SysNERV

System for Named-Entity Recognition and Visualization

Vedoucí projektu: doc. Ing. Zdeněk Žabokrtský, Ph.D. (zabokrtsky@ufal.mff.cuni.cz)

Počet řešitelů: 5 až 6

Termín dokončení: červen 2013

Rozpoznávání pojmenovaných entit

Cílem rozpoznávání pojmenovaných entit (Named-Entity Recognition, NER) je identifikace vlastních jmen v textu a jejich zařazení do předem stanovených kategorií, jako jsou například geografické názvy, jména osob, organizací apod. Tato úloha je motivovaná potřebami různých aplikací zpracování přirozeného jazyka (Natural Language Processing, NLP), například v oblastech získávání informací nebo strojového překladu.

Existující řešení

Jedním z dosud publikovaných řešení tohoto úkolu pro češtinu je rozpoznávač vyvinutý na MFF-UK[1], implementovaný jako modul pro systém TectoMT. Tento rozpoznávač klasifikuje pojmenované entity do několika desítek hierarchicky uspořádaných typů. Podobně jako u většiny jiných úkolů pro NLP je výhodné při vývoji použít anotovaná data, zejména pro trénování a následnou evaluaci. Zde byl jako trénovací data použit ručně anotovaný vzorek z Českého národního korpusu o velikosti 6,000 vět.

Pro angličtinu existuje mj. systém Stanford NER[2], naprogramovaný v jazyku Java. Stanford NER využívá model Conditional Random Fields. Na rozdíl od svého českého protějšku rozpoznává jen čtyři základní typy pojmenovaných entit.

Hlavní cíl projektu

Hlavním cílem projektu je převést laboratorní prototypy rozpoznávačů pojmenovaných entit pro češtinu a angličtinu do podoby robustních, „průmyslově“ použitelných aplikací s pohodlným uživatelským rozhraním. Pro zájemce o automatické rozpoznávání pojmenovaných entit budou vytvořena tři rozhraní: (1) webová služba, (2) grafické uživatelské rozhraní ve formě rozšíření do webového prohlížeče, (3) samostatně běžící aplikace.

Webová služba

Funkce implementovaných rozpoznávačů budou uživatelům zpřístupněny jako webová služba běžící na jednom z fakultních serverů. Umožní on-line rozpoznávání pojmenovaných entit pro potřeby dalších aplikací, a to bez lokální instalace jakéhokoliv dodatečného softwaru.

Rozpoznávač v prostředí webového prohlížeče

Citelným nedostatkem dosavadních řešení je neexistence grafického rozhraní. Chtěli bychom se proto zaměřit na vytvoření přívětivého grafického rozhraní. Uživateli bude vrácen vstupní text s přehledně označenými rozpoznanými pojmenovanými entitami.

Samostatně běžící aplikace

Rozpoznávač pojmenovaných entit ve formě lokálně běžící aplikace je potřebný pro uživatele, kteří pracují s citlivými informacemi (např. Policie ČR). Lokálně nainstalovaná aplikace nebude vyžadovat připojení k Internetu.

1 Ševčíková Magda, Žabokrtský Zdeněk, Krůza Oldřich: Named Entities in Czech: Annotating Data and Developing NE Tagger. In: Lecture Notes in Computer Science, Vol. 4629, No. XVII, Proceedings of the 10th International Conference on Text, Speech and Dialogue, Copyright © Springer, Berlin / Heidelberg, ISBN 978-3-540-74627-0, ISSN 0302-9743, pp. 188-195, 2007

2 Stanford Named Entity Recognizer, http://nlp.stanford.edu/software/CSF-NER.shtml

Další cíle projektu

Integrace českého a anglického NER do systému Treex

Jelikož je systém TectoMT, ve kterém byl rozpoznávač českých pojmenovaných entit původně implementován, v současnosti nahrazován novějším prostředím Treex, bude nutné prototyp českého NER z větší části reimplementovat. Zároveň vytvoříme i modul pro angličtinu. Oba moduly budou implementovány v jazyce Perl, který je hlavním jazykem používaným v prostředí Treex.

Rozšíření dat pro NER v češtině

Existující rozpoznávač pro češtinu využívá ručně anotovaná data, která ovšem nelze považovat za reprezentativní, neboť obsahují pouze věty s alespoň jednou pojmenovanou entitou. Aby výsledný natrénovaný rozpoznávač nenacházel příliš mnoho falešně pozitivních výskytů, bude nutné tato data lépe vyvážit (tj. rozšířit je o další věty).

Závěr

Výsledkem naší práce by neměla být pouze nová implementace rozpoznávačů pojmenovaných entit, ale také přehledné a uživatelsky přívětivé grafické rozhraní využívající nejmodernějších metod v oblasti tvorby webových stránek.