Interpreting Communicative Goals in Constrained Domains using Generation and Interactive Negotiation

Aurélien Max
Groupe d'Etude pour la Traduction Automatique
GETA-CLIPS
Grenoble, France
aurelien.max@imag.fr

Abstract

This article presents an approach to interpret the content of documents in constrained domains at the level of communicative goals. The kind of knowledge used contains descriptions of well-formed document contents and texts that can be produced from them. The automatic analysis of text content is followed by an interactive negotiation phase involving an expert of the class of documents. Motivating reasons are given for an application of this approach, *document normalization*, and an implemented system is briefly introduced.[1]

1 Introduction

A classical view on text interpretation is to have a syntactic parsing process followed by semantic interpretation derived from syntactic structures (Allen, 1995). In practice, however, building broad-coverage syntactically-driven parsing grammars that are robust to the variation in the input is a very difficult task. Sometimes, it may not be relevant to perform a fine-grained analysis of the semantic content of text. Indeed, there are cases where what should be recognized is the high-level communicative intentions of the author. Depending on the kind of interpretation that is targeted from a text, some semantic distinctions need not be recognized. For example, the two following sentences found in a drug leaflet may not carry significantly different communicative goals in spite of their clear semantic differences:

- Consult *your doctor* in case of pregnancy before taking this product.
- Consult *a health professional* in case of pregnancy before taking this product.

We have identified a domain of application, *document normalization*, where text interpretation can be limited in many cases to the interpretation of a text in terms of the communicative goals it conveys (Max, 2003a). We have defined document normalization as the process that first derives the normalized communicative content of a text in a constrained domain (e.g. drug leaflets), and then generates the normalized version of the text in the language of the original document. We considered three levels in a *normalization model* for documents in constrained domain:

1. **Communicative goals**: the communicative goals that can appear in a document in constrained domain belong to a predefined *repertoire*.
2. **Communicative structure**: the communicative structure describes the content of a document in terms of compatible communicative goals, as well as how these communicative goals are organized in a document.

[1] The author was affiliated to Xerox Research Centre Europe (XRCE) and GETA when this work was carried out under a PhD grant.

3. **Natural language**: the language used should be as comprehensible as possible. To this end, every communicative goal should be associated with an expression that could be considered as ”gold standard”.

Figure 1 shows a warning section found in the drug leaflet for a pain reducer. Manually deriving a normalized version of this document extract using a normalization model requires identifying the communicative goals present in the document, which may be deduced from textual evidence found at different places in the document. Once identified, these communicative goals must be compared with the normalized ones in the predefined repertoire. We consider the four following cases:

1. A communicative goal in the document is clearly identified as belonging to the predefined repertoire.
2. A communicative goal in the document belongs to the predefined repertoire, but several normalized communicative goals are in competition due to some evidence found in the document.
3. A communicative goal in the document does not belong to the predefined repertoire, but it is deemed close to a normalized communicative goal.
4. A communicative goal in the document cannot be matched with any normalized communicative goal.

Once the normalized communicative goals have been identified, the communicative structure can be built (provided there are no incompatibilities) and the corresponding normalized textual version produced. A possible normalized text corresponding to the input document of figure 1 is given on figure 2.

The very general *Warnings* section has been split into several subsections. Communicative goals that were expressed in the same sentence have been isolated and reformulated in separate sentences, as is the case for the communicative goal indicating that the product should not be taken in case of allergy to aspirin. This communicative goal was found in a complex sentence, *Do not take this product if you have asthma, an allergy to aspirin, stomach problems...*, and was reformulated as *DO NOT TAKE THIS DRUG IF YOU ARE ALLERGIC TO ASPIRIN* in the section about product warnings.

