KAFnotator: a multilingual semantic text annotation tool

Maurizio Tesconi, Francesco Ronzano, Salvatore Minutoli, Andrea Marchetti
Institute of Informatics and Telematics (IIT) CNR
Via G. Moruzzi, 1
56123 Pisa - Italy
{maurizio.tesconi, francesco.ronzano, salvatore.minutoli, andrea.marchetti}@iit.cnr.it

Carlo Aliprandi
Synthema S.r.l.
Language and Speech Solutions
Via Malasoma 24
56121 Ospedaletto (Pisa) - Italy
carlo.aliprandi@synthema.it

Abstract

At present, the availability of high quality annotated corpora is fundamental to carry out or to evaluate several Natural Language Processing and Text Mining tasks. To create consistently annotated corpora, direct human intervention represents a key factor: teams of manual taggers, usually composed by linguistically skilled people, are needed to refine existing annotations or to add new ones. As a consequence, manual corpora annotation is an expensive and a highly demanding task in term of involved resources.

In this paper we focus our attention on the annotation tools devoted to support and simplify as much as possible the job of manual taggers. In particular, after a brief description of the main issues concerning the process of manual annotation complemented by some relevant example of annotation tools, we describe KAFnotator: it is a Web-based environment useful to support manual semantic annotation of documents. It allows browsing texts annotated by exploiting the Knowledge Annotation Format (KAF) so as to refine and extend the meanings associated to their terms. Starting from a brief description of the KAF annotation format, we introduce KAFnotator also by providing an example of usage.

1 Introduction: text annotation tools

When we annotate a text we identify and characterize relevant items inside its contents with respect to the specific kind of annotation we deal with. In particular, in computational linguistics, there are different levels of annotation depending on the kind of informative content we want to identify inside a text: morphological, syntactic, semantic, discourse and pragmatic annotation. For all of these levels, we can build annotated corpora that are collections of annotated texts usually concerning the same domain: human intervention is needed to build high quality corpora and refine their annotations. Linguistically skilled people, referred to as manual taggers, are usually involved in this resource demanding task.

Many tools have been proposed to support and speed up manual tagging: they usually provide taggers with an accessible environment to browse a corpus together with its annotations so as to add new ones or to correct the existing ones. A corpus to be manually annotated has been often already enriched with a set of annotations obtained by exploiting automatic procedures: these annotations usually constitute the starting point of the job of manual taggers. The resulting manually annotated corpora are usually exploited as training data sets for automatic annotation procedures or as references to evaluate their results.

EtiFac (Branco, 2001) and **KCAT** (Ryu, 2000) represent two examples of corpus annotation tools dealing with the morphological and syntactic level of annotation: in particular they allow specifying the part of speech of words. They preprocess texts to be annotated by exploiting a set of syntactic rules so as to perform a first automatic annotation. Moreover EtiFac excludes from the terms to be annotated all the words belonging to closed classes. In order to allow taggers modifying texts annotations, they exploit desktop based applications, based for instance on text editor macros.

If we consider annotation tasks at the semantic level, we can for instance define inside a document through proper annotation formalisms the type of the Named Entities among those ones belonging to a specific group referred to as annotation schema or also we can specify the meaning of terms with respect to a collection of meanings usually represented by a lexical resource. (Agirre, 2006) describes the procedures and the special cases usually encountered while annotating a corpus with respect to the WordNet lexicon. (Laurenco, 2008) gives an example of a tool to annotate, with respect to an annotation schema made of 14 biological classes a corpus of biomedical documents.

OLLIE (Cunningham, 2003) is another example of Web-based collaborative annotation tool: it exploits a Java-based Web interface so as to manage the set of documents to be annotated as well as to deal with the annotations performed over a single document. It is integrated in the GATE[1] open-source language engineering infrastructure, developed by the Natural Language Processing Group at the University of Sheffield. **WordFreak** (Morton, 2003) is a Java-based document annotation infrastructure that can be easily adapted to different annotation schemas and applications. (Novak, 2007) describes a tool to manage large semantic networks so as to annotate natural language utterances in parallel text corpora.

Summarizing, all the analyzed examples of annotation tools are usually integrated with specific text processing environments so as to exploit their results; most of them are desktop based applications. They have little or no focus on the collaborative and distributed issues connected to corpora annotations.

In the rest of this paper we present KAFnotator: a Web-based semantic annotation tool useful to create and refine the term-to-meaning associations inside KAF annotated documents. This tool has been developed to support the manual annotation of a multi-lingual environmental corpus to be exploited as the reference corpus in the All-words Word Sense Disambiguation on a Specific Domain (WSD-domain) task at SemEval 2010[2]. Manual annotation consists in semantically annotating terms of the documents, referring them to WordNet synsets of the specific language.

In *Section 2* we briefly introduce KAF, the multilayered annotation format used by KAFnotator, defined in the context of the KYOTO system[3]. In *Section 3* we describe KAFnotator.

