Masarykova univerzita
Fakulta informatiky

Automatické seskupování zpravodajských článků týkajících se stejné události

Diplomová práce

Jiří Vejvoda

Brno, podzim 2013

Prohlášení

Prohlašuji, že tato diplomová práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

Vedoucí práce: doc. RNDr. Petr Sojka, PhD.

Poděkování

Zde bude uvedeno „poděkování“ ...

Obdobně jako poděkování se mohou vysadit shrnutí a klíčová slova pomocí prostředí Thesisacti a ThesisKeyWordsi.

Obsah

Úvod

Text ...

Kapitola 1

Uvedení do problematiky

Denně vychází velké množtví zpravodajských článků, jejich autoři od sebe navzájem často opisují, ten stejný článek je mnohdy rozesílán několika různými RSS kanály. K jedné světové události mohou vycházet články několik dní až týdnů podle závažnosti nebo atraktivity události. Orientace v takovém množství je čím dál náročnější.

Projekt si dává za cíl vytvoření systému pro automatické rozdělení množiny zpravodajských článků na podmnožiny podle událostí, které články popisují. Toto rozdělení má širokou škálu uplatnění, nejdůležitější je přetváření chaotických informací na strukturovanou formu, to mohou velmi ocenit zpravodajské služby, velikosti podmnožin mohou být také dobrým ukazatelem závažnosti události. Při studování nějaké konkrétní události je to snadný přístup ke všem relevantním informacím.

Na toto rozdělení do podmnožin je možné pohlížet také jako na relaci mezi dvojicemi článků na této množině. Dva články jsou v relaci, jedině když oba popisují tutéž událost. Z toho triviálně vyplývá, že tato relace je reflexivní a symetrická. Transitivitu můžeme jednoduše získat metodikou, kdy článek patří do clusteru ve chíli, kdy jeho skóre přesáhne definovanou hranici alespoň s jedním dalším článkem v daném clusteru. Tím získáme toto rozdělení ve formě relace ekvivalence.

Kapitola 2

Použité nástroje

2.1 Klasické NER nástroje

Pro systém jsem se rozhodl využít především nástroje pro rozpoznávání pojmenovaných entit (*NER*[1]). Ty totiž logicky nejvíce odhalují souvislosti mezi jednotlivými zpravodajskými články. Výhoda pojmenovaných entit je také v tom, že nejsou tolik jazykově zatížené jako ostatní slova v textu článku a systém by ve výsledku mohl fungovat na článcích v různých jazycích, to je ovšem podmíněno také tím, že musí být pro všechny jazyky podpora ve zvoleném NER nástroji.

Samotných NER nástrojů je větší množství, co se zpracování angličitny týče, určitě za zmínku stojí Stanford Named Entity Recognizer [3], dále pak Illinois Named Entity Tagger [4], případně ještě modul dostupný z NLTK knihovny [2] pro NLP v Pythonu.

Všechny jmenované nástroje mají tu výhodu, že je možné je nainstalovat a využívat ve specifickém programu relativně jednoduše, zajištění podpory pro další jazyk je většinou možné vytvořením dostatečně velkého označkovaného korpusu a naučením modelu pomocí tohoto korpusu. Nehrozí, že by nástroj během pár let příliš zestárl a celková efektivita programu, kde je využíván, by rapidně klesla. Tyto nástroje jsou ovšem určené především na vyznačení daných entit v textu, samotný text a druh entity neříká nic o tom, co daná entita zastupuje, je potřeba dodatečné zpracování už jen k určení, které entity v jednom textu označují stejnou věc, toto zpracování je netriviální už kvůli používání ruzných zkratek pro názvy organizací, míst (New York, NY, N. Y.), různé skloňování jmen, případně ne/uvádění křestních a prostředních jmen, opět jejich zkracování a pod.

1. Named Entity Recognition

2.2 NER nástroje s desambiguací entit

Tyto problémy se snaží řešit modernější nástroje, které jsou propojené s různými znalostními bázemi. Mezi takové nástroje patří například AlchemyAPI [1] od AlchemyAPI, Inc. a OpenCalais od Thomson Reuters. Vytváření a využívání znalostních bází je směr, jakým se aktuálně orientuje hodně aplikací zabývajících se NLP[2], jedná se o další krok k analýze sémantické roviny jazyka. Napojení na takovéto znalostní báze má hlavní výhodu v tom, že je možné pojmenované entity desambiguovat (s určitou pravděpodobností), pokud je pojmenovaná entita nalezena v databázi, uvede nástroj odkaz na tento záznam (URI), který už je v rámci databáze jedinečný. Samozřejmě se může stát, že některé označené entity se nepodaří propojit se záznamy v databázích ať už proto, že pravděpodobnost vyšla příliš malá, nebo záznam v databázi chybí.