The communicative goal warning about the risk of Reye’s syndrome in children is expressed in a long and complex sentence: *Children and teenagers should not use this medicine for chicken pox or flu symptoms before a doctor is consulted about Reye syndrome, a rare but serious illness reported to be associated with aspirin.* Considering the fact that no other communicative goals should be in competition with this one in this class of documents when Reye’s syndrome is involved, its identification can be quite simple.[2] In fact, it illustrates the fact that the interpretation of communicative goals within documents in constrained domains may not always require a very fine-grained semantic analysis, and that some indicators can already be quite informative.

However, it is unquestionable that in general identifying communicative goals and comparing them to predefined communicative goals clearly requires high-level interpretation capabilities, which would normally be those of an expert of the domain. With our application to normalize documents as target, we have proposed an approach to extract the communicative content of documents in constrained domains automatically. Considering that we wanted to obtain a practical normalization system, we further defined an approach to allow a human expert identifying the correct communicative content of a document from the set of hypotheses produced automatically.

This task should not be confused with text paraphrasing, for example for rewriting into a

[2] We do not claim that this is necessarily true in expert medical terms. Nonetheless, the normalization model that we used only considered this communicative goal involving Reye’s Syndrome.

Drug Interaction Precautions:
Do not take this product if you are taking a prescription drug for anticoagulation (thinning the blood), diabetes or gout unless directed by a doctor.

Warnings: Children and teenagers should not use this medicine for chicken pox or flu symptoms before a doctor is consulted about Reye syndrome, a rare but serious illness reported to be associated with aspirin. Do not take this product if you have asthma, an allergy to aspirin, stomach problems (such as heartburn, upset stomach, or stomach pain) that persist or recur, ulcers or bleeding problems, or if ringing in the ears or a loss of hearing occurs, unless directed by a doctor. Do not take this product for pain for more than 10 days unless directed by a doctor. If pain persists or gets worse, if new symptoms occur, or if redness or swelling is present , consult a doctor because these could be signs of a serious condition. As with any drug. If you are pregnant or nursing a baby, seek the advice of a health professional before using this product. It is especially important not to use aspirin during the last 3 months of pregnancy unless specifically directed to do so by a doctor because it may cause problems in the unborn child or complications during delivery. Keep this and all drugs out of the reach of children. In case of accidental overdose, seek professional assistance or contact a poison control center immediately.

Alcohol Warning: If you consume 3 or more alcoholic drinks every day, ask you doctor whether you should take aspirin or other pain relievers or fever reducers. Aspirin may cause stomach bleeding.

Figure 1: Example of a warning sections in a drug leaflet for a pain reducer

WARNINGS
Product warnings. DO NOT TAKE THIS DRUG IF YOU ARE ALLERGIC TO ASPIRIN. Do not take this product for more than 10 days unless directed by a health professional. Consult your doctor if pain persists or gets worse.
Alcohol. Do not take alcohol when you take this drug or ask your doctor for an alternative pain reducer.

Particular conditions. A doctor should be consulted before taking this drug if you have any of the following conditions:
- asthma
- stomach problems
- ulcers
- bleeding problems

Children and teenagers. CONSULT A DOCTOR BEFORE ADMINISTERING THIS PRODUCT TO A CHILD OR A TEENAGER, AS IT CAN INCREASE THE RISKS OF A SERIOUS ILLNESS CALLED REYE'S SYNDROME.

Pregnancy. Consult a doctor before taking this drug if you are pregnant. Using aspirin during the last 3 months of pregnancy may cause problems to the unborn child or complications during delivery.
Overdose. Stop taking this drug immediately and call a poison control control center or a health professional if you have taken too much of this drug.

Figure 2: Normalized text corresponding to the warning section of figure 1

controlled language (see e.g. (Nasr, 1996)). The main objective of our task is to identify which communicative goals from a given repertoire occur in a document, and to build a well-formed communicative structure that contains them.[3] Because the speech acts conveying a communicative goal (such as one that says that a doctor should be consulted before taking a given drug in case of pregnancy) can be performed under a wide range of surface forms, text paraphrasing would have to transform very different surface forms into the same target normalized text.

