Representing and Accessing Multilevel Linguistic Annotation using the MEANING Format

Emanuele Pianta
ITC-irst
38050, Povo
Trento, Italy
pianta@itc.it

Luisa Bentivogli
ITC-irst
38050, Povo
Trento, Italy
bentivo@itc.it

Christian Girardi
ITC-irst
38050, Povo
Trento, Italy
cgirardi@itc.it

Bernardo Magnini
ITC-irst
38050, Povo
Trento, Italy
magnini@itc.it

Abstract

We present an XML annotation format (MEANING Annotation Format, MAF) specifically designed to represent and integrate different levels of linguistic annotations and a tool that provides flexible access to them (MEANING Browser). We describe our experience in integrating linguistic annotations coming from different sources, and the solutions we adopted to implement efficient access to corpora annotated with the Meaning Format.

1 Introduction

It is well known that when using XML-based annotation schemes to represent multi layer annotations, it can be difficult to handle partially overlapping annotations. Annotating discontinuous elements may be considered as a variant of the same problem (Pianta and Bentivogli, 2004). Other difficulties can arise from the necessity of integrating manual and automatic annotations, as we will show in this paper.

One of the most effective solutions to the above mentioned problems is the so called *stand-off annotation*, based on the separation between textual data and annotations, and between various types of annotation, possibly pointing to same text. This approach has been systematically adopted in the design of MAF, a multilayer XML format developed for the EU-funded MEANING project, in the context of the creation of the Italian MEANING Corpus (Bentivogli et al., 2003).

In this paper we will describe our experience in the use of MAF, with special emphasis on how we solved issues related to representing annotation levels which come from different sources, and can possibly overlap. We will also give details about the solutions we adopted to allow for efficient access and human browsing of MAF standoff annotations.

The rest of the paper is organized as follows. Section 2 describes MAF and the types of annotations which have been represented with it. Section 3 reports on the integration into MAF of linguistic annotations coming from different sources. Section 4 illustrates the strategies adopted to make the information encoded in MAF quickly accessible. Finally, Section 5 presents the MEANING Browser, a tool for accessing and navigating corpora linguistically annotated with MAF.

2 The MEANING Format

Following the proposals for the ISO/TC 37/SC 4 standard for linguistic resources (Ide and Romary, 2002), the MAF scheme is based on annotation structures and data categories. Each type of *annotation structure* (nestable <struct> elements) corresponds to a specific kind of linguistic object (e.g. tokens, lexical units, multiwords), and each instance of a linguistic object is identified by a unique identifier. *Data categories* (<feat> tags) represent attributes of the linguistic objects. Different representation levels are contained in separate documents, or document sections. The XLink and XPointer syntax is used to represent relations between elements in different XML documents, and IDREFs attributes for relations within the same document.

2.1 First version

The first version of the MEANING Format has been used to represent seven kinds of information: orthographic features, the structure of the text, morphosyntactic information, multiwords, syntactic information, named entities, and word senses.

Annotation levels are related to each other following a hierarchy of annotation levels, which reflects a theoretically grounded hierarchy of linguistic objects. The basic (orthographic) annotation level, representing tokens, is implemented with pointers to the character positions in the hub corpus. Then the morphosyntactic level, representing word-related morphological information, contains pointers to the tokens, whereas the multiword level points to the words described at morphosyntactic level.

The following example shows how the morphosyntactic features of the Italian word "andare" (to go) are represented.

```
<struc   type="w-level" id="w_12"
         xlink:href="#xpointer(id('t_10'))">
   <feat type="lemma">andare</feat>
   <feat type="stem">and</feat>
   <feat type="pos">v</feat>
   <feat type="elra-tag">VF</feat>
   <feat type="mood">inf</feat>
   <feat type="tense">pres</feat>
 </struc>
```

MAF also specifically addresses the problem of *discontinuous units*, such as for instance non-contiguous multiwords; see "*andarci* veramente *piano*" (*take it* really *easy*). A detailed study of how standoff annotation allows for an elegant treatment of this phenomenon can be found in (Pianta and Bentivogli 2004).

2.2 Second version

The first version of the MEANING Format has recently been extended within the FU-PAT ONTOTEXT project (Magnini et al. 2005).

Within this project, we are creating the Italian Content Annotation Bank (I-CAB), a corpus of Italian news stories annotated with different kinds of semantic information. Annotation is being carried out manually, as we intend I-CAB to become a benchmark for automatic Information Extraction and Ontology Population tasks, including recognition and normalization of various types of entities, temporal expressions, relations between entities, and relations between entities and temporal expressions (e.g. the relation *date-of-birth* connecting a person to a date).

