Ontea: Pattern based Semantic Annotation Platform

Michal Laclavík, Martin Šeleng, Ladislav Hluchý

laclavik.ui@savba.sk, http://ontea.sourceforge.net/, http://ikt.ui.sav.sk/

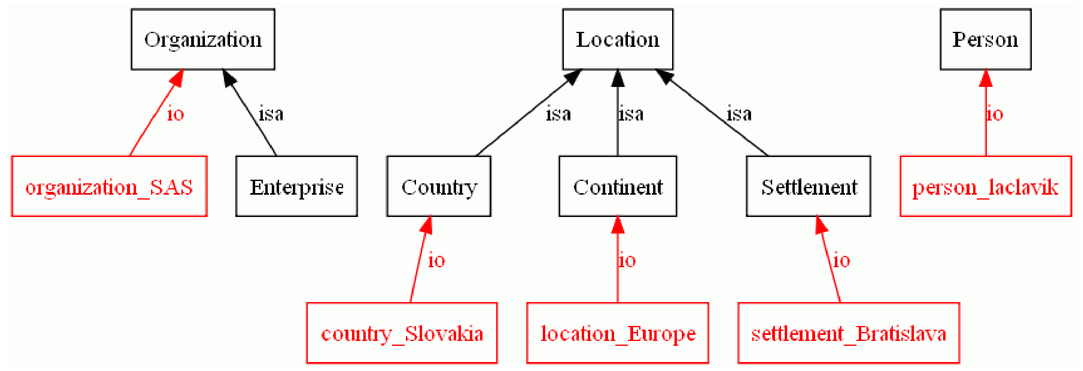
Institute of Informatics, Slovak Academy of Science, Dúbravská cesta 9, 845 07 Bratislava, Slovakia

Motivation

To create semantic meta data from texts or documents

Approach

Even unstructured text documents contain patterns, which can be used to extract various objects from text. Ontea is platform which exploits pattern approach to find or create key – value pairs from text. Pairs can be further processed and transformed to ontology instances or other metadata type.



Pattern Types

Regular Expressions – current implementation

Further possibilities:

- Integration of existing pattern solutions such as C-PANKOW or GATE
- Advanced patterns used in Information Extraction and Information Retrieval
- Keyword based
- XPath – for XML or HTML documents

Example	Text	Patterns – Regular Expressions
1 English	Michal Laclavik works for Slovak Academy of Sciences located in Bratislava, the capital of Slovakia.	\\b(\\p{Lu}[a-z]+ +\\p{Lu}[a-z]+|\\p{Lu}[a-z]+)\\b
2 Slovak	Automobilka KIA sa rozhodla investovať pri Žiline, kde vybudovala svoju prvú továreň v Európe. Kia Motors Slovakia, s.r.o. P.O.Box 2, 01301 Teplička nad Váhom Slovakia	\\b([\\p{Lu}][-&\\p{L}]+[ ]*[-&\\p{L}]*[ ]*[-&\\p{L}]*)[,]+s\\.r\\.o\\. *Organization* \\b[0-9]{3}[ ]*[0-9]{2} +(\\p{Lu}[^\\s,.]+[ ]*[^0-9\\s,.]*[ ]*[^0-9\\s,.]*)[ ]*[0-9\\n,]+ *Settlement* (v|pri) +(\\p{Lu}\\p{L}+) *Location*

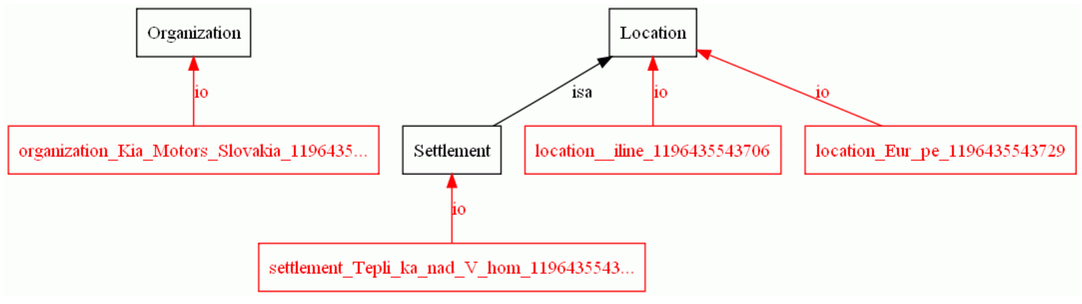
#	Method	Annotation Results
1	Ontea	Michal Laclavik *Person* Slovak Academy *Organization* Bratislava *Settlement* Slovakia *Country*
2	Ontea create	Žiline *Location* Európe *Location* Kia Motors Slovakia *Organization* Teplička nad Váhom *Settlement*
2	Ontea create, lemmatization	Žilina *Location* Európa *Continent* Kia Motors Slovakia *Organization* Teplička nad Váh *Settlement*