Through document normalization, we want to enforce 4 properties of document well-formedness that should be encodable into the normalization model used:

- Well-formedness of the communicative structure of documents: sentences should be well articulated to form a coherent discourse.
- Consistency of the communicative content: incompatible communicative goals should not coexist in the same document.
- Completeness of the communicative content: communicative content imposed by some communicative goal must be present.
- Comprehensibility and coherence of the language used: readers should be able to identify easily the communicative intentions across documents of the same class.

Text paraphrasing into a controlled language at the level of the sentence would only enforce the last property, because if controlled language rules can enforce some level of semantic well-formedness, they cannot guarantee the three other properties.

2 Automatic analysis of the communicative content of a document in constrained domain

Several approaches have already been experimented to analyze the content of documents in constrained domains, which can vary depending on the amount of surface analysis of the text. One type of approach uses information extraction techniques such as pattern matching that use strong predictions on the content and attempt to fill templates derived from a model of the domain (e.g. (Blanchon, 2002)), thus not giving too much importance to syntactic structure. Another type of approach first performs a syntactic analysis of the text, from which semantic dependencies can be extracted. The system presented in (Brun and Hagège, 2003) derives normalized predicates encoding the meaning of documents from semantic dependencies found by a robust parser. This allows obtaining identical semantic interpretations for paraphrases such as *ProductX is a colorless, non flammable liquid* and *ProductX is a liquid that has no colour and that does not burn easily*.

These approaches require an encoding of templates and extraction or normalization rules that may be difficult to build and to maintain. Furthermore, if they seem appropriate for extracting surface semantic information, interpreting communicative goals using these techniques may be more difficult. Indeed, communicative goals can be expressed with different surface texts carrying semantic differences that may not bear any significance for our purpose and may not always be considered as paraphrases. In the following examples from pain reducer leaflets, it may be acceptable that a particular normalization model consider the three following sentences as carrying one and only communicative goal:

1. This product should not be taken for more than 14 days without first consulting a health professional.
2. If pain persist after 14 days, consult your doctor before taking any more of this product.
3. If symptoms persist for 2 weeks, stop using this product and see a physician.

[3] It is true, however, that document normalization of a given document with very particular properties relative to a normalization model could be achieved by text paraphrasing at the level of the sentence, but this is too specific to us.

In order to be able to identify communicative goals, we believe that it is important to consider them within a well-formed communicative structure. Therefore, we think that the central objects for analysis should be well-formed descriptions of document communicative content[4], as it may be counterproductive to spend too much effort on the fine-grained analysis of surface text. If semantic dependencies can be expressed in these descriptions, then the space of possible contents will filter out incompatible communicative goals and thus disambiguate without always requiring a more fine-grained semantic analysis.

We have proposed an approach for the deep content analysis of documents in contrained domain, *fuzzy inverted generation* (Max and Dymetman, 2002). Well-formed document content representations are produced for the class of the input document. From these representations, normalized texts are generated, and a score of semantic similarity taking into account common descriptors is computed between the normalized texts and the text of the input document. The underlying hypotheses are, as we said earlier on, that considering well-formed content representations can restrict the space of the communicative goals to consider, and that the presence of informative textual indicators can help identifying communicative goals.

However, the space of content representations being potentially huge, a heuristic search can be performed to find the candidate representations with the best global scores. Moreover, in order to better cover the space of possible texts, the generation of the text can be done non-deterministically, so that several texts will compete over the input document from the same content representation. Figure 3 shows how several texts produced from a content representation can span several documents from the space of possible texts. The content representation that corresponds to the text with the highest similarity score with the input document is then considered to be the most likely candidate.