To fulfill I-CAB annotation needs, we extended MAF, by adding a number of new linguistic annotation levels, i.e.:

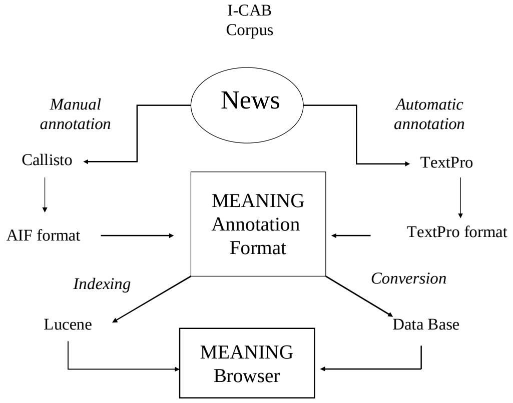
- temporal expressions
- entities of type person and organization
- mentions (i.e. the textual expressions referring to the entities)

According to the hierarchical approach to representing relations between annotation levels in the first version of the MEANING Format, temporal expressions and entity mentions are represented with pointers to morphosyntactic level entities. Entities, instead, are represented with pointers to entity mentions.

To manually annotate temporal expressions we followed the TIMEX2 markup standard, while to mark entities and mentions we relied on the ACE entity detection task guidelines. To perform the annotation task we used Callisto (http://callisto.mitre.org).

3 Converting linguistic annotations into MAF

The manual annotations produced through Callisto, which is related to novel annotation levels such as temporal expressions and entity mentions, had to be integrated with more traditional annotations which are performed automatically with the TextPro tool, an automatic linguistic analysis Tool Suite developed at ITC-irst.



As one can see in the above figure, two different annotation processes (automatic and manual) produce two different formats which must be converted and integrated into MAF in order to be accessed by the MEANING Browser (or any other NLP tool).

3.1 From TextPro format to MEANING Format

TextPro takes a raw text as input and carries out basic processing tasks such as tokenization, morphological analysis, PoS tagging, lemmatization, and multiword recognition. The results of TextPro analyses are represented in a table, where each token is on a row, and columns contain multiple annotation levels. Converting from the TextPro to the MEANING Format requires retrieving the character positions of tokens in the hub corpus, which are not directly available in the TextPro output.

3.2 From AIF format to MEANING format

The Callisto manual annotation tool produces a coding format called AIF (Atlas Interchange Format), which implements a stand-off XML annotation scheme.

When using the Callisto graphical interface, all annotations of temporal expressions and entity mentions are carried out by selecting a sequence of contiguous characters. As a consequence, all AIF annotations make reference to character positions.

However, from Section 2.2 we know that in MAF temporal expressions and entity mentions make reference to morphosyntactic linguistic objects, not characters. This implies that, to go from AIF to the MEANING Format, we need to translate annotations making reference to the position of characters into annotations that point to morphological entities. More precisely, we need to substitute pointers to character positions with pointers to morphosyntactic objects which have been marked automatically by TextPro. Carrying out this step will also achieve the integration of manual and automatic annotations.

The integration step is possible because the MAF hierarchy of annotation levels points, at the lowest level, to character positions. By following the hierarchy of links relating the various annotation levels it is always possible to trace back a linguistic object to some sequence of characters in the raw text, and in the opposite direction, given a string, we know what linguistic objects correspond to it. Summing up, the integration of AIF annotations into MAF requires that, given the character positions contained in the AIF annotation of some string, we substitute the pointers to characters with the pointers to the linguistic objects that cover the same string.

4 Data Access

MAF turned out to be a flexible and expressive means to represent and integrate multiple levels of linguistic annotation. This was achieved mainly thanks to the adoption of the standoff annotation approach. However accessing and retrieving information spread in possibly very large repositories (hundreds of thousands) of XML files may be a challenging task even for Database Management Systems specifically designed to handle XML. To solve this problem we first analyzed existing native XML databases such as eXist, and Apache Xindice, but found that what was available at the time did not suited our needs. For this reason we approached the access problem through a two-fold strategy:

- converting XML data into a relational database
- indexing XML data and accessing them through a search engine (LUCENE)

The conversion of MAF data into a relational database is based on the following strategy. Each annotation level is mapped into a table, where rows represent instances of the relevant linguistic object (e.g. words), and columns represent its attributes (e.g. lemma, PoS, etc). Specific columns contain the object identifiers and the pointers to objects of other types/tables.

