition (DTD, see (Goldfarb 1998)). The DTD was inspired mainly from the TEI Guidelines (1997). It can be found at http://www.itri.brighton.ac.uk/projects/iconoclast/.

Understandably, the program performs very well on documents with a distinguishable hierarchy and a simple grid and very badly on documents with a typographically non-distinguishable hierarchy and a complex grid. We manually evaluated ten randomly selected leaflets from the PILS and found that the program has a precision for nesting (determining the hierarchy) of about 50% and for sectioning (identifying the sections) of about 75%.

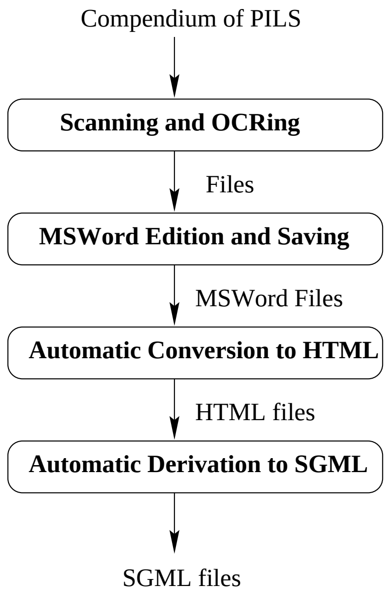

Figure 1: Conversion to Logical Structure Representation

Issues for the logical structure derivation

Layout attributes polysemy

There is a many-to-many correspondence between layout attributes and logical elements. For instance, italics can be used to emphasise or to signal foreign words; a blank line can be used to set off a warning, a list or a paragraph. Thus, in some cases, the text following a blank line marking the end of a list could either belong to the same paragraph or start a new one. This problem is prevalent in the PILS corpus which contains 2050 lists, 1210 of which are followed by a block of text.

The example below illustrates the problem. The first block of text following the first list presumably belongs to the same topical unit (paragraph) whereas the block of text following the second list, although visually balancing the second "paragraph" with the first one (both would consist of block + list + block), contains the referring expression *this list of possible events*, which semantically refers to both lists.

> As with all medicines undesirable events are sometimes experienced. With 'Sorbichew' these may include:
>
> headache
> flushing of the face
> dizziness
> weakness
>
> These may occur at the start of treatment but tend to become less as treatment continues.
>
> Other effects which may occur less frequently include:
>
> nausea and vomiting
> dizziness on standing up
> rash
>
> Do not be alarmed by this list of possible events. You may not have any of them.
>
> *(Sorbichew, Zeneca)*

A program such as the one described in the previous section which automatically derives the SGML logical structure obviously fails to consider these visual and linguistic issues. All it is capable to do is to identify blocks of texts. Thus, the excerpt below is made of six blocks, two of which are eventually recognised as lists. Additional heuristics could be added such as grouping together a list and the block of text preceding it which ends with a colon. This kind of superficial information is fairly limited because the lead text of a list can assume various forms and places (e.g., as a heading, as text following the list or preceding the list or both). Therefore, more linguistic information is called for.

Continuum of logical elements

There is a continuum of logical elements in the PILS corpus, from more layout-oriented to more content-oriented. This continuum of visual organisation was noted by Bernhardt (1985) across different genres.

In our corpus, problems arise for deciding whether a visual element belongs to a certain logical class because it does not have the prototypical features of that class. For instance, compare the cases below, which illustrate a continuum from heading-like to less heading-like segments. The last case does not display any of the characteristic features of a heading: it is not signalled typographically and it does not have the typical heading-like syntax as in the first example (noun phrase). The question arises as to whether it is a heading because it signals the theme of what follows, and, more importantly, it is the kind of information the reader will be looking for in PILS. In fact, the *headings* in the example do not reflect the structural hierarchy of the document, but rather, simply provide the reader with *reading access points* (Waller 1988).

PRODUCT LICENCE HOLDER

Elixir Limited, Manchester, UK.

Product licence holder: Elixir Limited, Manchester, UK.

