META-MultilanguagE Text Analyzer

Pierpaolo Basile, Marco de Gemmis, Anna Lisa Gentile,
Leo Iaquinta, Pasquale Lops, Giovanni Semeraro

[1]Department of Computer Science, University of Bari, Italy
{basilepp,degemmis,gentile,iaquinta,lops,semeraro}@di.uniba.it

Abstract

Natural Language Processing *(NLP)* has a significant impact on many relevant Web-based and Semantic Web applications, such as information filtering and retrieval. Tools supporting the development of NLP applications are playing a key role in text-based information access on the Web.

In this paper, we present META *(MultilanguagE Text Analyzer)*, a tool for text analysis, designed with the aim of providing a general framework for NLP tasks over different languages. The system implements both basic and advanced NLP functionalities, such as Word Sense Disambiguation. After describing the main ideas behind the architecture of META, we discuss some results about the processing of different corpora in English and Italian. Finally, we show how META has been integrated in a recommender system for content-based information filtering.

Index Terms: Natural Language Processing, Information Filtering, Document Indexing

1. Introduction

A vast portion of the Web consists of text documents, thus methods for automatically analyzing text have great importance in the context of the Web.

Several techniques have been developed within the fields of Information Retrieval (IR) and Information Filtering (IF), and include indexing, scoring, and categorization of textual documents. Filtering and retrieval systems deal with the ranking of textual documents in order of relevance. Retrieval refers to the selection of documents from a fixed set, whereas filtering typically refers to selection of relevant documents from a stream of incoming data. Retrieval systems are generally concerned with satisfying a users one-off information need (query); filtering systems are usually applied to attaining information for a users long term interests (profiles). Categorization or classification of documents is another useful technique, somewhat related to IR and IF, that consists of assigning a document to one or more predefined categories. A classifier can be used, for example, to distinguish between relevant and irrelevant documents (where the relevance can be personalized for a particular user or group of users), or to help in the semiautomatic construction of large Webbased knowledge bases or hierarchical directories of topics like the Open Directory[1].

In this scenario, the development of robust tools for both basic and more complex NLP tasks is becoming crucial. This paper describes META (*MultilanguagE Text Analyzer*), an infrastructure for processing textual documents over different languages. The main features of the proposed tool are:

- The system is designed to clearly separate low-level tasks (such as data storage, location and loading of language resources) from data structures and algorithms.
- The tool provides a baseline set of NLP components (Tokenizer, POS-tagger, ...) that can be extended and modified by the user according to the tasks to be accomplished.
- The architecture was conceived so that language-independent components for both basic and more complex tasks, such as Word Sense Disambiguation, can be easily included.
- Indexing structures produced by the META can be exported in different formats, thus allowing an easy integration with both IR and IF systems.

The rest of the paper is structured as follows. We first describe the META architecture in Section 2, then we provide some detail about document representation in Section 3. Some application scenarios are reported in Section 4, while a brief description of related work is given in Section 5, while conclusions and future work close the paper.

2. System Architecture

The architecture of META is depicted in Fig. 2, in which the three main components of the system are showed:

1. COLLECTION MANAGER - This component provides the tools for the import of documents in different formats (HTML, PDF, DOC,...), allows the user to organize them in collections, and includes algorithms for the segmentation of documents, that is each document is logically viewed as structured in different sections (e.g., a scientific paper can be structured into: *title*, *abstract*, *authors*, *body* and *references*). The COLLECTION MANAGER allows also the annotation of sections with tags stored in a domain ontology;
2. NLP ENGINE - This engine is devoted to the management of different NLP annotators. An annotator is a component that performs a specific NLP task (e.g. tokenization, stop word elimination, POS-tagging). The NLP ENGINE schedules the annotators, loads the lexical resources required for each annotator, and runs the annotator over all the documents into a collection.
3. EXPORT MANAGER - This component is able to export the results carried out by the NLP ENGINE into different formats, according to the user's request (XML, RDF, specific DBMS, ...).

The whole process of document analysis performed by META is described in the following. The COLLECTION MANAGER imports the documents to be processed from the user's

[1]http://dmoz.org/

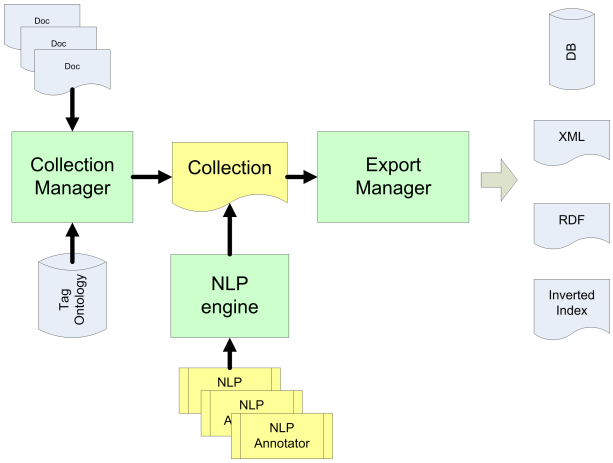

Figure 1: META conceptual architecture

file system (HTML, DOC, RTF, PDF) and groups them in a *collection*. Each document is assigned with a unique identifier (ID) in the collection, then segmentation is performed and the raw text is extracted from the original document. In this stage, it is also possible to associate both collections and single documents with tags stored in a domain ontology. After these preliminary steps, the documents are ready for the next stage performed by the NLP ENGINE.