Once MAF data are stored in a relational database, they can be accessed quite efficiently. However, when the access to data requires joins of many tables, access times become incompatible with various kinds of applications, such as on-line corpus browsing. For this reason we tried to complement the use of a relational database with the exploitation of the indexing capability of the LUCENE search engine (http://lucene.apache.org/). To this extent we modified the LUCENE analyzer so as to be able to parse XML structures. In this way LUCENE can be configured in order to index any XML structure.

The fast access capabilities of a relational database combined with the extended indexing capabilities of LUCENE enabled us to implement a browser of MAF annotated corpora.

5 The MEANING Browser

The MEANING Browser can be used by humans to navigate any corpus encoded with MAF. The browser is built upon an API which can be used by any automatic system.

In the following, we are going to demonstrate how I-CAB texts and their annotations can be accessed through the MEANING Browser.

The first kind of access to the corpus is word-oriented, and amounts to a concordancer, i.e. a tool able to provide all the occurrences of a certain word in the corpus. The user can alternatively search for all occurrences of a *word form*, or a *lemma*, possibly constraining the search to a certain PoS. Free combinations between these constraints are allowed. The system will return a KWIC-like concordance of all the tokens in the corpus that match the request, within a chosen word window. By clicking on the magnifying glass, one can see the sentence in which the searched word occurs (see Appendix 1).

By clicking on a specific icon a new window is opened where the whole text is displayed and its linguistic annotations are made accessible. A number of graphical widgets allow the user to highlight the desired annotations: e.g. nouns, verbs, multiwords, temporal expressions, mentions of a specific entity.

In Appendix 2 the browser is used to show both nouns (automatically annotated) and entity mentions (from manual annotation). Appendix 3 shows time expressions and discontinuous multiwords; see how the multiword “ha rassegnato … le dimissioni” (*he resigned)* is made discontinuous by the occurrence of a time expression ieri (*yesterday*). The browser will also give morphosyntactic information about single words composing multiwords (governo, *government*). From the same window one can access the XML files encoding multiple annotation levels for the same document.

References

Bentivogli, L., Girardi, C., Pianta, E. 2003. The MEANING Italian Corpus. In *Proceedings of the Corpus Linguistics 2003 conference*, Lancaster, UK.

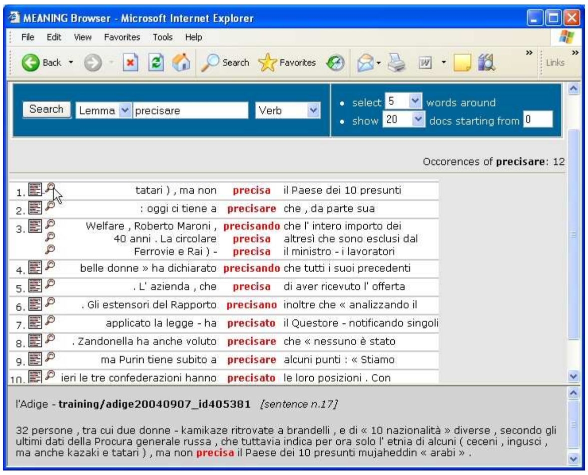
Ide, N. & Romary, L. 2002. Standards for Language Resources. In *Proceedings of LREC 2002*, Las Palmas, Canary Islands, Spain.

Magnini, B., Negri, M., Pianta, E., Romano, L., Speranza, M., Serafini, L., Girardi, C., Bartalesi, V., Sprugnoli, R. 2005. From Text to Knowledge for the Semantic Web: the ONTOTEXT Project. In *Proceedings of SWAP 2005 Workshop*, Trento, Italy.

Pianta, E. and Bentivogli, L. 2004. Annotating Discontinuous Structures in XML: the Multiword Case. In *Proceedings of the LREC 2004 Satellite Workshop on "XML-based richly annotated corpora"*, Lisbon, Portugal.