2.2.1 Znalostní báze

Rozsah znalostních bází neustále roste a samotné databáze se liší svým zaměřením, ať už tématickým nebo lokálním/globálním a samozřejmě rozsahem. Dalším specifikem je, jak rychle údaje v databázi stárnou, pro naše účely budou nejdůležitější databáze s geografickými názvy, známými osobnostmi, společnostmi a údálostmi. Geografická data netrpí tolik zmíněným stárnutím, kdežto mezi známé osobnosti a společnosti přibývají často zmiňované entity velmi rychle a často se tomu tak děje právě na základě událostí, které popisují aktuální zpravodajské články, které má tento projekt za cíl zpracovávat. Je tu tedy kladen velký důraz na aktuálnost dat ve znalostní bázi, přesněji je nutné aby existovaly záznamy o entitách, které vznikly nebo se staly známými až v poslední době. Znalostní báze je tedy nejlepší využívat přímo online s nejaktuálnějšími záznamy.

Základní stavební jednotkou dat uvnitř databází jsou RDF[3] trojice, ty reprezentují vždy vztah dvou entit a celou databázi je pak možné si nejjednodušeji představit jako orientovaný graf. Definice těchto trojic má spíše abstraktní formu, každá databáze používá svou vlastní implementaci, je však přesně nadefinována jejich serializace a díky tomu je možné propojovat RDF trojice mezi různými databázemi.

Mezi hlavní znalostní báze patří DBpedia vyvíjená Lipskou univerzitou a Svobodnou univerzitou Berlín, první verze byla vydána na začátku

2. Tento trend je možné pozorovat i u napojení Googlem vytvořeného Knowledge Graph na jeho vyhledávač v roce 2012
3. Resource Description Framework

roku 2007. Databáze je vytvářena ze stránek Wikipedia.org, používá se automatická extrakce strukturovaných dat především z tabulek se základními informacemi. Stejně jako klasická wikipedia je i DBpedia vícejazyková, anglická verze pokrývá 3,77 milionu věcí, z toho 2,35 milionu je zařazeno do ontologie, celkově DBpedia sestává z 10,3 milionů věcí ve 111 různých jazycích. DBpedia obsahuje 1,89 miliardy RDF trojic a je tak největší volně dostupnou znalostní bází.

Další rozsáhlá volně dostupné znalostní báze je Freebase založená roku 2007 společností Metaweb[4], ta je na rozdíl od DBpedie nejen automaticky vyextrahovaná z wikipedie, využívá automatickou extrakci i z dalších zdrojů a hlavně je část jejích dat přímo ručně vytvořena uživateli. Freebase obsahuje téměř 40 milionů věcí a přes 337 milionů RDF trojic.

2.2.2 Linked Data

Velké množství ruzných znalostních bází přineslo potřebu propojit je do jednoho celku, to je možné pokud jsou databáze stavěny na nějaké formě RDF trojic, nebo jsou do tohoto formátu alespoň převoditelné. Tak vznikl komunitní projekt Linking Open Data[5] k němu dále Linked Data[6], ty si dávají za cíl propojit entity z různých databází (tzv. SameAs links propojující různé URI v jednotlivých databázích zastupující stejnou věc), takto je popropojované opravdu velké množství různých databází, jak je vidět na grafu v Appendixu, hlavním středem dat je DBpedia, současně ale vznikají souvislejší komponenty na pár místech v grafu podle zájmových oblastí různých databází. Jak je vidět, rozsah takto propojených znalostních bází je opravdu velký a pro účely tohoto projektu je velmi dobře využitelný.

2.3 AlchemyAPI

AlchemyAPI je webová služba (*SaaS API*[7]) stejnojmenné společnosti poskytující širokou škálu nástrojů v oblasti zpracování přirozeného jazyka. Mezi hlavní nástroje patří rozpoznávání pojmenovaných entit a jejich desambiguace pomocí URI z databází z Linked Data, rozpoznávání výroků (typu „Takové kroky rozhodně nemůžeme nechat bez odpovědi“, řekl mluvčí

4. V roce 2010 koupena společností Google, která Freebase použila jako jeden ze základních prvků při tvorbě svého Knowledge Graph.
5. `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`
6. `linkeddata.org`
7. Software as a Service Application Programming Interface

společnosti), rozpoznávání klíčových slov, extrakce konceptů a další.

Kapitola 3

Přístup distributivní sémantiky

Kapitola 4

Funkce příslušnosti ke stejné události

4.1 Použitelné příznaky

4.2 Hlavní složky funkce

4.3 Výsledná funkce

Kapitola 5

Vlastnosti systému

5.1 Škálovatelnost

5.2 Napojení na databázi

5.3 Distribuovatelnost

Kapitola 6

Vyhodnocení systému

Kapitola 7

Závěr

Literatura

[1] Inc. AlchemyAPI. Alchemyapi, 2013.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009.

[3] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, 2005.

[4] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *CoNLL*, 6 2009.