2 KAF: KYOTO Annotation Format

KAF (Knowledge Annotation Format) is a language neutral annotation format representing both morpho-syntactic and semantic annotation of documents through a stand-off multilayered structure (Marchetti et al., 2009).

KAF has been defined and developed in the context of the KYOTO Project, an Asian-European project that develops a community platform for modeling knowledge and finding facts across languages and cultures. KAF provides an open XML reference format for the representation of knowledge and facts, encapsulating the following annotations:

- Tokenizaton and word form segmentation;
- POS tagging;
- Lemmatization and Term Extraction;
- Constituency and Dependency Tagging;
- Named Entity Recognition (NER);
- Word Sense Disambiguation (WSD);
- Semantic Role Labeling (SRL)-

KAF provides a set of separate layers for:

- Sequences of words, sentences and pages;
- Sequences of terms;
- Sequences of constituent chunks;
- Sequences of semantic roles.

Each of these layers is interconnected through identifiers, so that each level of analysis can be related to the next level.

KAF adopts a stand off strategy for annotating the source text and is compatible with LAF, even if it imposes a more specific standardization of the annotation format itself.

A proper XML Syntax to represent KAF annotated documents has been defined, setting up a specific XML Schema.

We can distinguish in KAF three macro-layer of annotation (see Fig. 1):

- the *morpho-syntactic layer*: it groups all the language-specific textual annotations. Tokens, sentences and paragraphs are identified in a specific document. Terms made of words or multi-words are pointed out along with their POS. In this layer also functional dependences are

[1] GATE Web Site, http://gate.ac.uk/

[2] SemEval 2010 Web Site, http://semeval2.fbk.eu

[3] KYOTO Project Web Site, http://www.kyoto-project.eu/

represented as well as chunks that are constituents and phrases;

- the *level-1 semantic layer*: it includes linear annotation of expressions of time, events, quantities and locations;
- the *level-2 semantic layer*: it is mainly devoted to represent facts, in a non linear annotation context, thus possibly aggregating evidences from the lower layers of multiple textual sources.

KAFnotator mainly deals with the results of the morpho-syntactic annotation layer of KAF. In particular when a text is annotated at this layer, the following elements are identified: *word forms* inside a specific sentence of a defined paragraph, *terms* and *compound terms*, dependency relations between couples of terms and chunks (like Compound Nouns, Verbal Phrases or Propositional Phrases), spanning one or more terms.

For each term it is possible to specify its lemma, its part of speech and, optionally, its Named Entity type. Also *the link of a term to an external reference* like an entity of a knowledge

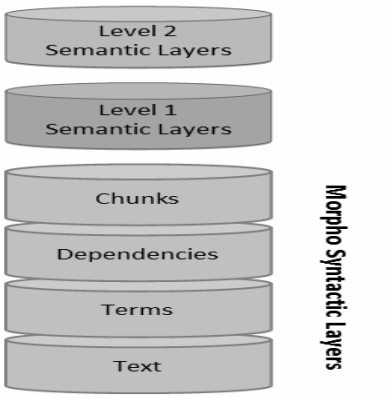

Figure 1. The Layers of KAF

base, a class of a ontology or one of its instances can be represented: in particular this KAF expressive feature is exploited *to connect terms to the meaning they refer to represented as a concept (or synset) in WordNet*. These links can be created or refined thanks to KAFnotator, as better described in the following Section.

3 KAFnotator

KAFnotator is a web based tool that helps users to add semantic annotation to KAF files.

Thanks to this system, users can quickly browse, show terms, choose sense and annotate them with simple click and drop-down lists that speed up the process when performing repetitive and recurring procedures.

The process of annotation consists:

1. Assign a language to each user;
2. Assign roles to users (choose between "blind annotator" and "adjudicator");
3. Start blind annotation process: each tagger tags all occurrences;
4. Resolve conflict: an adjudicator decides which is the final selection when there is a disagreement.

This is the list of features that made KAFnotator rather unique among the other semantic annotation tools:

- *Role Based Access*: users can have several roles and the system shows an appropriate interface and specific functionalities.
- *Multilingual*: each user can be assigned to a language and the terms into KAF files are disambiguated using correct WN by means of DEBVisDic server.
- *Multiversion*: for each annotation a new version is created, thus allowing rollback mechanisms.
- *XML based*: this tool is completely based on native XML database.

3.1 Architecture

The architecture of system is based on a client-server model. KAF files are stored in a XML repository and the system can access DebVisDic Server by means of a REST API in order to obtain a list of synsets related to a given term.

