Analýza zpravodajských textů a jejich komentářů napříč jazyky

Josef Steinberger

Katedra informatiky a výpočetní techniky, Nové technologie pro informační společnost, Fakulta aplikovaných věd, Západočeská univerzita v Plzni, Univerzitní 8, 306 14 Plzeň

jstein@kiv.zcu.cz

Abstrakt. Tento příspěvek představuje projekt MediaGist, jehož cílem je vytvoření online systému, který analyzuje a propojuje zpravodajské články a jejich komentáře v pěti jazycích. Umožňuje novinářům detekovat a prozkoumávat zpravodajská témata, která jsou kontroverzně reportována nebo diskutována napříč různými jazyky/zeměmi. Sumarizace a analýza polarity textu jsou dvě hlavní technologie použité v textové analytice. Polarita slouží k výpočtu kontroverze a souhrny pomáhají zkoumat rozdíly.

Klíčová slova: analýza textu, sumarizace, analýza polarity textu.

1 Úvod

Zpravodajské portály publikují tisíce článků každý den v různých jazycích. Vyznat se v takto ohromném množství informací je bez automatických nástrojů nemožné. Existuje řada agregátorů a analyzátorů zpravodajství a každý má své silné stránky. Google News shlukuje články a zobrazuje příběhy podle zájmů čtenáře. IBM Watson News Explorer přináší analytický způsob čtení zpráv skrze vizualizaci pomocí *linked data*. Europe Media Monitor (EMM) shlukuje zpravodajství téměř v reálném čase ve více než padesáti jazycích [1]. Ovšem existuje ještě další hodnotný zdroj informací na zpravodajských portálech: komentáře k článkům, ze kterých lze dolovat názory veřejnosti na témata zpravodajství. Zahrnutí komentářů do analýzy přináší mnoho nových možností pro novináře, agentury, které studují veřejné mínění a částečné také pro čtenáře. Kontroverzní témata, jako např. migrační krize nebo skandál s emisemi VW, a jejich vnímání v různých zemích může být zdrojem pro další zpravodajství. Zaměření na tyto témata přináší více čtenářů a bohaté diskuze na zpravodajské portály. Mezinárodní agentury a politické instituce mohou využít srovnání vnímání témat v různých zemích k analýze veřejného mínění. Krosjazyčně organizované články a komentáře mohou také využít sami čtenáři, kteří žijí v multikulturním prostředí. Mohou rychle najít a pochopit různé pohledy na kontroverzní témata.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 84-88.

MediaGist[1] [8] je online systém, který je postaven na funkcionalitě agregátorů zpráv, navíc ale přidává dimenzi komentářů. To, že propojuje shluky článků v různých jazycích a analyzuje také komentáře, umožňuje objevovat a prozkoumávat témata, která jsou kontroverzně reportována nebo diskutována v různých zemích.

Následující sekce je věnována technologii, na které je MediaGist postaven, třetí sekce ukazuje stručně jeho funkcionalitu a závěr rozebírá další směry vývoje.

2 Technologie

Zpracování dat v MediaGistu začíná crawler. Ten sbírá články a komentáře pod nimi z předdefinovaných zpravodajských portálů[2]. Pro každý článek je vytvořen RSS soubor, který putuje dále přes další moduly zpracování přirozeného jazyka.

Nejprve jsou rozpoznány pojmenované entity (NER), jak v článku, tak v jeho komentářích, a je jim přiřazeno krosjazyčné ID. NER modul je založen na JRC-Names[3], což je vícejazyčný seznam jmen osob a organizací. Různé varianty téhož jména jsou propojeny stejným ID [9]. Množina výskytů entit je rozšířena detektorem referencí, který rozpoznává dva druhy referencí: části jmen (např. „Zeman“ nebo definitní popisy, např. „český prezident“) [5].

Dalším krokem je přiřazení polarity ke každému článku, komentáři a také každému výskytu entity. Skóre polarity je v intervalu <-100,+100>. Analyzátor polarity používá vícejazyčné a porovnatelné slovníky vzniklé automatickou triangulizací (viz [7]). V případě přiřazení polarity článku nebo komentáři počítá subjektivní slova. V případě přiřazení výskytu entity se omezuje pouze jeho okolí. Navíc obsahuje pravidla pro měření intenzity subjektivního výrazu nebo negace [6]. Přestože strojového učení by lépe predikovalo polaritu, v současné době nemáme k dispozici vhodná trénovací data ve všech cílených jazycích.