First, the *NLP Engine* detects the document language; this step is strictly required in order to load the right lexical resources for each language. Then, the NLP engine normalizes (for example, all formatting characters are removed) and tokenizes the text. At this stage, each document is turned in a list of tokens. Each token can be associated with a set of *annotations*. An annotation is a pair *(annotation_name,value)*, which specifies the kind of annotation and the corresponding value (e.g., the position of the token in the text). Annotations are produced by different components called NLP ANNOTATORS, whose scheduling is managed by the NLP ENGINE.

Currently, the following annotators have been developed and included in META:

1. *Stop words elimination*: all commonly used words are deleted;
2. *Stemming*: it is the process of reducing inflected (or sometimes derived) words to their stem. In META, we adopt the *Snowball stemmer* [2];
3. *POS-tagging*: it is the process of assign a part-of-speech to each token. We develop a JAVA version of *ACOPOST tagger* [3] using Trigram Tagger T3 algorithm. It is based on Hidden Markov Models, in which the states are tag pairs that emit words;
4. *Lemmatization*: it is the process of determining the lemma for a given word. We use WordNet Default Morphological Processor (included in the WordNet distribution) for English. For the Italian language, we have built a different lemmatizer that exploits the *Morph-it!* morphological resource [4];
5. *Entity Recognition Driven by Ontologies*: it is the process of finding ontology instances into the text;
6. *Word Sense Disambiguation* (WSD): it is the problem of selecting a sense for a word from a set of predefined possibilities, by exploiting a sense inventory that usually comes from a electronic dictionary or thesaurus. We have implemented a WSD algorithm, called JIGSAW [1], able to disambiguate both English and Italian text.

At the end of the pipeline ran by the NLP ENGINE, the output could be exported in different formats by the EXPORT MANAGER. This component is devoted to turn the internal output produced by META into different formats such as XML or RDF.

3. Document representation

The internal representation of META is a collection that contains a list of documents. Each document is subdivided into segments, each one corresponding to a specific part of the document. Documents are composed by one segment at least. Each segment contains a list of token, each one associated with one annotation at least. An annotation represents a particular feature extracted during text processing (e.g. token, stemming, lemma, entity, sense, ...). The logical structure of a document in depicted in Fig. 3.

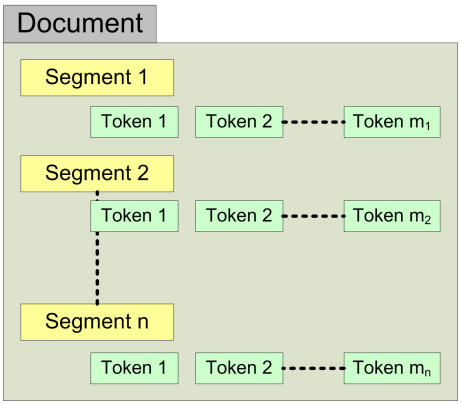

Figure 2: Conceptual document structure

For example, if the user want to analyze the text '*In this paper we present META (MultilanguagE Text Analyzer), it's a tool for text analysis which implements some NLP functionalities.*'. The NLP ENGINE executes the following operations: tokenization, stemming, pos-tagging, lemmatization and WSD. The output of the system is a list of tokens and corresponding annotations. Fig. 3 shows the logical structure for the token *paper*. In particular, the token has a *sense annotation* produced by the WSD annotator, whose value is *n12660433*, the number which identifies the WordNet synset assigned by JIGSAW.

The snapshot of the META GUI that represents the output of the above example is showed in fig. 3. The GUI of the system allows the visualization of the output by using a table format: tokens are represented in rows and annotations in columns. Also, from the GUI it is possible to access EXPORT MANAGER functionalities.