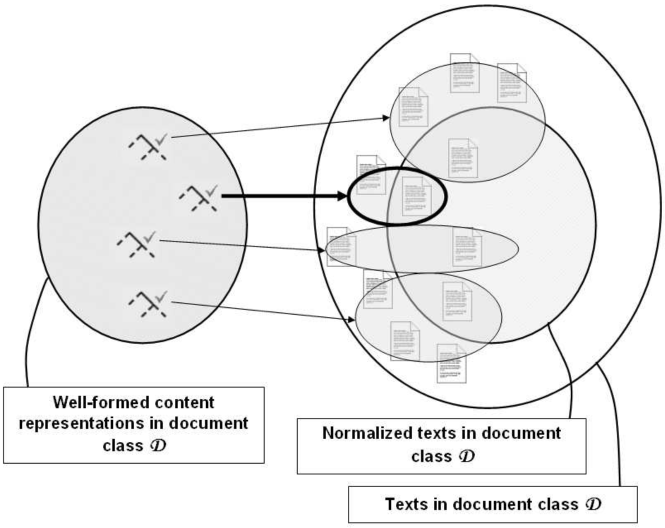

Well-formed content representations in document class $\mathcal{D}$
Normalized texts in document class $\mathcal{D}$
Texts in document class $\mathcal{D}$

Figure 3: Fuzzy inverted generation

3 Interactive validation of the correct communicative content

Relying solely on information retrieval techniques to associate a normalized content representation to an input document is unfortunately unlikely to yield good results, even if linguistically-oriented techniques can improve accuracy (Arampatzis et al., 2000). We have advocated an interactive approach to text understanding (Dymetman et al., 2003) where the input text is used as a source of information to assist the user in *re-authoring* its content. Following fuzzy inverted generation, an *interactive negotiation* can take place between the system and its hypotheses (the candidate content representations) on the one hand, and a human expert on the second. A naive way would be to let the expert choose which hypothesis is correct based on the normalized text associated with each one of them. But this would be a tedious and error-prone process. Rather, underspecifications from analysis can be found by building a compact representation of the candidates, and then used to engage in negotiations over local interpretation issues.

[4] This is under the assumption that the input documents are semantically well-formed and complete, but if they are not then the model used can indicate for what reasons they are ill-formed, and document normalization can be used to correct those documents so that they become valid relative to the normalization model.

Using interactive validation with generated texts has already been used in several domains: for example, (Blanchon, 1994) proposed disambiguation dialogues involving reformulations for dialogue-based machine translation; (Overmyer et al., 2001) proposed a text that can be used to inspect the domain object model automatically built from a text describing a software engineering domain model. In the following section, we introduce our implementation of a prototype system for interactive document normalization based on the two presented approaches.

4 Interactive document normalization system

Systems implementing controlled document authoring (Hartley and Paris, 1997) are based on an interaction with an author who makes semantic choices that define the content of a document, from which multilingual textual versions can be produced. Therefore, these systems integrate resources that can be used to represent document content and to generate textual versions of the documents. The MDA system developed at XRCE (Dymetman et al., 2000; Brun et al., 2000) uses a formalism inspired from Definite Clause Grammars (Pereira and Warren, 1980) that encodes both the abstract semantic syntax of well-formed documents and the concrete syntax for the documents in several languages.[5] MDA grammars contain the definition of *semantic objects* of a given *semantic type*, which are used to build typed abstract semantic trees. Importantly, the formalism can encode the three levels for a normalization model that we described in our introduction: semantic objects can be of any granularity and can thus be communicative goals; the communicative structure is described by the abstract semantic syntax, which can be used to express semantic dependencies across subtrees; and the text generated is entirely under control, so normalized texts can be associated with communicative goals.

