Ontea: Platforma pre sémantickú anotáciu založenú na vzoroch*

Michal Laclavík, Martin Šeleng, Ladislav Hluchý

Ústav informatiky, Slovenská akadémia vied,
Dúbravská cesta 9, 845 07 Bratislava
laclavik.ui@savba.sk

Abstrakt. Sémantický web sa snaží aby webu nerozumel len človek ale aj stroje. Problémom je nedostatok formalizovaných metadát na webe, ktoré sa dajú vytvoriť ručne ako aj automatickými riešeniami pre takzvané značkovanie alebo sémantickú anotáciu. Automatické riešenia s najväčšou úspešnosťou anotácie sú založené na rôznych učiacich algoritmoch z oblasti umelej inteligencie. Je ich však ťažké reálne použiť v praxi kvôli neexistencii trénovacích množín a ich nákladnej tvorbe. Cestou na tvorbu automatických riešení pre sémantickú anotáciu sa javia prístupy založené na rozpoznávaní vzorov ako štruktúra dokumentu, jazyk a podobne. Príspevok sa zaoberá popisom platformy Ontea, ktorá umožňuje generovať sémantické metadáta na základe vzorov.

Kľúčové slová: sémantická anotácia, extrakcia informácií, metadáta

1 Úvod

Semi-automatické riešenia pre sémantickú anotáciu vychádzajú aj z oblasti extrakcie informácií „Information Extraction" – IE [1], kde boli konferenciami MUC definované úlohy pre extrakciu informácií ako napríklad rozpoznávanie názvoslovných entít „Named entity recognition" – NE. Rozpoznávanie NE ako aj ďalšie úlohy z MUC konferencií, sú podobné ako sémantická anotácia.
Semi automatické anotačné riešenia je možné rozdeliť na dve skupiny podľa výsledkov anotácie ktoré produkujú a to:

- Identifikácia inštancií zo znalostnej bázy
- Automatické napĺňanie ontológií inštanciami

Semi-automatické metódy sa zameriavajú najme na identifikáciu inštancií alebo vytváranie sémantických metadát pre ich ďalšie počítačové spracovanie. Podrobnejší prehľad metód pre sémantickú anotáciu je uvádzame v [2] [3].

2 Ontea

Metóda použitá v platforme Ontea [2] [3] je porovnateľná najmä s metódami C-PANKOW, KIM[2] a SemTag. Pracuje nad textom príslušným k nejakej aplikačnej doméne, ktorá je popísaná doménovým ontologickým modelom a používa regulárne výrazy na hľadanie vzťahov medzi textom a sémantickým modelom. Okrem toho že

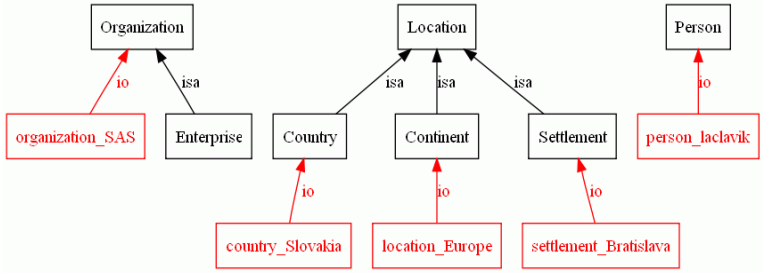
* This work is supported by projects NAZOU SPVV 1025/2004, RAPORT APVT-51-024604, SEMCO-WS APVV-0391-06, VEGA 2/7098/27.
[2] http://www.ontotext.com/kim/semanticannotation.html

obsahuje implementáciu vzorov nad regulárnymi výrazmi, jej architektúra bola vytvorená tak, aby bolo možné jednoducho implementovať ďalšie metódy založené na vzoroch ako wrapre, riešenia využívajúce štruktúru dokumentu, jazykové vzory ako u C-PANKOW a ďalšie.

2.1 Príklad použitia

Nástroj Ontea môže byť použitý v troch rôznych scenároch podľa potreby aplikácie:

- *Ontea*: hľadanie relevantných inštancií v korporatívnej pamäti
- *Ontea creation*: vytváranie nových inštancií objektov nájdených v texte
- *Ontea IR*: ako v predchádzajúcom prípade ale s využitím nástroja RTFS alebo Lucene na zistenie relevancie takto vytvorenej inštancie



Obr. 1. Jednoduchá ontológia použitá v príkladoch

Na demonštráciu použijeme nasledovné texty uvedené v tabuľke 1 a ontológiu s niekoľkými inštanciami (obr. 1). Jeden text v slovenčine a jeden v angličtine. V tabuľke sú uvedené aj príklady použitých vzorov založených na regulárnych výrazoch. Pre angličtinu bol použitý len jeden regulárny výraz na vyhľadanie jedného alebo dvoch slov začínajúcich veľkým písmenom.

Tabuľka 1. Text príkladov

Príklad	Text	Vzory – regulárne výrazy
1	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	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*

V prípade anglického textu išlo iba o vyhľadanie inštancií, ktoré sú v znalostnej báze obrázok 1 a prvý riadok tabuľky 2. Na slovenskom texte bol ukázaný prípad keď inštancie nie sú len vyhľadávané v rámci znalostnej bázy ale sú aj vytvárane. Pre slovenčinu sú v príklade použité 3 vzory:

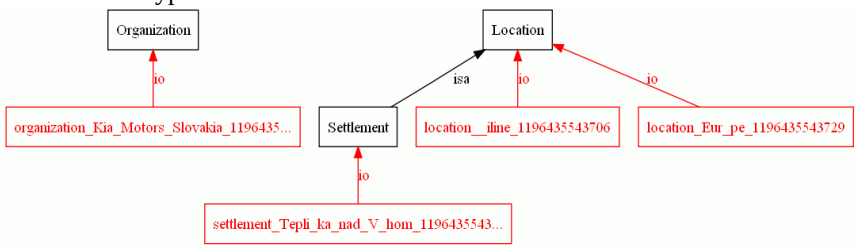
- Hľadanie organizácie podľa „s.r.o.“,
- hľadanie sídla podľa PSČ a
- hľadanie geografickej lokality podľa predložiek v a pri a nasledujúcim slovom s veľkým písmenom.

