Zpracování přirozeného jazyka v rámci projektu InteCom

Pavel Král[1,2], Ladislav Lenc[2], Josef Steinberger[1,2], Tomáš Brychcín[2], Pavel Přibáň[2] a Jakub Sido[2]

[1] Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni
Univerzitní 8, 306 14 Plzeň
[2] Nové technologie pro informační společnost - NTIS, FAV ZČU v Plzni
Technická 14, 306 14 Plzeň
{pkral,llenc,jstein,brychcin,pribanp,sidoj}@kiv.zcu.cz

Abstrakt. Cílem tohoto příspěvku je seznámení čtenáře s úlohami, kterými se zabývá naše výzkumná skupina v rámci projektu „Výzkum a vývoj inteligentních komponent pokročilých technologií pro plzeňskou metropolitní oblast (InteCom)“ a jejich širším kontextem. Jedná se především o sémantickou analýzu textu a její využití v návazných úlohách v oblasti zpracování přirozeného jazyka (NLP). V rámci článku budou rovněž představeny relevantní problémy v dané oblasti a možnosti aplikací výsledků výzkumu. Dále bude čtenář seznámen s cíli a výstupy, které jsou naplánovány v rámci tohoto projektu. Na závěr popíšeme výsledky, které byly dosaženy během prvního roku a půl řešení projektu.

Klíčová slova: InteCom · sémantická analýza · zpracování přirozeného jazyka

1 Představení projektu

Cílem výzkumného záměru Inteligentní komponenty pokročilých technologií je vyvinout IT technologie škálovatelné cenou a výkonovými parametry, které převedou stávající teorie, poznatky, principy, metody a algoritmy z oblasti automatizace, robotiky, umělé inteligence, monitorování, diagnostiky a zpracování signálů do formy, která významně urychlí a zviditelní možnost jejich uplatnění v praxi. Záměr se skládá ze tří výzkumných témat:

- robotické a řídící technologie,
- diagnostické a rozhodovací technologie,
- monitorovací technologie.

Výsledky projektu jsou orientovány k aplikacím ve výrobě, dopravě, energetice, zdravotnictví, veřejné správě, telekomunikacích a dalších oblastech lidské činnosti. Jejich úspěšné uplatnění v produktech a procesech konkrétních podniků a institucí významně posílí a rozšíří spolupráci výzkumného centra NTIS s aplikační sférou.

Projekt je financovaný Ministerstvem školství, mládeže a tělovýchovy z operačního programu EU Výzkum, vývoj a vzdělávání v rámci výzvy Předaplikační výzkum pro ITI plzeňské metropolitní oblasti.

Zpracování přirozeného jazyka se řeší v rámci druhého výzkumného tématu, diagnostické a rozhodovací technologie.

2 Zpracování přirozeného jazyka

Klíčovou oblastí, kterou se budeme zabývat v rámci řešení tohoto projektu, je sémantická analýza textu [10,7] a její využití v návazných úlohách zpracování přirozeného jazyka (NLP). Sémantická analýza se zabývá způsoby reprezentace významu přirozeného jazyka a lepší reprezentace významně přispěje ke zlepšení výsledků řady dalších NLP úloh např. klasifikace dokumentů, rozpoznávání pojmenovaných entit, analýza polarity textu, automatická sumarizace, strojový překlad a další.

Sémantická analýza (porozumění) jednotlivých slov v textu a zároveň strojové učení s učitelem dnes dosahují velmi dobrých výsledků. Tyto metody bohužel mají své limity: sémantická analýza jednotlivých izolovaných slov neuvažuje kontext, který je ale pro význam a pochopení textu klíčový. Strojové učení s učitelem vyžaduje ruční tvorbu trénovacích dat a dalších jazykových prostředků, což je časově i finančně velmi nákladné. V obou případech nastává problém při změně cílového jazyka nebo i při adaptaci na jinou doménu. Výše uvedené limity je možno překonat pomocí metod bez učitele [11] (nebo s jeho částečnou pomocí) a s využitím modelů schopných analyzovat více jazyků [6]), které jsou zároveň jazykově / doménově téměř nezávislé. Výborných výsledků lze také dosáhnout s využitím (hlubokých) neuronových sítí [3].

3 Relevantní problémy

Většina přístupů v oblasti analýzy textu (vč. automatické klasifikace dokumentů, přiřazení klíčových slov, detekce pojmenovaných entit, apod.) využívá sémantickou informaci pouze ve velmi omezené míře. Tyto metody zpravidla neberou v úvahu kontext a jeho strukturu (vztahy mezi slovy a jejich slovosled) a sémantiku textu často vnímají jen jako bag-of-words.

Tyto přístupy jsou zpravidla založené na pravidlech nebo na strojovém učením s učitelem, protože obojí má v současné době dostatečný aplikační potenciál. Nedostatkem těchto metod analýzy přirozeného jazyka je proto ale jejich jazyková / aplikační závislost.

