Project no. FP6-507752

MUSCLE

Network of Excellence
Multimedia Understanding through Semantics, Computation and Learning

DN 4.3.1: Text Analysis Tools

Due date of deliverable: 30.11.2006
Actual submission date: 19.02.2007

Start date of project: 1 March 2004 Duration: 48 months

Deliverable Type: PU
Number: DN4.3.1
Nature: P
Task: WP4

Name of responsible: Andreas Rauber, TU Vienna-IFS

Revision 1.0

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Text Analysis Tools

Within the MUSCLE Network of Excellence on multimedia understanding, datamining and machine learning researchers have developed a range of tools for text analysis, text annotation, Natural Language Processing text classification and semantic indexing. This deliverable (DN 4.3) of WP4 represents an inventory of current text analysis tools:

Part-of-Speech Tagger, Spatial Query Extractor

Bilkent University, Ugur Gudukbay, Ozgur Ulusoy

Bilkent University has extended the freely available English part-of-speech MontyTagger to further recognize spatial relationships is search queries. They have integrated this function in their video search engine BilVideo.

In their extension their tool can extract from natural language queries the following spatial relations:

- topological relations that describe order in 2D space (disjoint, touch, inside, contain, overlap, cover, coveredby)
- directional relations that describe the neighborhood of objects (directions: north, south, east, west, northeast, northwest, southeast, southwest and neighborhood: left, right, below, above)
- 3D relations that describe object positions in 3D space (infrontof, strictlyinfrontof, behind, strictlybehind, touchfrombehind, touchedfrombehind, samelevel)

For example, given a video search query such as "Retrieve segments where James Kelly is to the right of his assistant", this system will extract the spatial relation right(JamesKelly, assitant) that can sent to a further query processing engine.

The tool can be experimented through the Web as a Web client. There is a

demo video for query processing purposes together with some examples, queries and tutorials.

Website: http://www.cs.bilkent.edu.tr/~bilmdg/bilvideo

Presentation in MUSCLE BSCW

Updatable Probabilistic Latent Semantic Indexing

AUTH, Constantine Kotropoulos

Probabilistic latent semantic indexing (PLSI) is a semantic space reduction method that folds documents and the concepts that appear in them into a smaller dimensioned semantic space which can then be used to index and classify new documents. Building a reduced semantic space is time consuming, order O(N^3). AUTH has implemented a new method for updating PLSI when new documents arrive. The new method incrementally adds the words of any new document in the term-document matrix and derives the updating equations for the probability of terms given the class (i.e. latent) variables, as well those of documents given the latent variables. This quick updating would be useful in a web crawler where the term-document matrix must be refreshed very often.

Website: http://www.aiia.csd.auth.gr/EN/

Presentation in MUSCLE BSCW

Natural Language Processing Tools, OWL version of WordNet

CEA LIST, Olivier Mesnard

The CEA has a suite of natural language processing tools for the following languages: English, French, Italian, Spanish, German, Chinese, and Arabic. Alpha versions exist for Hungarian, Japanese, and Russian. They perform the following functions:

- language identification and text encoding identification
- UNICODE translation of codesets
- tokenization, dividing input stream into individual words
- morphological analysis (recognizing conjugated word forms and providing their normalized dictinoary-entry forms)
- part-of-speech tagging (choosing the grammatical function of each word in a text)
- entity recognition (identifying people, organizations, place names, products, money, time)

- dependency extraction (recognizing subject-verb-object relations, and modifier relations)

These function allow the transformation of raw text into symbolic knowledge that can be used to describe, index and access textual information, such as that associated with image captions, or in raw descriptions.

The CEA has also developed an OWL ontology version of the WordNet lexical hierarchy. A reduced version of this ontology restricted to all the picturable objects in WordNet (30 Mbytes) is available from the CEA LIST.

Website of commercial version of these tools: http://www.new-phenix.com

Presentation in MUSCLE BSCW

SOMLib Java Package

TU-WIEN - IFS, Andreas Rauber

TU Vienna - IFS has developed a software for analyzing text documents and organizing them on a Self Organizing Map (SOM) - a representation of a reduced semantic dimension, bringing similar documents, or objects closer together on a two or three dimensional plane. The SOMLib Java Package is a collection of JAVA programs that can be used to create SOMLib library systems for organizing text collections. The package includes

- Feature Extraction
- Feature space pruning
- Feature vector creation
- Feature vector normalization
- SOM training
- SOM Labeling
- libViewer template generation.

Website: http://www.ifs.tuwien.ac.at/~andi/somlib/download/index.html

Quick Reference:
http://www.ifs.tuwien.ac.at/~andi/somlib/download/java_package/

Semi-automated Corpus Annotator

CNRS, Fathi Debili

The CNRS has produced an interface for semi-automatic annotation of large quantities of text. Originally developed for the Arabic language, it is currently being used to tag English and French texts, and can be adapted to any language. Example of its applications are creating training set for part-of-speech taggers, or creating tagged corpora for other types of linguistic research.

Presentation in MUSCLE BSCW

UTIA Text Classifier

UTIA, Jana Novovicova

Text categorization (also known as text classification) is the task of automatically sorting a set of documents into predefined classes based on its contents. Document classification is needed in many applications including e-mail filtering, mail routing, spam filtering, news monitoring, selective dissemination of information to information consumers, and automated indexing of scientific articles. The Prague-based team of UTIA has produced a method for text classification using Oscillating Search which, unlike traditional approaches, evaluates feature groups instead of individuals and which improves classification accuracy in experiments.

Paper describing the work:
http://staff.utia.cas.cz/novovic/files/CIARP06_NSP.pdf