Výsledky možno vedieť v tabuľke 2 v druhom a treťom riadku. Inštancie vytvorené z druhého riadku sú aj na obr. 2.

Tabuľka 2. Výsledok anotácie

Príklad	Metóda	Výsledky anotácie
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, lematizácia	Žilina *Location* Európa *Continent* Kia Motors Slovakia *Organization* Teplička nad Váh *Settlement*

Pri výsledkoch tretieho riadku tabuľky 2 bola použitá lematizácia nájdených textov čo umožňuje korektnejšie vytvorenie inštancií v prvom páde v prípadoch „Žilina“ a „Európa“ zároveň sa však pokazila lokalita „Teplička nad Váhom“ kde vzniklo „Teplička nad Váh“. Tu by bolo možné použiť lematizáciu iba na niektoré vzory a nie na všetky. Takisto si v tomto prípade môžeme všimnúť že „Európa“ už nie je typu *Location* ale *Continent* pretože algoritmus ju našiel v znalostnej báze pričom bolo použité odvodzovanie keďže *Continent* je subclass *Location*. Pri vytváraní sa teda algoritmus najprv pozerá do znalostnej bázy či už neexistuje inštancia takého typu alebo odvodená.



Obr. 2. vytvorené inštancie na slovenskom texte.

3 Architektúra

Základnými stavebnými prvkami nástroja sú nasledovné Java rozhrania a objekty ktoré sú rozširované a implementované:

- *ontea.core.Pattern*: interface pre adaptáciu na rôzne techniky vyhľadávania vzorov. V súčasnosti je implementované hľadanie vzorov pomocou regulárnych výrazov *PatternRegExp*.
- *onetea.core.Result*: class reprezentujúci výsledok anotácie pomocou vzorov – teda inštancie. Jeho rozšírenia sú rôzne typy inštancií podľa implementácie v ontológii (jena, sesame) alebo ako páry hodnoty a typu.
- *ontea.transform.ResultTransformer*: interface ktoré po na implementovaní obsahuje rôzne typy transformácie medzi výsledkami anotácie. Teda môže transformovať množiny výsledkov a zapojiť do transformácií rôzne scenáre

anotácie napríklad: relevanciu, lematizáciu výsledkov, transformáciu do ontologickej bázy založenej na Sesame alebo Jena.

3.1 Integrácia

Uvedené príklady ilustrujú použitie externých metód na Lematizáciu, prácu s ontologickým modelom a ďalšie. Metóda Ontea môže byť integrovaná s nasledujúcimi softvérmi, čo bolo aj implementačne overené:

- *Nalit*: Identifikácia jazyka.[5].
- *Tvaroslovnik*: Lematizácia. [4].
- *Lucene*: relevancie vytvorených inštancií napr. pomocou Lucene [3].
- *Sesame*: transformácia entít do ontologického API ako Sesame alebo Jena.

4 Záver

V článku v krátkosti opisujeme súčasný stav problematiky automatickej sémantickej anotácie ako aj platformu Ontea , ktorá je založená na vyhľadávaní vzorov. Okrem sľubnej úspešnosti metódy opísanej v [2] a [3] je metóda zaujímavá aj samotnou architektúrou, ktorá umožňuje jednoduchú integráciu s inými metódami ako aj rozšírenia. Metóda je opísaná na jednoduchých príkladoch na ktorých je vidieť funkcionalitu a použitie metódy.

V ďalšej práci by sme chceli metódu sprístupniť pod Open Source licenciou ako aj prispieť k jednoduchšej tvorbe vzorov a možnosti anotovať veľké kolekcie textov.

Literatúra

1. Cunningham, H. (2005). Information Extraction, Automatic. Encyclopedia of Language and Linguistics, 2nd Edition
2. Laclavik M., Seleng M., Gatial E., Balogh Z., Hluchy L.: Ontology based Text Annotation OnTeA; Information Modelling and Knowledge Bases XVIII. IOS Press, Frontiers in AI, Vol. 154, ISBN 978-1-58603-710-9, ISSN 0922-6389, (2007) 311–315
3. Laclavik M., Ciglan M., Seleng M., Hluchy L.: Ontea: Empowering Automatic Semantic Annotation in Grid; to appear in proceedings of PPAM 07, Springer-Verlag
4. Stanislav Krajči, et al.: Hľadanie základného tvaru slovenského slova na základe spoločného konca slov; In: WIKT 2006; ISBN 978-80-969202-5-9; March 2007
5. Peter Vojtek, Mária Bieliková (2007), Comparing Natural Language Identification Methods based on Markov Processes. In: Slovko 2007, Bratislava

Annotation:

Ontea: Platform of Pattern based Semantic Annotation

Web documents are structured but their structure is understandable only for humans, which is the major problem of the Semantic Web. Semantic Web can be exploited only when computer understandable metadata will reach critical mass. Semantic metadata can be created manually or by automated annotation or tagging tools. Automated semantic annotation tools with best results are built on different machine learning algorithms which require training sets. Other approach is to use pattern based semantic annotation. In this paper we describe Ontea platform for pattern based semantic annotation.