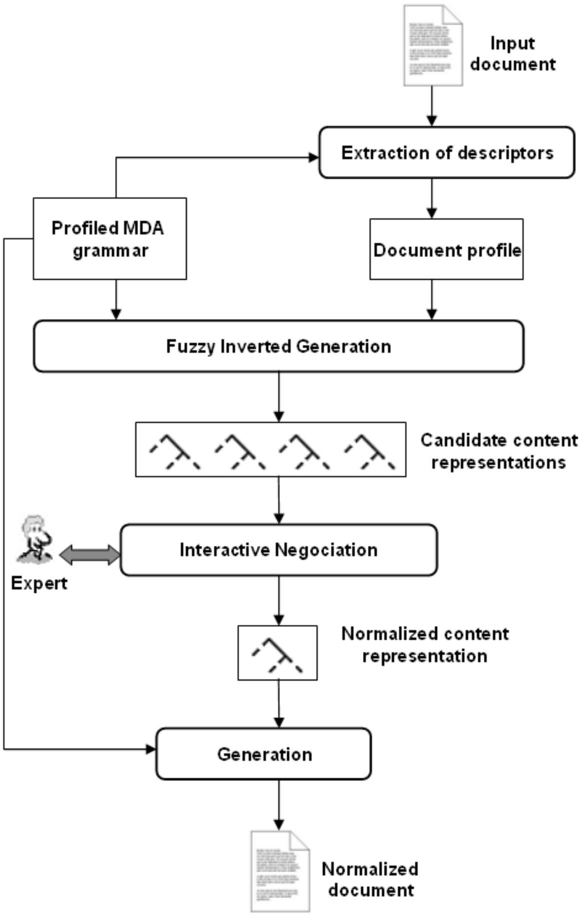
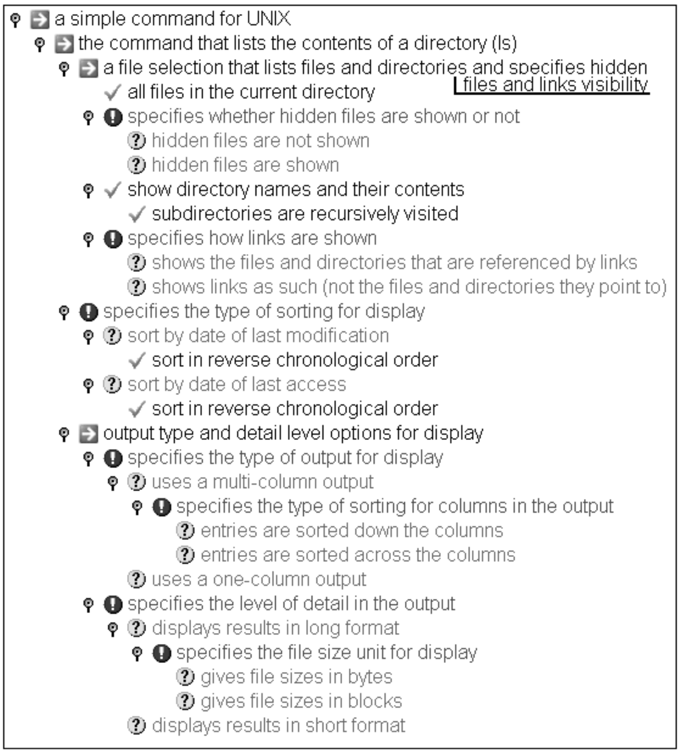

Input document
Extraction of descriptors
Profiled MDA grammar
Document profile
Fuzzy Inverted Generation
Candidate content representations
Expert
Interactive Negociation
Normalized content representation
Generation
Normalized document

Figure 4: Architecture of our document normalization system

For the reasons given above, we used the formalism of MDA for our implementation. The architecture of our normalization system is shown on figure 4. Textual descriptors (WordNet synsets in our current implementation) are first extracted from the text of the input document to build the *profile* of the input document. The MDA grammar used was previously compiled offline in order to associate profiles to each semantic objects and types described in the grammar. Fuzzy inverted generation is then performed from the profile of the document and the profiled grammar. Details on the implementation using MDA grammars have been described elsewhere (Max, 2003a; Max, 2003b).

