ALVIS format of annotations
Task WP3.2
DRAFT - version 2.0

Guillaume Vauvert

March 30, 2005

Contents

Questions

This paper proposes a format of annotations and raises several issues:

1. About schemas (section 3.3): What is the more suitable language of schema (XML Schema, Relax NG) ? What tool to use to validate XML documents ? What's about our idea to use meta-schema ? How to validate partially annotated documents (partially may mean a part of the text, or a part of layers of annotations) ? How to organize partial schemas for partial validation ?

2. About processing: We assumed that a corpus is processed document by document. Do you validate this hypothesis ? Is it OK for statistical processing ?

3. About annotation content (section 1.3): What kind of semantic information do you need ?

4. About correction (section 3.2): Is our proposition enough powerfull and not too complex ?

5. About data volume (section 4.1): Be aware that this format considerably increases the document size. Is it OK ? Could compressed documents be used instead ?

6. About roles in Alvis annotating task (section **??**): WP5 and WP7 will do the following tasks: tagging of tokens, words, sentences, sections, paragraphs, titles, abstracts. From the Copenhaguen discussion, we assumed that WP7 is in charge of the structural annotations: sections, paragraphs, titles and abstracts (that requires to access to the original document). We propose that WP7 also assures tokenization (according to our definition, it is an automatic and not ambiguous process). The linguistic annotations must be processed in the following order: tokenization, named entities, other words, sentences. WP5 makes tag the words (which requires linguistic ressources). The main issue of tagging the sentences is to determine if a point (".") joined to a sequence of letters is a part of the word (Mr., www.ibm.com) or if it is a full stop marker. Tagging marks words or to tag sentences is the same issue: to determine the role of the point ".".

So, to make both in the same time is probably the best solution. Among these NLP steps, the precise role of WP7 and WP5 is not clear yet.

7. About the “light versions” of annotated documents (section **??**): How to access quickly (without parsing all the document) the desired annotations ? For validation, see “About schemas” above.

Chapter 1

Introduction

From a Natural Language Processing point of view, Alvis is mainly a process of annotation. Some packages add annotations (WP5 and WP7), some other use them for their tasks (WP2 and WP3 for classification and retrieval, WP6 for acquisition). So, the format of annotation is a central problem in Alvis.

1.1 Our goal

Our goal is to provide a format of annotation (called "alvis format") that enables tools to exchange annotations. Actually, each tool provides some annotations in a *adhoc* format that may vary.
Tools are assumed to be impossible (close-source tools) or hardly (compexity of tools) to modify. So, the format translation must be transparent for them.
Here is the annotation process we propose (see figure 1.1):

1. some annotations (that refer to a particular text) are available as input for a tool in "alvis format";
2. a subset of annotations is selected:only a sub part of the annotation may be required (a level of annotation, a part of the text);
3. it is translated (to be understood by the tool) into the tool's input format; some informations needed for further alignment may also be stored;
4. the tool processes the annotations;
5. it produces a new separate set of annotations in the tool's output format (or a set of annotations that are mixed with the input annotations);
6. this output is translated into the "alvis format", eventually using the alignment information stored before;