Language Processing Chains in ATLAS

Anelia Belogay[1], Dan Cristea[2], Eugen Ignat[2], Diman Karagiozov[1], Svetla Koeva[3], Maciej Ogrodniczuk[4], Adam Przepiórkowski[4], Polivios Raxis[5], Cristina Vertan[6]

[1]Tetracom Interactive Solutions, [2]Alexandru Ioan Cuza University,
[3]Institute for Bulgarian Language, Bulgarian Academy of Sciences,
[4]Institute of Computer Science, Polish Academy of Sciences, [5]Atlantis Consulting, [6]Hamburg University

Abstract

The aim of this demo is to present the platform for multilingual content management developed by the ATLAS project (Applied Technology for Language-Aided CMS, a European CIP ICT-PSP project, Grant Agreement 250467, `http://www.atlasproject.eu`).

Keywords: ATLAS, UIMA, linguistic processing chains, language resources and tools, natural language processing, NLP

1 Introduction

The ATLAS project unifies complex language processing into a common software platform and provides several sample solutions to demonstrate its capabilities. Mechanisms such as automatic annotation of important words, phrases and names, text summarization, categorization and computer-aided translation intend to facilitate the process of manipulating heterogeneous multilingual content in complex IT solutions.

Languages in scope of the project and represented by the consortium partners are Bulgarian, German, Greek, English, Polish and Romanian.

2 i-Publisher

The first illustration of ATLAS potential is i-Publisher (`http://i-publisher.atlasproject.eu`) – a powerful Web-based application for creating, running and managing small and enterprise content-driven Web sites. It offers e.g.:

- dynamic content propagation,
- multilingual versioning,
- retrieval of similar documents in different languages,
- dynamic interlinking of documents based on automatically extracted content,
- point-and-click Web-based user interface for building reusable content-driven Web sites.

3 UIMA-based language processing chains

Language processing in ATLAS is carried out by means of language processing chains integrating several existing tools (mainly open source) into a scalable and distributable annotation framework – Unstructured Information Management Architecture (UIMA).

Each processing chain splits texts into sentences and tokens, provides information about POS tags and lemmata, annotates noun phrases and named entities (date, time, money and percentage expressions, persons, organizations, locations) as well as provides additional information used on further stages of document processing (e.g. synset ID for summarization or normalized NE versions for visualization).

4 i-Librarian and EUDocLib

i-Librarian (`www.i-librarian.eu`) and EUDocLib (`http://eudoclib.atlasproject.eu`) are two Web sites demonstrating ATLAS-based linguistic processing. The first one is a free online library that assists authors, students, young researchers, scholars, librarians and executives to easily create, organise and publish various types of documents; the second is a publicly accessible repository of EU documents from the EUR-LEX collection which provides easier access to relevant documents in the user's language.

Both sites are capable of:

- processing documents in supported languages in order to automatically categorize, summarize and annotate content with important noun phrases and named entities,
- providing better content navigation (such as list of similar documents) based on interlinked text annotations,
- providing machine-translated excerpts of documents and using them for document categorization and clustering.

5 The „Polish EUDocLib" demo

The Polish variant of EUDocLib is a language processing chain-powered Web site offering search and browsing of around 1000 acts of Parliament automatically annotated with a set of ATLAS-integrated tools for Polish:

- Morfeusz – a morphological analyser for Polish,
- Pantera – a rule-based Brill tagger of Polish,
- Spejd – an engine for shallow parsing using cascade grammars,
- plNER tool – a statistical CRF-based named entity recogniser.

Basing on the annotations, the Web application provides for each document a set of recognized named entities, important noun phrases (with their weights) and a list of similar documents. For presentation, base forms of multiword units are generated and manually assigned categories are used.