Document Content Analysis through Inverted Generation

Aurélien Max
Groupe d'Etudes pour la Traduction Automatique
(GETA CLIPS IMAG)
Xerox Research Centre Europe (XRCE)
Grenoble, France
aurelien.max@xrce.xerox.com

Marc Dymetman
Xerox Research Centre Europe (XRCE)
Grenoble, France
marc.dymetman@xrce.xerox.com

Abstract

A paradigm for the deep content analysis of documents in restricted domains is proposed, along with an implementation technique relying on the emergent field of interactive natural language generation.

A paradigm for document content analysis

The paradigm sees the formal specification of constrained content representations as a prerequisite for making sense of the documents and makes use of these representations for predicting textual aspects of the documents. Access to syntactic information is resorted to only when it is deemed necessary for disambiguating between two well-formed content representations, thus achieving a better division of labor between the highly constrained content space and the much looser syntactic/textual space. The approach relies on a mechanism for producing intermediate structures from the content representations which can be used to perform a fuzzy match with the text of the document to be analysed. The space of well-formed content representations is then heuristically searched based on the fuzzy similarity measure until good enough matches with the text are found. If several candidates remain at this stage, attempts are made to disambiguate between them using shallow syntactic and semantic clues (as may be provided by a shallow parser). If some decisions cannot be made reliably by the system, a human expert may be asked to disambiguate between candidates.

The paradigm reverses the traditional picture on content analysis, which tends to view it primarily as a parsing process, where gradually larger syntactic units (at the level of the sentence, then at the level of the discourse) are built and where semantic interpretation is typically done in a compositional manner on the basis of the syntactic structures found (see for example (Allen 1995)). In that picture, document content emerges so to speak as an object *derived* from well-formed syntactic constructs, and the central tool is a syntactically-oriented grammar. In our view, on the contrary, the central tool should be a formal specification of what counts as a valid semantic object, and a mechanism should be provided to generate textual predictions from such objects. We then view content analysis as a kind of *fuzzy inverted generation*, where a textual structure predicted (deterministically or non-deterministically) by the generator is compared with the text to be analysed. This is illustrated in figure 1.

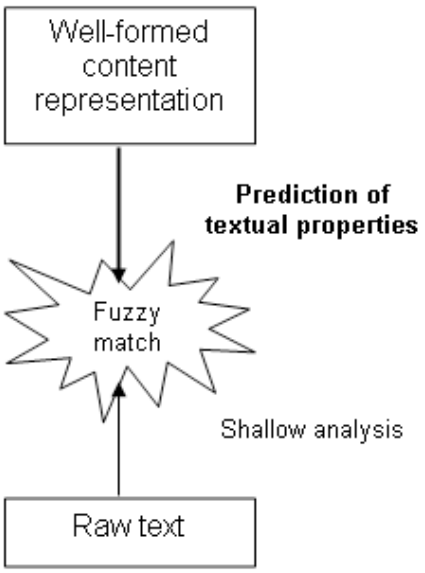

Figure 1: Deep content analysis through inverted generation

Implementation

A concrete implementation of the paradigm is currently under way at XRCE, taking as its starting point our Multilingual Document Authoring (MDA) system (Brun, Dymetman, & Lux 2000). MDA is a kind of interactive generation system (see (Power & Scott 1998; Coch & Chevreau 2001)) which uses an internal document-content representation, respecting certain predefined constraints, to guide an author during the formulation of a new document. The text of the document is generated automatically from the content representation, and serves as the interface to the user, through the use of *content selection*