U každého článku můžou být i tisíce komentářů. Dalším krokem je tedy jejich automatická sumarizace. Sumarizace v MediaGistu je založena na extraktivním přístupu založeném na latentní sémantické analýze, který používá jak slovní tak entitní příznaky [3]. Tento krok výrazně redukuje velikost dat, která se posílají dalším modulům. Obohacené RSS soubory pak vstupují do fáze shlukování.

Každé čtyři hodiny, pro každý jazyk samostatně, načte shlukovací modul soubory článků publikované během aktuálního týdne a vytvoří jednojazyčné shluky. Je použito aglomerativní hierarchické shlukování se strategií *group average* [2]. Články jsou reprezentované *log-likelihood vektory* a podobnostní funkcí je *kosinus*. Od tohoto kroku obsahuje RSS informace o všech článcích shluku. Krosjazyčný linker pak propojí nejpodobnější shluky mezi jazyky. Linker používá dva typy příznaků: přítomnost

[1] MediaGist je dostupný na adrese: *http://mediagist.eu*. Video, které jej představí lze shlédnout zde: *https://www.youtube.com/watch?v=ONtKw_l6_X4*.

[2] Momentálně MediaGist sbírá data z 8 zdrojů v pěti jazycích: angličtina (theguardian.com), čeština (idnes.cz, ihned.cz, novinky.cz), italština (corriere.it, repubblica.it), francouzština (lemonde.fr) a němčina (spiegel.de).

[3] https://ec.europa.eu/jrc/en/language-technologies/jrc-names.

stejných entit a výskyt stejných deskriptorů z thesauru *EuroVoc*[4] [10]. Posledním krokem je vytvoření souhrnu článků a souhrnu komentářů (které již byly sumarizovány na úrovni článku) pro každý shluk.

RSS soubor tak obsahuje všechny informace pro prezentační vrstvu založenou na technologii java servletů a JSP: http://mediagist.eu.

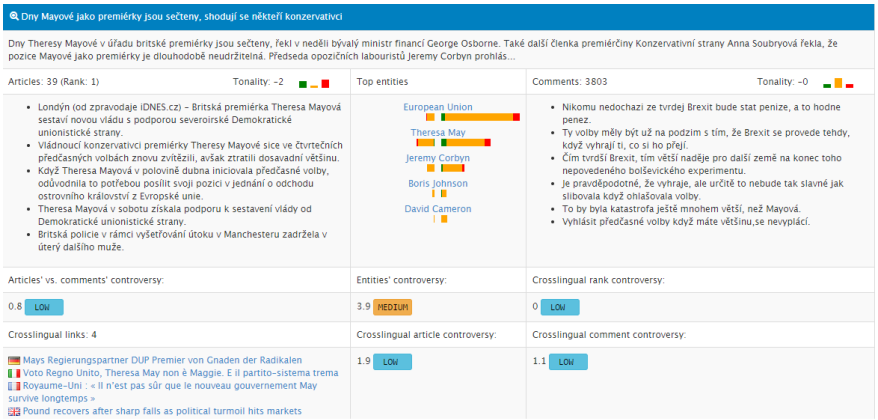
3 Funkcionalita

Systém obsahuje dva pohledy k prozkoumávání dat z médií: pohled na shluky článků (~témata) a pohled na entity. Po výběru jazyka, týdne a řazení dat se zobrazí detaily pohledu a také nejvýznamnější témata/entity analyzovaného výběru v levém panelu.

U každého tématu je zobrazen název a popis, který odpovídá centrálnímu článku. (viz obr. 1). V levé části jsou informace o článcích a v pravé o komentářích. Na obou stranách jsou zobrazeny vygenerované souhrny a agregované údaje o polaritě. Centrální část pohledu ukazuje entity a distribuci jejich polarity v článcích a v komentářích. Dole se nachází odkazy na asociované shluky v dalších jazycích.

MediaGist počítá také několik skóre kontroverze:

- Články vs. komentáře: směrodatná odchylka polarit na obou stranách.
- Entitní kontroverze: porovnává polaritu entit v článcích a komentářích.
- Kontroverze pořadí: porovnává významnost tématu v různých jazycích (dle počtu článků).
- Kontroverze mezi články v různých jazycích: vysoké číslo odpovídá velkému rozdílu polarit článků mezi jazyky.
- Kontroverze mezi komentáři v různých jazycích: vysoká hodnota indikuje témata, která jsou diskutovaná s různou polaritou napříč jazyky (~zeměmi).