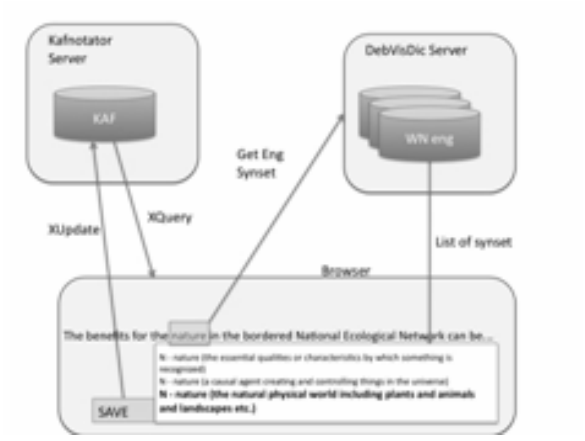

Figure 2: Overall Architecture

We developed the system using XQuery for visualization and XUpdate for annotation, the XML database is eXist and we have used JQuery for user interface. Ajax has been used for interaction between browser and server.

3.2 Annotate corpora with KAFnotator

After user authentication, a list of document is presented and the user can select a KAF file to begin the annotation process.

For the specific goals of the SemEval 2010 WSD-domain task, we limited the annotation to specific part of speech, namely Nouns and Verbs (highlighted in blue). When the user selects a term, the tool asks him to choose the sense among a list of synsets related to the term.

The term is marked in yellow after annotation and the information about sense, user and timestamp are saved into KAF. If the user has adjudicator role, the terms in conflict will be highlighted and the user can decide the correct sense.

4 Conclusions

We have developed KAFnotator, a Web-based tool for semantic annotation of documents. KAFnotator supports manual annotation of multi-lingual documents, allowing quick browsing, terms identification and sense annotation using underlying semantic resources like Ontologies and Wordnets. The tool, that will be used to annotate reference corpora for the SemEval 2010 WSD-domain task, speeds up the semantic annotation procedure via the automatic exploitation of Wordnet senses that are automatically referred to terms in the corpora.

Compared with other annotation tools, KAFnotator is characterized by the following features: it is a collaborative tool, with specific functionalities and an appropriate interface depending on the role of the user; each annotated document is saved and stored with its version, thus allowing a rollback mechanisms; the tool is completely based on native XML technology and database. Moreover KAFnotator benefits from the underlying structure of KAF (Knowledge Annotation Format), a multilayered annotation format that provides a robust and rich backbone for morpho-syntactic and semantic information.
We expect that these features will not only reduce repetitive works, but also improve quality and effectiveness of annotations.
Further details can be obtained from the Semeval-2010 WSD-domain task website[4].

The work described here is funded by the the European Commission (KYOTO FP7 ICT-2007-211423).

[4] http://xmlgroup.iit.cnr.it/SemEval2010/

References

Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Pociello E. and Quintian M. 2006. *Improving Basque WordNet by Corpus Annotation*. In the Proc. of the 3rd International WordNet Conference, South Jeju Island, Korea.

Agirre E., Lopez de Lacalle O., Fellbaum C., Marchetti A., Toral A. and Vossen P. 2009. *SemEval-2010 Task 17: All-words Word Sense Disambiguation*. In the Proc. of NAACL Workshop on Semantic Evaluations (SEW-2009). Boulder, Colorado.

Bosma W., Vossen P., Soroa A., Rigau G., Tesconi M., Marchetti A., Aliprandi C., Monachini M. 2009. *KAF: a generic semantic annotation format*. In the Proc. of the 5th International Conference on Generative Approaches to the Lexicon, Pisa, Italy.

Cunningham H., Tablan V., Bontcheva K., Dimitrov M. 2003. *Language engineering tools for collaborative corpus annotation*. In the in Proceedings of Corpus Linguistics 2003 Conference, Lancaster, United Kingdom.

Horta Branco A. and Ricardo Silva J. 2001. *EtiFac: A facilitating tool for manual tagging*, in the Proceedings of the Encontro Nacional da Associação Portuguesa de Linguística, Lisbon, Portugal.

Laurenco A., Carneiro S., Carreira R, Rocha M., Rocha I. And Ferreira E. 2008. *A Tool for Automatic and Manual Annotation of Biomedical Documents*, in the Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine, Turku, Finland.

Marchetti A., Minutoli S., Ronzano F. and Tesconi M. 2009. *Wikyoto Knoledge Editor: the collaborative environment to manage Kyoto Multilingual Knowledge Base*. In the Proc. Of the 6th International Conference on Knowledge Management (ICKM 09), Hong Kong, China.

Morton T. and LaCivita J. 2003. *WordFreak: An Open Tool for Linguistic Annotation*. In the Proc. of the 2003 Human Language Technology Conference - Demonstration, Edmonton, Canada.

Novak V. 2007. *Large Semantic Network Manual Annotation*, in the Proc. of the 2007 Human Language Technology Conference - Demonstration, New York, USA.

Won-Ho Ryu, Jin-Dong Kim, Hae-Chang Rim and Heui-Scok Lim. 2000. *KCAT: A Korean Corpus Annotating Tool Mininmizing Human Intervention*. In the Proc. of the 17th Conference on Computational Linguistics, Morristown, NJ, USA.