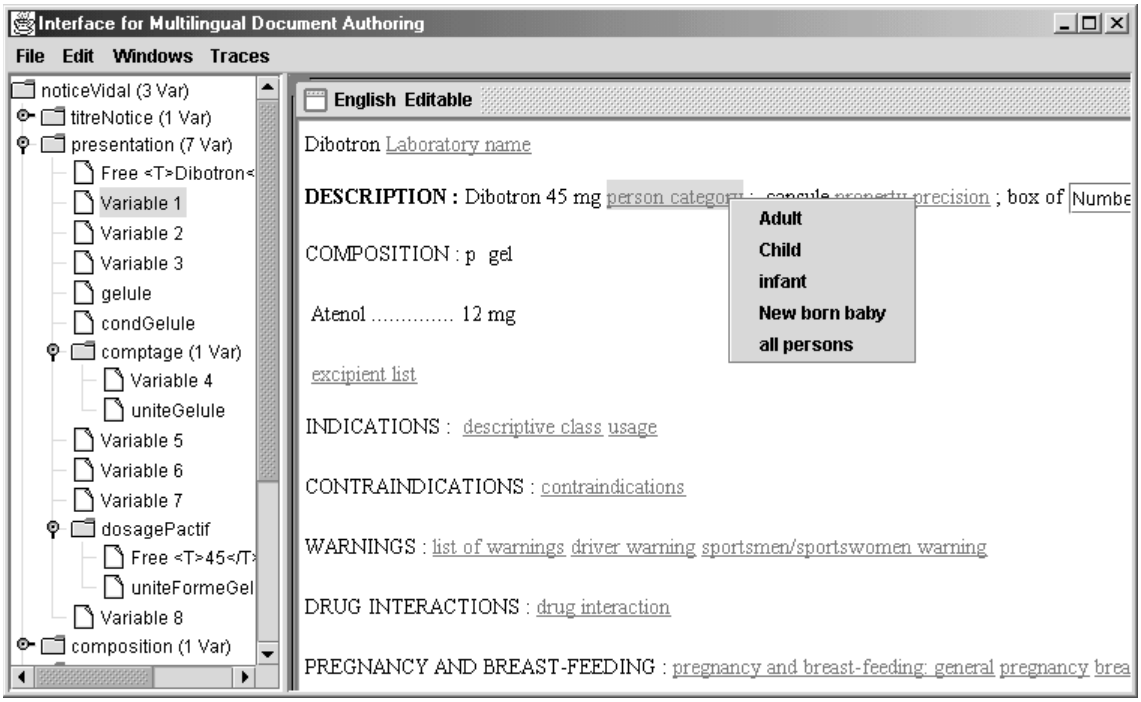

Figure 2: Drug description authoring using the MDA system

menus interspersed with the evolving text of the document (see figure 2).

In MDA, modelling the semantics of a class of documents is done by writing a kind of *content grammar* (logic-programming based) that recursively defines well-formed document representations. The internal coherence of these representations is guaranteed through constraint checking in the content grammar. In order to produce the actual texts from the content representations, a so-called *realization grammar* is used (or sometimes several realization grammars when one is interested in multilingual output).

An important methodological point in MDA (and one important for the present discussion) is that the content representations are not intended to focus on the detailed semantics of the words used in the documents (what we have called the *microsemantics*), but rather on the high-level communicative goals which are performed by the document (the *macrosemantics*). To illustrate this, let us look at two sentences that one could encounter in a drug leaflet:

1. *Do not take this drug in case of pregnancy.*
2. *Pregnant women are advised that this drug may be dangerous for their health.*

From the point of view of microsemantics, these sentences are completely different (lexicon, speech acts, ...); however, within the context of a drug leaflet, they carry the same communicative goal (have the same macrosemantics): pregnant women should not take the drug. A possible common representation for these sentences in MDA might then be dangerous(drug237,pregnantWoman).[1]

Although an MDA model is constructed with the principal aim of helping a user author a new document, it can be used to analyze an existing *raw* document of the same type. *In substance, we use the raw document as an information source for reconstructing the set of choices that an author would have made in order to describe interactively a new document*.

Let us sketch how this can be done. Conceptually, the content grammar can be seen as enumerating a collection of well-formed content representations and the realization grammar as taking such representations and constructing textual documents from them. At this level, an MDA system can then be seen as a kind of formal generative device for enumerating a collection of *virtual documents* and associating content representations with them. Now, it should not be expected in general that the raw document to be analysed coincide with any one of the virtual MDA-generated documents (the generative device is seriously undergenerating relative to the real corpus of documents in the class). However it is often possible to compare the raw document with a virtual document using a fuzzy similarity measure analoguous to the ones used in information retrieval. Conceptually we want to retrieve the virtual document which is *closest* to the given raw document and consider its associated content as an approximation to the real content of the raw document.

[1]For more details on these notions refer to (Dymetman, Lux, & Ranta 2000; Brun & Dymetman 2002).

In the implementation which is currently under way, the matching is done in two passes. In the first pass, a heuristic search is performed in the space of partially instanciated content representations with the goal of minimizing a certain rough distance between the lexical profile associated with the raw document and a lexical profile computed for the partially instanciated representation (the lexical profiles are expanded by adding synonyms and morphologically related words to allow better predictive behavior). If more than one candidate representation is kept at this stage, disambiguation is attempted in a second pass by using a more refined similarity measure which requires more computing effort (such as using shallow syntactic information or evaluating thesaurus distance between two words).

The rationale for this approach is that the space of possible content representations is vastly more restricted than the space of possible texts, and that even superficial similarities between textual elements predicted by the generator and elements in the input text will be indicative of commonality of content. So if the generator is prepared to realize a certain valid concept C1 as the text (1) it should be prepared to accept the same concept as an analysis of the *similar* text *Pregnant women should not take this drug*, unless another valid concept in its repertoire competes with C1 in terms of similarity with the input text.