Used Technology

- Regular Expressions
- RDF, OWL
- Java

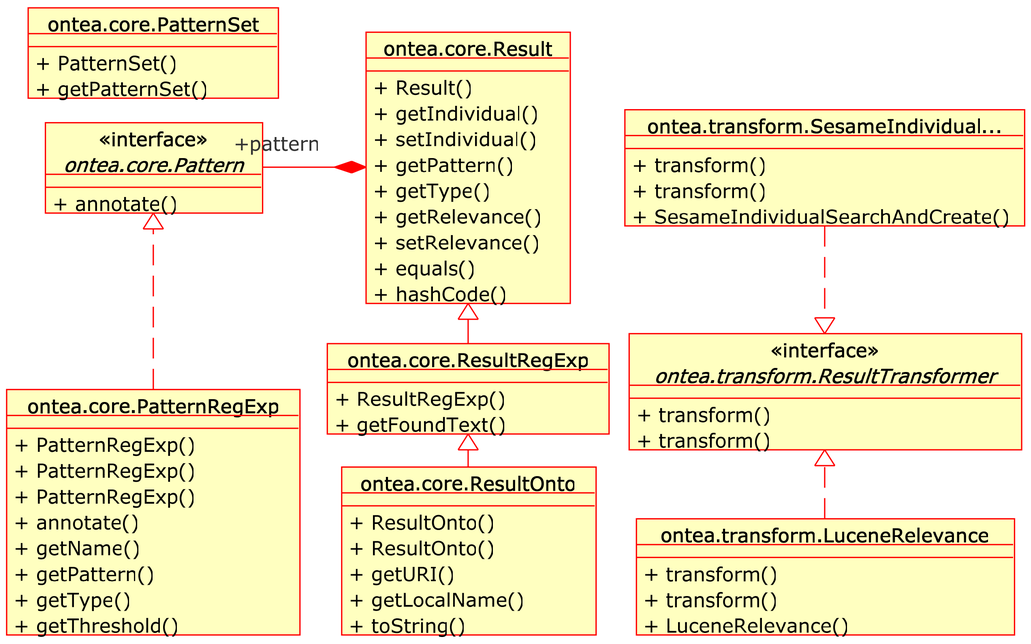
Used Tools

- Lucene
- Hadoop
- Sesame, Jena



Features

- Identification of concept instances from the ontology
- Automatic population of ontologies with instances
- Identifying relevance, when creating instances using information retrieval techniques
- Large scale semantic annotation of documents or texts using Google's MapReduce architecture.



Architecture

- **ontea.core.Pattern:** Interface to adopt different pattern annotation techniques. Current implementations include regular expression pattern matching
- **onetea.core.Result:** Class representing results of pattern annotation – instances of defined type
- **ontea.transform.ResultTransformer:** Interface transforming results of annotation to different type or quality of results e.g. concrete ontology mapping, knowledge base implementation or result quality checking

Integration with External Tools

- **Nalit:** Text Language Identification
- **Morphonary:** Lemmatization of Slovak developed at UPJŠ Košice
- **Lucene:** New instance relevance identification.
- **Hadoop:** Large scale semantic annotation using MapReduce Architecture
- **Sesame, Jena:** Transformation of found key - value pairs into RDFS or OWL instances in Sesame or Jena API

Acknowledgment: This work is partially supported by NAZOU SPVV 1025/2004, Commius FP7-213876, SEMCO-WS APVV-0391-06 and VEGA 2/7098/27