Obr. 1. Ukázka hlavního shluku v češtině v týdnu 11.-17. června 2017.

[4] http://eurovoc.europa.eu.

U entit systém ukazuje nalezené varianty jména a jejich frekvence, agregovanou polaritu v článcích a v komentářích a také nejčastější zabarvené pojmy, které pomáhají vysvětlit detekovanou polaritu. Protože máme entity také propojené mezi jazyky, lze spočítat jejich kontroverzi mezi reportováním o dané entitě (články) a veřejném mínění (komentáře) v různých jazycích (~zemích).

4 Závěr

V projektu MediaGist se pracuje na jazykových technologiích, které pomohou detekovat kontroverzi v mezinárodním zpravodajství. Detekce polarity textu identifikuje kontroverzní témata a entity mezi jazyky a skrze sumarizaci je možné data detailněji prozkoumávat. Mimo vylepšení jednotlivých technologií (detekce polarity textu, sumarizace, NER, detekce referencí, shlukování, propojení mezi jazyky, výpočet kontroverze) je cílem dalšího vývoje rozšířit množství dat jak vertikálně (množství zdrojů/jazyků), tak horizontálně (historická data). Systém momentálně konzumuje nefiltrované komentáře. Odstranění osobních útoků, nerelevantních k tématu, a textů od trolů [4] povede ke zvýšení přesnosti detekce názorů uživatelů online světa.

Literatura

1. Atkinson, M. and E. van der Goot: Near real time information mining in multilingual news. In: Proceedings of the 18th International WWW Conference (2009), 1153-1154.
2. Hastie, T., R. Tibshirani, and J. Friedman: The Elements of Statistical Learning. Springer-Verlag, 2009.
3. Kabadjov, M., J. Steinberger, and R. Steinberger: Multilingual statistical news summarization. In: Multilingual Information Extraction and Summarization, volume 2013 of Theory and Applications of Natural Language Processing, Springer (2013), 229-252.
4. Mihaylov, T., G. Georgiev, and P. Nakov: Finding opinion manipulation trolls in news community forums. In: Proceedings of the 19th CoNLL, ACL (2015), 310-314. ACL.
5. Steinberger, J., J. Belyaeva, J. Crawley, L. Della-Rocca, M. Ebrahim, M. Ehrmann, M. Kabadjov, R. Steinberger, and E. Van der Goot: Highly multilingual coreference resolution exploiting a mature entity repository. In: Proceedings of the 8th RANLP Conference, Incoma Ltd. (2011), 254-260.
6. Steinberger, J., P. Lenkova, M. Kabadjov, R. Steinberger and E. van der Goot: Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In: Proceedings of the 8th RANLP Conference, Incoma Ltd. (2011), 770-775.
7. Steinberger, J., M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, P. Lenkova, R. Steinberger, H. Tanev, S. Vzquez, and V. Zavarella: Creating sentiment dictionaries via triangulation. In: Decision Support Systems (2012), 53(4), 689-694.
8. Steinberger, J.: MediaGist: A cross-lingual analyser of aggregated news and commentaries. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, ACL (2016), 145-150.
9. Steinberger, R., B. Pouliquen, M. Kabadjov, J. Belyaeva, and E. van der Goot: JRCNames: A freely available, highly multilingual named entity resource. In: Proceedings of the International RANLP Conference. Incoma Ltd. (2011).

10. Steinberger, R.: Multilingual and cross-lingual news analysis in the europe media monitor (EMM). In: Multidisciplinary Information Retrieval, LNCS 8201, Springer (2013), 1-4.

Poděkování: Tento článek byl podpořen projektem MediaGist, EUs FP7 People Programme (Marie Curie Actions), č. 630786.

Annotation:

A crosslingual analyser of news and their commentaries

The paper introduces project MediaGist, which builds an online system for analysing and linking news and commentaries in five languages. It is designed to assist journalists to detect and explore news topics, which are controversially reported or discussed in different countries. Sentiment analysis and summarization are key technologies used for text analytics. Sentiment analysis provides a basis to compute controversy scores and summaries help to explore the differences.