This is the basic intuition, but it needs to be refined. First, not all words in the texts play an equally informative role for fuzzy matching: in a model for package leaflets, *pregnancy* will be a better predictor than *drug*. Second, some amount of non-determinism in the generator might help: being able to generate from concept C1 both (1) and the superficially somewhat distant (2) would allow to *catch* texts in the vicinity of both (1) and (2), with a better precision in the presence of possible competitors. Thirdly, generating along with a text some kind of intermediate structure (a shallow syntactic-semantic representation of the text) could improve the quality of matching, assuming availability of a shallow parser for getting a comparable structure out of the input text. Finally, in some cases, all hypotheses for filling some part of the content representation will lead to poor matches, and in such cases, the content representation should be left locally unspecified, in which case the system may resort to human disambiguation.

Challenge questions

How does my model adjust to a new domain?

The kind of model we have presented relies on a formal description of the valid semantic structures for a given class of documents. This requires that such descriptions be written in the formalism used after a thorough corpus study has been conducted. At XRCE, several types of models for documents in the medical domain, where structure and content are of high importance, have been written. Smaller experiments have been conducted in the domain of UNIX commands, where the model can predict and generate natural language descriptions of a subset of commands, and where document content analysis can serve to reconstruct the semantics of a command. More work needs to be done to describe more precisely the characteristics of the domains that are suitable to be modelled by our proposed approach.

How does my model adjust to previously unknown material?

The aim of our approach is to be as robust as possible to the linguistic variety in the input. This is done by allowing fuzzy matching between a text predicted by the model and the input text, this procedure being quite permissive relative to linguistic variation. In order to improve the quality of the matches, it can be useful to permit non-deterministic realizations of a given concept. These realizations are determined by looking at what linguistic formulations frequently occur in a corpus. For example, the model may generate the two following alternative realizations to express one and the same concept:

If pregnant or breast-feeding ask a health professional before use.

As with any drug, if you are pregnant or nursing a baby, seek the advice of your doctor before using this product.

These two sentences do not convey any relevant communicative differences for our purposes; generating them both makes the model cover more precisely the space of possible formulations for the concept.

How can knowledge in my model be inspected, assessed, and hand-edited except in a near-full-scale trial?

Let's note that once the content representation of a document is reconstructed, it can be used to produce a *normalized* reformulation of the input text simply by using the generator in the usual mode. This normalized document can be used by a human inspector to check that the content analysis system has correctly understood the input, without resorting to an examination of the underlying content representation. This provides a very useful debugging tool to check the correctness of the understanding process.

How well does the model perform in a large-scale trial?

We have not conducted any large-scale trial yet. An interesting question would be to evaluate empirically (for a given corpus) how many paraphrases for the same concept are necessary before the system is capable of automatic discrimination between different candidates.

What additional knowledge or theory does my model need to perform better?

Currently, the information used by the fuzzy parsing procedure is based on bags of words weighted by a certain

measure of informativity. We are also investigating possible refinements of this approach to account for some structural aspects of the predictions (e.g. linguistic dependencies, multiword expressions). But we need to evaluate the costs and benefits of adding such complexity to the procedure.

We have also designed the system so that it can support human disambiguation when it cannot decide between several candidates. At a later stage, we would like to reuse the knowledge brought in by the user at this stage to iteratively improve the performance of our model.

Perspectives and conclusions

We have sketched an approach to document content analysis which gives central focus to a formal specification of what may count as well-formed content representations in a given domain, and which attempts to reconstruct such representations by inverting the textual predictions made by a language generator. An implementation is under way, which attempts to use the input document as a source for reconstructing the choices made by an author interacting with the generation system. The approach is best adapted to situations in which rich a priori constraints on content representations can be formulated, but we believe that for such cases, focusing on the modelling of the content representations and on the texts that these representations can generate provides an interesting alternative to approaches putting more emphasis on syntax-driven parsing and bottom-up information extraction techniques.

Acknowledgements

Thanks to Christian Boitet, Caroline Brun, and Pierre Isabelle for several discussions. Aurélien Max' PhD work is supported by a CIFRE grant from ANRT.

References

Allen, J. 1995. *Natural Language Understanding*. Benjamin/Cummings Publishing, 2nd edition.

Brun, C., and Dymetman, M. 2002. Rédaction Multilingue Assistée dans le Modèle MDA. In *Multilinguisme et Traitement de l'Information*. Hermès.

Brun, C.; Dymetman, M.; and Lux, V. 2000. Document Structure and Multilingual Authoring. In *Proceedings of INLG 2000, Mitzpe Ramon, Israel*.

Coch, J., and Chevreau, K. 2001. Interactive Multilingual Generation. In *Proceedings of CICLING 2001, Mexico City, Mexico*.

Dymetman, M.; Lux, V.; and Ranta, A. 2000. XML and Multilingual Document Authoring: Convergent Trends. In *Proceedings of COLING 2000, Saarbrucken, Germany*.

Power, R., and Scott, D. 1998. Multilingual Authoring using Feedback Texts. In *Proceedings of COLING/ACL-98, Montreal, Canada*.

Reiter, E., and Dale, R. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.