The set of abstract semantic trees extracted by fuzzy inverted generation is then used to build

[5] This is achieved by developing parallel grammars that share the same abstract semantic syntax, but specify concrete syntax for a particular language.



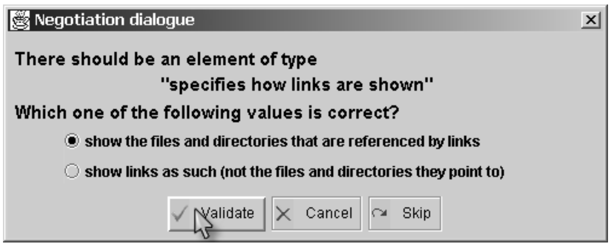
```
➔ a simple command for UNIX
  ➔ the command that lists the contents of a directory (ls)
    ➔ a file selection that lists files and directories and specifies hidden
                                                files and links visibility
        ✓ all files in the current directory
      ❗ specifies whether hidden files are shown or not
          ❓ hidden files are not shown
          ❓ hidden files are shown
      ✓ show directory names and their contents
          ✓ subdirectories are recursively visited
      ❗ specifies how links are shown
          ❓ shows the files and directories that are referenced by links
          ❓ shows links as such (not the files and directories they point to)
    ❗ specifies the type of sorting for display
      ❓ sort by date of last modification
          ✓ sort in reverse chronological order
      ❓ sort by date of last access
          ✓ sort in reverse chronological order
    ➔ output type and detail level options for display
      ❗ specifies the type of output for display
        ❓ uses a multi-column output
          ❗ specifies the type of sorting for columns in the output
              ❓ entries are sorted down the columns
              ❓ entries are sorted across the columns
          ❓ uses a one-column output
      ❗ specifies the level of detail in the output
        ❓ displays results in long format
          ❗ specifies the file size unit for display
              ❓ gives file sizes in bytes
              ❓ gives file sizes in blocks
          ❓ displays results in short format
```

Figure 5: Factorized abstract semantic tree

a compact representation (a factorized abstract semantic tree) for interactive negotiation with an expert. The output of this phase is a single abstract semantic tree, such as the one shown on figure 5 that is used for interactive validation. The ✓ icon represents a semantic object that dominates a semantic subtree containing no underspecifications; the ➔ icon represents a semantic object that does not take part in any underspecification, but which dominates a subtree that contains at least one; the ❗ icon represents a semantic type that is underspecified, that is for which at least two semantic objects are in competition; finally, the ❓ icon denotes semantic objects in competition, which are ordered for a given type by decreasing score of plausibility.

The MDA grammar used for analysis can then be used to produce the text associated with this tree, which corresponds to the normalized version of the input document that was validated by the expert.

The interface of our system displays an enumeration of all the underspecifications found in the compact representation. They are ordered by decreasing score, where the score can indicate the average score of the objects in competition, or the inverse of the average number of candidates per object in competition. Therefore, the expert can choose to resolve first underspecifications that contain likely objects, or underspecifications that involve few candidates so that the validation of an object will prune more candidates from the compact representations. Clicking on an underspecification in the list triggers a negotiation dialogue similar to that on figure 6. The semantic type on that dialogue, specifies how links are shown, is not supported by any evidence in the input document. The expert can however choose a value for it.



There should be an element of type
"specifies how links are shown"
Which one of the following values is correct?

● show the files and directories that are referenced by links
○ show links as such (not the files and directories they point to)

Figure 6: Example of a negotiation dialogue

5 Perspectives

We have presented a practical approach to content analysis at the level of communicative goals, in which a strong emphasis is put on document content well-formedness. Providing the expert is willing to spend enough time, the communicative content of a document can be interactively built. The better the system performance, the less time is needed to identify the correct candidate content representation. The fact that the expert can read the corresponding normalized text (on the MDA view) can help guarantee that the whole validation process was carried out correctly.