[2] http://snowball.tartarus.org/
[3] http://acopost.sourceforge.net/
[4] http://sslmitdev-online.sslmit.unibo.it/linguistics/morph-it.php

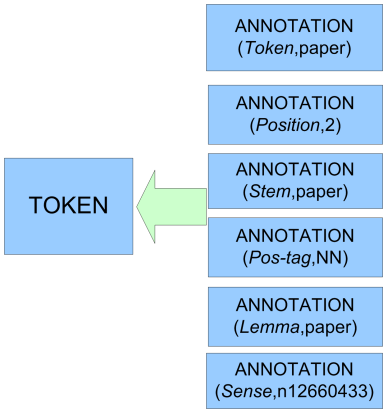

Figure 3: Conceptual token structure

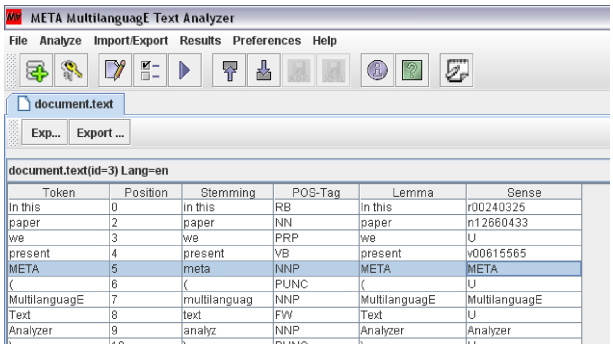

Figure 4: META GUI snapshot

4. META @ Work

META has been employed for the processing collection of documents in different scenarios, in order to evaluate its performance:

1. Disambiguation of a whole collection of document in English;
2. Disambiguation of a whole collection of document in Italian;
3. Indexing of a whole collection of scientific papers for personalized filtering;

In the following, we describe each one of the scenario in which we system was tested.

4.1. WSD on English

JIGSAW, the WSD algorithm included in META, has been tested in the context of SemEval 1-Task 1 competition [2]. This task is an application-driven one, where the application is a fixed Cross-Lingual Information Retrieval (CLIR) system. Participants must disambiguate text by assigning WordNet synsets, then the CLIR system must perform both the expansion to other languages and the indexing of the expanded documents; the final step is the retrieval (in batch) for all the languages. The retrieved results are taken as a measure of the disambiguation accuracy. The dataset consisted of 29,681 documents, including 300 topics (short text). Results are reported in Table 1. Besides the two systems (JIGSAW and PART-B) that participated to SEMEVAL-1 Task 1 competition, a third system (ORGANIZERS), developed by the organizers themselves, was included in the competition. The systems were scored according to standard IR/CLIR measures as implemented in the TREC evaluation package[5].

system	*IR documents*	*IR topics*	*CLIR*
no expansion	0.3599		0.1446
full expansion	0.1610	0.1410	0.2676
1st sense	0.2862	0.1172	0.2637
ORGANIZERS	0.2886	0.1587	0.2664
JIGSAW	0.3030	0.1521	0.1373
PART-B	0.3036	0.1482	0.1734

Table 1: SEMEVAL-1 Task1 Results

All systems showed similar results in IR tasks, while their behaviour was extremely different on CLIR task. Probably, the negative results of JIGSAW in CLIR task depends on complex interaction of WSD, expansion and indexing. Contrarily to other tasks, the task organizers do not plan to provide a ranking of systems on SEMEVAL-1 Task 1. As a consequence, the goal of this task - what is the best WSD system in the context of a CLIR system? - is still open.

4.2. WSD on Italian

An important applications scenario is EVALITA [6], that is an initiative devoted to the evaluation of Natural Language Processing tools for Italian. In this context, we have evaluated META for Italian language. Experiments were performed by using the instructions for EVALITA WSD All-Word-Task. The dataset consisted of about 5000 words. Precision and Recall are reported in Table 2.

SYSTEM	*P*	*R*	*attempted*
UniBa_Basile (JIGSAW)	0.560	0.414	73.95%
1st sense (baseline)	0.669	0.669	100%

Table 2: JIGSAW results on EVALITA All-Words Task

The results are encouraging as regards precision, considering that our system exploits only ItalWordNet as knowledge base. JIGSAW was compared only with the *baseline* (for all words, the first sense in ItalWordNet is selected), which achieves very high results. In Table 3 the precision for each POS-tag is showed. It is possible to notice that the precision is quite acceptable for nouns, and very high for proper nouns because generally they have only a sense. The results show that the verb disambiguation is very hard due to high polysemy. High precision is achieved for adjectives and adverbs, but recall is lower due to POS-tagger errors. The process of WSD requires lemmatization and POS-tagging, which introduce errors, thus influencing the recall. We estimated lemmatization and POS tagging precision respectively to 77,66% and 76,23%. More details are reported in [3].

