Minireferát: From Hyperlinks to Semantic Web properties using Open Knowledge Extraction (4IZ440)

Vypracoval: Marek Bečica, 2016

Pro minireferát jsem si vybral článek From Hyperlinks to Semantic Web properties using Open Knowledge Extraction. Článek se zaměřuje na to, že v současné době existuje potřeba na automatické porozumění strukturovaného, semi-strukturovaného a hlavně přirozeného textu počítačem. To se v současné době snaží vyřešit mnoho technologií Knowledge Extraction (extrakce znalostí), které dokáží ve strukturovaném a semi-strukturovaném textu rozpoznat entity a vztahy mezi nimi. Problém ale nastává u přirozeného textu, kde dosud neexistovala dostatečně účinná a univerzální metoda, jak z textu vyextrahovat entity a vztahy mezi nimi. Existující metody byly chybové (ať už v nalezení nesprávných nebo nenalezení správných vztahů) a nedokázaly pracovat s různými druhy textu. Jak článek upozorňuje, styl psaní textu např. pro Wikipedii se velmi liší od stylu psaní blogu a ten se velmi liší od psaní příspěvku na Twitteru. Metoda, která tak zvládne relativně dobře vyhledat vztahy v jednom typu textu úplně pohoří v jiném.

Článek se zaměřuje na to, jak by měla vypadat metoda, která si dovede poradit s oběma problémy. Dále je potřeba aby metoda dokázala pracovat nejen se vztahy v rámci jednoho textu, ale aby uměla propojit spojitosti v textech pomocí hyperlinků. Tyto vztahy by poté měly být formalizovány pomocí RDF trojic, tak aby objekt, subjekt i predikát měli URI a jednalo se o formalizované vlastnosti RDF/OWL, aby bylo možné je znovupoužít a napojit na stávající slovníky. Autoři článku se zaměřili na to, jak obohatit Open Information Extraction (OIE) o abstrakci. OIE je totiž zaměřeno primárně na extrakci, tedy na rozdělení textu na smysluplné fragmenty, které by se daly použít pro zjištění vztahů mezi nimi. Tuto novou metodu nazývají Open Knowledge Extraction (OKE), která k OIE přidává formální sémantickou reprezentaci textu a vytváří výstup, který je v souladu s principy a požadavky sémantického webu. Tuto metodu OKE poté implementují pomocí systému, který nazývají Legalo. Ten dokáže zpracovat anglický text na sadu RDF trojic, které odpovídají relevantním a faktickým vztahům z daného textu a jsou pojmenovány co nejblíže tomu, jak by je pojmenoval člověk. Kvalitu extrakce znalostí z textu poté ověřují na textech ze stránek z Wikipedie a porovnávají jej s výsledky získanými od lidí pomocí crowdsourcingu. Ukazuje se, že Legalo podává nad očekávání dobré výsledky.

Metoda OKE by měla fungovat bez zásahu člověka, v otevřené doméně a umět zpracovat všechny druhy textů. Toho má dosáhnout pomocí několika kroků. V prvním kroku nalezne v textu pár entit a zhodnotí, zda mezi nimi neexistuje nějaký relevantní vztah (tzn. takový, který podporuje dostatek rozpoznaných stop v textu). Vztah může být nejen binární, ale i n-ární. Pokud je vztah mezi entitami potvrzen, v dalším kroku je vygenerován predikát pro tento vztah s takovým názvem, který by se měl co nejvíce podobat tomu, jak by takový vztah definoval člověk pro Linked Data slovník. V posledním kroku je poté tento vztah formalizován jako OWL objekt s TBox axiomem (na konceptuální úrovni) a ABox axiomem (na faktické úrovni).

Tyto tři kroky jsou v systému Legalo realizovány pomocí šesti základních kroků. První je abstraktní krok, při kterém je vytvořena vnitřní formální reprezentace věty textu do grafu. V dalším kroku jsou identifikovány relevantní vztahy mezi dvěma entitami a daný vztah musí být potvrzen v závislosti na obsahu dané věty. V dalším extraktivním kroku jsou nalezeny relevantní názvy pro predikát. V následujícím abstraktním kroku je vygenerován popisek predikátu (v souvislosti s tím Legalo spoléhá

na další zdroje, jako DBpedia, Schema.org, WiBi a VerbNet). Následuje další abstraktní krok, ve kterém je vytvořena formální definice predikátu v rozsahu lingvistického důkazu a formální reprezentace. V posledním kroku je úprava názvů a definic, tak aby odpovídala existujícím vlastnostem sémantického webu (k tomu používá tři sémantické zdroje: službu Watson, agregátor Linked Open Vocabularies LOV a systém Never-Ending Language Learning NELL).

Systém Legalo je založen na dvou pracovních hypotézách:

- Legalo je schopen posoudit, zda v dané větě existuje relevantní vztah mezi dvěma entitami v závislosti na zbytku věty.
- Legalo je schopný vygenerovat predikát pro relevantní vztah mezi dvěma entitami vyjádřenými v dané větě. Popisek predikátu vygenerovaný systémem Legalo je velmi podobný popisku, který by člověk definoval pro slovník Linked Data pro daný výskyt v textu.

Pro zhodnocení těchto dvou hypotéz byl použit korpus vytvořený Google research, který obsahoval pět datasetů dedikovaných pro extrakci specifických vztahů – místo narození, navštěvování nebo absolvování nějaké instituce, místo smrti, datum smrti a stupeň vzdělání. Datasety obsahují úryvky z Wikipedie, které obsahují více než jednu větu. Toto zhodnocení tedy ověřuje pouze schopnost práce s textem stylu Wikipedia, zatímco Legalo by měl být schopný pracovat s různými texty (blogy, Twitter), může se značně lišit výkonnost mezi nimi. Pro ověření výkonnosti byla použita crowdsourcingová platforma Crowdflower, kde byly osoby nejprve otestovány, zda jsou vhodné pro daný výzkum. Při porovnání výsledků vyšlo najevo, že Legalo dosáhl překvapivě vysokého výsledku v rámci první hypotézy – F-míra = 0,92. Pro druhou hypotézu Legalo vykazoval podobnost s názvy vytvořenými člověkem s jistotou 68-76 %, což je také velmi dobrý výsledek.

V době psaní článku probíhaly další experimenty, které mají nalézt nové způsoby úprav vztahů, tak aby přesněji navazovaly na existující ontologie. Dále probíhají další experimenty pro zvýšení škálovatelnosti Legalo na velikost a rozmanitost současného webu. Autoři studie mají také další plány pro budoucnost, kde by chtěli aby, Legalo zvládl pracovat nejen s jednoduchým textem, ale i s textem na několika úrovních (několik textů propojených hyperlinky) a aby dovedl ověřit výsledky pomocí nejnovějších metod.

Bibliografie:

Presutti, Valentina, Nuzzolese, Andrea Giovanni, Consoli, Sergio, Gangemi, Aldo and Reforgiato Recupero, Diego, 2016, From hyperlinks to Semantic Web properties using Open Knowledge Extraction. *Semantic Web*. 2016. Vol. 7, no. 4, p. 351-378. DOI 10.3233/sw-160221. IOS Press