Appendix 1

Kwic Concordancer


MEANING Browser - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Links

Search | Lemma | precisare | Verb

- select 5 words around
- show 20 docs starting from 0

Occorences of **precisare**: 12

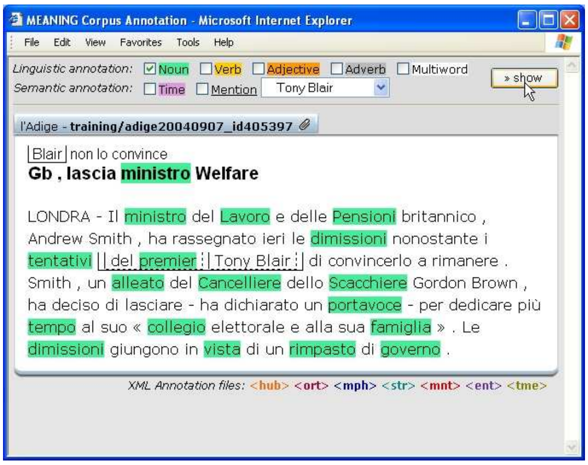
1.	tatari) , ma non	**precisa**	il Paese dei 10 presunti
2.	: oggi ci tiene a	**precisare**	che , da parte sua
3.	Welfare , Roberto Maroni , 40 anni . La circolare Ferrovie e Rai) -	**precisando** **precisa** **precisa**	che l' intero importo dei altresì che sono esclusi dal il ministro - i lavoratori
4.	belle donne » ha dichiarato	**precisando**	che tutti i suoi precedenti
5.	. L' azienda , che	**precisa**	di aver ricevuto l' offerta
6.	. Gli estensori del Rapporto	**precisano**	inoltre che « analizzando il
7.	applicato la legge - ha	**precisato**	il Questore - notificando singoli
8.	. Zandonella ha anche voluto	**precisare**	che « nessuno è stato
9.	ma Purin tiene subito a	**precisare**	alcuni punti : « Stiamo
10.	ieri le tre confederazioni hanno	**precisato**	le loro posizioni . Con

l'Adige - **training/adige20040907_id405381** *[sentence n.17]*

32 persone , tra cui due donne - kamikaze ritrovate a brandelli , e di « 10 nazionalità » diverse , secondo gli ultimi dati della Procura generale russa , che tuttavia indica per ora solo l' etnia di alcuni (ceceni , ingusci , ma anche kazaki e tatari) , ma non **precisa** il Paese dei 10 presunti mujaheddin « arabi » .

Appendix 2

Browsing nouns (in grey, automatic annotation) and entity mentions (Tony Blair, manual annotation)


MEANING Corpus Annotation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
» show

Linguistic annotation: ☑ Noun □ Verb □ Adjective □ Adverb □ Multiword
Semantic annotation: □ Time □ Mention Tony Blair

l'Adige - **training/adige20040907_id405397**

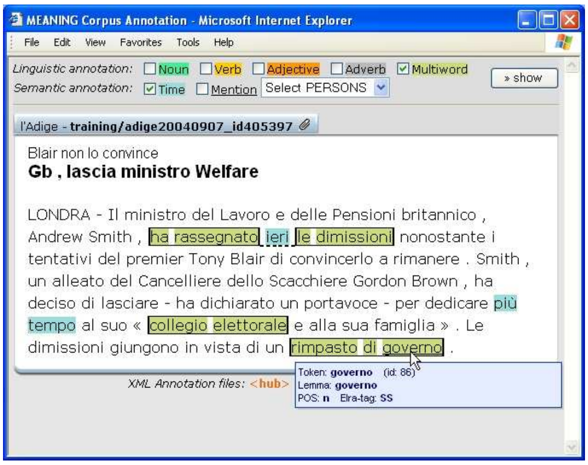
Blair non lo convince
Gb , lascia ministro Welfare

LONDRA - Il ministro del Lavoro e delle Pensioni britannico , Andrew Smith , ha rassegnato ieri le dimissioni nonostante i tentativi del premier Tony Blair di convincerlo a rimanere . Smith , un alleato del Cancelliere dello Scacchiere Gordon Brown , ha deciso di lasciare - ha dichiarato un portavoce - per dedicare più tempo al suo « collegio elettorale e alla sua famiglia » . Le dimissioni giungono in vista di un rimpasto di governo .

XML Annotation files: <hub> <ort> <mph> <str> <mnt> <ent> <tme>

Appendix 3

Browsing discontinuous multiwords (ha rassegnato … le dimissioni, *he resigned*), time expressions (ieri, *yesterday*) and word information (governo)


MEANING Corpus Annotation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
» show

Linguistic annotation: □ Noun □ Verb □ Adjective □ Adverb ☑ Multiword
Semantic annotation: ☑ Time □ Mention Select PERSONS

l'Adige - **training/adige20040907_id405397**

Blair non lo convince
Gb , lascia ministro Welfare

LONDRA - Il ministro del Lavoro e delle Pensioni britannico , Andrew Smith , ha rassegnato ieri le dimissioni nonostante i tentativi del premier Tony Blair di convincerlo a rimanere . Smith , un alleato del Cancelliere dello Scacchiere Gordon Brown , ha deciso di lasciare - ha dichiarato un portavoce - per dedicare più tempo al suo « collegio elettorale e alla sua famiglia » . Le dimissioni giungono in vista di un rimpasto di governo .

XML Annotation files: <hub>

Token: **governo** (id: 86)
Lemma: **governo**
POS: **n** Elra-tag: **SS**