4.3. META in an Information Filtering Scenario

META has been used as Content Analyzer into a content-based recommender system [4]. The recommender automatically infers the user profile, a structured model of the user interests,

[5] http://trec.nist.gov/
[6] http://evalita.itc.it/

$POS-tag$	P	R	$attempted$
$NOUN$	0,556	0,444	79,96%
$VERB$	0,375	0,283	75,60%
$OTHERS$	0,676	0,321	47,55%
$PROPERNOUN$	0,913	0,724	79,25%

Table 3: JIGSAW results for each POS-tag on EVALITA All-Words Task

from documents that were already deemed relevant by the user. The profile is used to filter new documents and to produce personalized suggestions. We used META in the indexing phase for the extraction of both lexical and semantic features from documents. The learning algorithms embedded in the recommender are able to infer user profiles from the feature produced by META. The system produced both a classical Bag-Of-Word (BOW) document representation and a new representation that we call Bag-of-Synset (BOS). In this model, a document is represented by a vector of WordNet synsets recognized by the WSD procedure.

5. Related Work

The design of META was strongly inspired by GATE-General Architecture for Text Engineering [7] developed by the NLP group at Sheffield University. GATE [5] is an infrastructure for developing and deploying software components that process human language. GATE helps scientists and developers in three ways: by specifying an architecture for language processing software; by providing a framework, or class library, that implements the architecture and can be used to embed language processing capabilities in various applications; by providing a development environment built on top of the framework made up of convenient graphical tools for developing components. The goal of GATE is to enable users to develop and deploy language engineering components and resources in a robust fashion. On the other hand, META is a tool for the management of documents collections, the organization of multi-lingual NLP pipelines, and the storage of processed documents.

The main differences between META and GATE are:

1. META provides powerful tools for both the management of collections and document segmentation and annotation;
2. META provides an NLP pipeline that allows the development of NLP annotators not only for English;
3. META is oriented toward semantic indexing of documents, by making easier the integration of WSD algorithms;
4. META was not designed specifically for information extraction or text mining as GATE, but it is possible to convert the output produced by META in several formats. Therefore, META is prepared for the export also in formats required by external mining tools like WEKA[8].

UIMA [9] is a framework for NLP developed by IBM. The UIMA framework is an open, industrial-strength, scalable and extensible platform for building analytic applications or search solutions that process text or other unstructured information to find the latent meaning, relationships and relevant facts buried within. It enables developers to build analytic modules and to compose analytic applications from multiple analytic providers, encouraging collaboration and facilitating value extraction for unstructured information. UIMA is able to deal with both text and other media format such as videos and images.

In conclusion, META is more useful for NLP tasks that require the indexing of documents and the extraction of semantic features from text.

[7] http://gate.ac.uk

[8] http://www.cs.waikato.ac.nz/ml/weka/

[9] http://uima-framework.sourceforge.net/

6. Conclusion and Future Work

NLP tools has a crucial role in the success of Semantic Web technologies because they provide an automated way to extract semantic features from text. In this paper, we described META, a tool that support the development of NLP applications. This component allows the management of collection of documents and provide an engine able to run different NLP operations on documents. The output of this operations could be exported in different way or could be used in Information Retrieval, Information Filtering or Information Extraction tasks.

An ongoing work in which META is involved is the adoption of the system as indexer for a semantic search engine designed and developed in our lab. This search engine provides different document representations that we call *levels*. Each level has a local scoring function, then a global ranking function is defined in order to merge the results produced by local scoring functions. META is adopted to build the different levels of document representation. at the moment, we have three levels: keyword, synset and entity.

As future work, we plan to develop new components able to carry out statistical report on the extracted features. We are working also in order to provide tools for the evaluation of NLP algorithms included in the META pipeline.

7. Acknowledgements

We would like to thank Franco Grieco for his help in the design and development of the first release of META, and all the students who contributed to improve META.

8. References

[1] P. Basile, M. de Gemmis, A. Gentile, P. Lops, and G. Semeraro, "Jigsaw algorithm for word sense disambiguation," in *SemEval-2007: 4th Int. Workshop on Semantic Evaluations*. ACL press, 2007, pp. 398–401.

[2] E. Agirre, B. Magnini, o. Lopez de Lacalle, A. Otegi, G. Rigau, and Vossen, "Semeval-2007 task 1: Evaluating wsd on cross-language information retrieval," in *SemEval-2007: 4th Int. Workshop on Semantic Evaluations*. ACL press, 2007, pp. 1–6.

[3] P. Basile and G. Semeraro, "Jigsaw: an algorithm for word sense disambiguation," *Rivista dell'Associazione Italiana per l'Intelligenza Artificiale*, vol. IV(2), pp. 53–54, 2007.

[4] G. Semeraro, M. Degemmis, P. Lops, and P. Basile, "Combining learning and word sense disambiguation for intelligent user profiling." in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence IJCAI-07*, 2007, pp. 2856–2861, m. Kaufmann, San Francisco, California. ISBN: 978-I-57735-298-3.

[5] H. Cunningham, "Information Extraction, Automatic," *Encyclopedia of Language and Linguistics, 2nd Edition*, 2005.