

Semantic Annotation based on Regular Expressions

Laclavik M., Gatial E., Balogh Z., Habala O., Nguyen G., and Hluchy L.

Institute of Informatics, Slovak Academy of Sciences
Dubravska cesta 9, 845 07 Bratislava, Slovakia
laclavik.ui@savba.sk

Abstract. In this paper we describe a solution for creation of ontological metadata from text documents using a semantic annotation. A lot of solutions are known for creation of ontological metadata from text documents. Our solution is based on a regular expression. The solution uses a regular expression - ontology pairs to detect ontology concepts and annotate the document with them. The solution requires having ontology describing the problem domain. The paper explains two possible applications for using such technique.

1 Pattern Ontology Model and Annotation Algorithm

The instances of the Pattern class are used to define and identify relations between text and domain ontology, where the pattern property contains the regular expression which describes textual representation of the ontology element. The examined text is processed with the regular expression for every pattern and when it is found: the detected ontology element of hasClass or hasInstance represents text in the chosen problem domain. Moreover, when the hasClass property exists in the Pattern, the RDQL query is constructed and processed to find the individuals that match the condition:

- individual is the class of hasClass
- a property of individual contains the matched word

Figure 1 shows results of the annotation based on patterns from Figure 2. We will describe an algorithm by an illustration example from the Znalosti project where a job offer text is analyzed and annotated by regular expressions patterns (Figure 2). System will find related ontology elements from domain ontology. In this example the job offer location - New York and USA are identified by a regular expression $([A - Za - z]^+)$ a $([-A - Za - z0 - 9]^+)$ + $[\]$ + $[-A - Za - z0 - 9]^+$, because individual *locNY* has the property title *NewYork*, *locUS* has the property title *USA*. Similarly other ontology elements are detected. Some regular expressions search for ontology individuals, other ontology classes and others such as *pattFullTime* (Figure 2) annotate a job offer by a concrete individual *jtPermanent* if expression $(Full[-]Time)$ is found. Systems detect ontology elements based on domain ontology. In this example it is ontology of job offers.

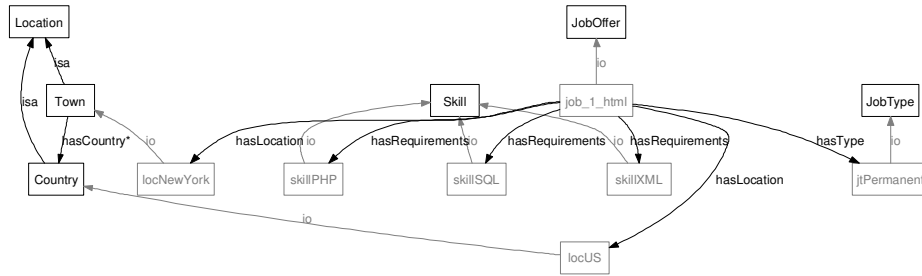


Fig. 1. Job Offer Individual with its properties detected by regular expressions showed in Figure 2 from Znalosti job offer application

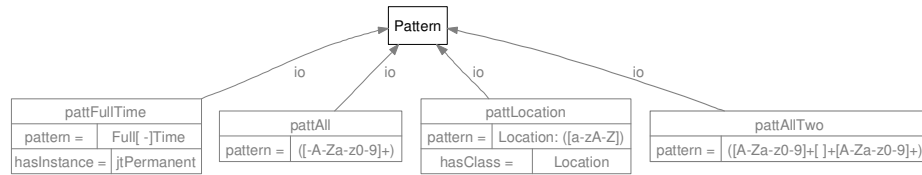


Fig. 2. Fragment of pattern ontology from Znalosti job offer application

2 Conclusion

The described solution is used in K-Wf Grid [2] and the Znalosti project to detect relevant structured knowledge described by a domain specific ontology model in the unstructured text. The main difference between existing annotation solutions such as Anotea [1] is detection of ontology elements from existing domain ontology, while other annotation solutions try to create such ontology. In the Znalosti project our solution is used to detect structured information about job offers. In the K-Wf Grid project our solution is used to detect a user context/problem from the text description as well as annotate user knowledge entered in a form of text notes[3]. The main disadvantage is that regular expression patterns need to be manually created. In our future work we will focus on semi-automatic pattern creation.

Acknowledgments: This work is supported by the project K-Wf Grid EU RTD IST FP6-511385, VEGA No. 2/3132/23, APVT-51-024604 and SPVV 1025/2004.

References

- [1] Annotea Project, <http://www.w3.org/2001/Annotea/>, (2001)
- [2] K-Wf Grid Consortium: K-Wf Grid Project Website, <http://www.kwfgird.net/>, (2005)
- [3] Laclavik M., Gatjal E., Balogh Z., Habala O., Nguyen G., Hluchy L.: Experience Management Based on Text Notes (EMBET), Challenges Conference, 19 - 21 October 2005, Ljubljana, Slovenia , (2005)