We now need to grow our grammars for Unix commands and drug leaflets, and to enrich our test corpus of annotated documents (raw text and abstract semantic structure)[6] for these classes in

[6] Documents for the test corpus can be obtained by using

order to be able to carry out evaluation. Evaluation should be performed on two aspects. First, the performance of fuzzy inverted generation could be measured, for a given normalization model and on a given source of documents, by the position and relative score of the candidate content representation corresponding to the normalized document. Second, we want to evaluate the usability of our user interface supporting interactive negotiation. An evaluation corresponding to the number of steps and the time needed to obtain the normalized version of a document would be a good indicator.

Moreover, we plan to implement the possibility for the expert to add new formulations found in documents to better match communicative goals in subsequent normalizations. It will then be interesting to evaluate the impact of this kind of supervised learning on system performance and user acceptance. Our next challenge will be to investigate how our approach can be applied to documents in less-constrained domains for which normalization models cannot be entirely built a priori.

Acknowledgments

Many thanks to Marc Dymetman and Christian Boitet for their supervision of this work.

References

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

Avi Arampatzis, Th. P. van der Weide, P. van Bommel, and C.H.A Koster. 2000. Linguistically Motivated Information Retrieval. *Encyclopedia of Library and Information Science*, 69.

Hervé Blanchon. 1994. *LIDIA-1: une premire maquette vers la TA interactive pour tous*. Phd thesis, Université Joseph Fourier, Grenoble.

Hervé Blanchon. 2002. A Pattern-based Analyzer for French in the Context of Spoken Language Translation: First Prototype and Evaluation. In *Proceedings of COLING-02, Taipei*.

Caroline Brun and Caroline Hagège. 2003. Normalization and Paraphrasing using Symbolic Methods. In *Proceedings of the 2nd International Workshop on Paraphrasing (IWP2003) at ACL-03, Sapporo, Japan*.

Caroline Brun, Marc Dymetman, and Veronika Lux. 2000. Document Structure and Multilingual Authoring. In *Proceedings of INLG 2000, Mitzpe Ramon, Israel*.

Marc Dymetman, Veronika Lux, and Aarne Ranta. 2000. XML and Multilingual Document Authoring: Convergent Trends. In *Proceedings of COLING 2000, Saarbrucken, Germany*.

Marc Dymetman, Aurélien Max, and Kenji Yamada. 2003. Towards Interactive Text Understanding. In *Proceeding of ACL-03, interactive posters session, Sapporo, Japan*.

Anthony F. Hartley and Cécile L. Paris. 1997. Multilingual Document Production - From Support for Translating to Support for Authoring. *Machine Translation*, 12:109–128.

Aurélien Max and Marc Dymetman. 2002. Document Content Analysis through Inverted Generation. In *Actes de l'atelier Using (and Acquiring) Linguistic (and World) Knowledge for Information Access du AAAI Spring Symposium Series, Université Stanford, Etats-Unis*.

Aurélien Max. 2003a. *De la création de documents normalisés à la normalisation de documents en domaine contraint*. Phd thesis, Université Joseph Fourier, Grenoble.

Aurélien Max. 2003b. Reversing Controlled Document Authoring to Normalize Documents. In *Proceedings of the EACL-03 Student Research Workshop, Budapest, Hungary*.

Alexis Nasr. 1996. *Un modèle de reformulation de phrases fondé sur la théorie Sens-Texte. Application aux langues contrôlées*. Phd thesis, Université Paris 7.

Scott P. Overmyer, Benoit Lavoie, and Owen Rambow. 2001. Conceptual Modeling through Linguistic Analysis using LIDA. In *Proceedings of the 23rd international conference on Software Engineering, ICSE, Toronto, Canada*.

Fernando Pereira and David Warren. 1980. Definite Clauses for Language Analysis. *Artificial Intelligence*, 13.

our system on each document, or by re-creating the content with the MDA system. Building a significant corpus is a time-consuming task that we have not finished yet.