Product licence held by: Elixir Limited, Manchester, UK.

Product licence held by: Elixir Limited, Manchester, UK.

The role of linguistic as well as graphical information in the identification of a logical type can be illustrated further with the case of lists. In some PILS leaflets, bullets are used to mark paragraphs throughout the document as if to provide *individual* access to these logical objects. In other leaflets, list items are presented as paragraphs and the only visually as well as linguistically unifying property is their syntactic parallelism. The first qualify as lists in our analysis whereas the second do not.

Genre-specific properties

Aside the paragraph-size list items mentioned above, the one-item lists and the one-sentence paragraphs, the PILS corpus display other properties which differ from standard expository texts. An important one is that the visual units do not necessarily correspond to the logical units (i.e., the argument structure). Thus, contrary to linguistic consensus (e.g., (Brown & Yule 1983)), the paragraph in some leaflets do not necessarily signal a new topic or subtopic; it might simply continue the previous one. One might therefore wonder the communicative function of the paragraph break in such cases.

In the example below, the topical continuity between the last two blocks of texts is emphasised by the use of a pronoun in the last block referring back to an entity in the previous block.

AFTER USING YOUR PATCHES

These patches sometimes cause unwanted effects in some people:

- Headaches, nausea or breast tenderness
- Cramping pains in the calf.
- Feeling slightly bloated.
- Slight redness and itching of the skin where a patch has been [..]

These effects are often mild and may wear off after a few weeks' treatment.

If they are very troublesome and do not improve tell your doctor. *(Estraderm TTS, CIBA)*

The communicative function of this paragraph break is made clear if one tries to reunite these two blocks. Indeed, a concessive marker is required at the beginning of the second block as illustrated below. This example illustrates clearly the rhetorical equivalence between laid-out and running prose.

These effects are often mild and may wear off after a few weeks' treatment. *However,* If they are very troublesome and do not improve tell your doctor.

Conclusion

We have tried with mild success to automatically annotate a corpus of information leaflets with logical structure. Indeed, for many leaflets, we have ended up with the visual rather than the logical structure of the document. The problem stems from the many different layout and linguistic features with which a logical element can be marked in this particular genre. It is also due to the many different communicative functions assumed by the same visual units, which means that an understanding of content is required in order to assign a logical type to a visual element. These issues are not new to the document design community (see for example similar observations made by (Norrish 1987)). However, if we are to automatically produce patient information leaflets, we need to understand the relative contributions of layout and language to the communicative goals of the author. The analysis of our annotated corpus is a first step towards this goal.

References

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

Bernhardt, S. 1985. Text structure and graphic design: the visible design. In Greaves, J. . W., ed., *Systemic Perspectives on Discourse*, volume 2. NJ: Ablex.

Brown, G., and Yule, G. 1983. *Discourse Analysis.* Cambridge Textbooks in Linguistics.

Gilreath, C. 1993. Graphic Cuieing of Text: The Typographic and Diagraphic Dimensions. *Visible Language* 3(27):336–361.

Goldfarb, C. 1998. *The SGML Handbook.* Oxford University Press.

Norrish, P. 1987. The graphic translatability of text. Technical report, British Library Board and University of Reading.

Summers, K. 1995. Near-wordless document structure classification. In *Proceedings of the International Conference in Document Analysis and Retrieval (ICDAR–95)*, 462–465.

TEI-P3. 1997. *Guidelines for Electronic Text Encoding and Interchange*, volume I and II. The Association for Computers and Humanities (ACH), The Association for Computational Linguistics (ACL), The Association for Literary and Linguistic Computing (ALLC). C.M. Sperberg-McQueen and Lou Burnard (eds). Available from the Text Encoding Initiative Home Page at http://www.uic.edu:80/orgs/tei/.

Waller, R. 1988. *The Typographic Contribution to Language: Towards a Model of Typographic Genres and Their Underlying Structures.* Ph.D. Dissertation, Department of Typography & Graphic Communication, University of Reading.