Dalším důležitým problémem je analýza specifických textů. Jedná se zejména o komentáře v sociálních mediích, které jsou zpravidla velmi krátké, obsahují velké množství překlepů a nespisovných slovních spojení. V případě češtiny často chybí diakritika a správná kapitalizace. Dále sem patří analýza textu z obrazového (zpravidla PDF) formátu, kde je třeba před samotnou analýzou rozpoznat text. Převod do textu může být problematický v případě nízké kvality naskenovaných stránek nebo v případě velkého množství obrázků a tabulek.

4 Aplikační potenciál

S rostoucím množstvím textových dat (na Internetu, ale i v interních systémech) je stále více důležité jejich inteligentní zpracování. Enormní množství textu generované uživateli skrývá velmi hodnotné informace, které analýza textů může odhalit. Mnoho aplikací dnes používá fulltextové vyhledávání, detekci spamů, filtrování obsahu webových stránek, strojový překlad do jiných jazyků, detekci polarity textu recenzí atd. Analýza velkých textů je stále více populární kvůli velkému potenciálu jak v privátním tak ve veřejném sektoru. Firmy se potřebují vyznat v obrovském množství interních dat (často v podobně PDF), potřebují také chápat potřeby svých zákazníků, kteří je píší online, aby lépe cílily své marketingové kampaně a vytvářely lepší produkty. Vládní organizace musí monitorovat globální problémy společnosti (např. uprchlické krize). Řešení si navíc musí poradit s různými jazyky. Analýza textu je také „core business“ hlavních internetových společností, např. Facebook, Google, Twitter, Baidu, Yahoo. Sémantická analýza textu výrazně přispěje ke zlepšení výsledků ve všech výše uvedených aplikačních oblastech.

5 Cíle a výstupy

V rámci projektu se zaměříme na překonání potřeby vytvářet a trénovat jazykově závislé modely, zaměříme se na sémantickou cross-linguální analýzu delších textů (slovních spojení, vět, odstavců, apod.) pomocí metod učení bez učitele (příp. s jeho částečnou pomocí) nebo s využitím neuronových sítí. Výstupem budou metody, které v experimentech prokáží schopnost zpracovávat více jazyků a zároveň se sníží potřeba vytvářet označená trénovací data. Zvýší se tak potenciál pro nasazení našich aktuálních metod do praxe. Použitelnost vyvinutých metod v praxi bude ověřena na datech z reálného prostředí, viz např. následující usecase.

Řada IT subjektů v současné době intenzivně řeší digitalizaci dokumentů (tištěných, ale i ručně psaných) s následným ukládáním do databáze. Další skupina organizací již dokumenty v databázi uložené má. V obou případech ale zpravidla chybí jejich jednoduché zpřístupnění uživatelům (rychlé nalezení dokumentu dle různých kritérií, tzv. inteligentní vyhledávání). Naším cílem je vyvinout metody pro rozpoznání textu z naskenovaných dokumentů, jejich analýzu, zaindexování do fulltextové databáze a umožnění rozšířeného inteligentního vyhledávání nad jejich obsahem vč. metadat. Zaměříme se nejen na tištěné dokumenty, ale i na dokumenty ručně psané, u kterých předpokládáme využití zejména hlubokých neuronových sítí. V rámci analýzy obsahu dokumentů bude automaticky provedena kategorizace dokumentů, přiřazena klíčová slova, určeny pojmenované entity (jako např. osoby, instituce, data, apod.), vytvořen automatický souhrn obsahu dokumentu a v případech, kde to je relevantní, bude určen sentiment dokumentu. Při zpracování textu (tj. napříč všemi použitými metodami analýzy dokumentů) bude provedeno sémantické zpracování obsahu dokumentů, kde budou využity nejnovější poznatky z metod strojového učení a výpočetní lingvistiky.

6 Dosažené výsledky

Nejprve jsme navrhli novou metodu pro reprezentaci textu, která rozšiřuje stávající metody reprezentace slov o globální informaci z Wikipedie [9]. Překonali jsme tak state-of-the-art v oblasti mapování slov na vektory reálných čísel. V oblasti vícejazyčné (cross-linguální) sémantické analýzy jsme vymysleli metodu transformace, která doplňuje nejlepší transformační metody o vážení [1]. Náš přístup překonává ostatní metody v úloze sémantické podobnosti vět na několika souborech dat v různých jazycích. Nové metody byly úspěšně využity ve dvou NLP úlohách a to pro automatické rozpoznávání dialogových aktů [5] a pro vícejazyčné slovní analogie [2].

Dále jsme navrhli a implementovali základní metody pro segmentaci a rozpoznávání tištěných textů (OCR) založené na hlubokých neuronových sítích typu CNN a LSTM [4] . V současné době probíhá ověřování funkčnosti těchto metod na reálných historických datech. Zároveň jsme vytvořili systém pro predikci emocí v krátkých textech, který rovněž využívá neuronové sítě typu LSTM. Funkčnost systému byla ověřena na reálných datech, tj. krátkých zprávách (tweetech) ze sociální sítě Twitter [8].

7 Závěr

V rámci tohoto článku jsme seznámili čtenáře s úlohami, kterými se zabývá naše výzkumná skupina v rámci projektu InteCom a jejich širším kontextem. Zároveň byly představeny relevantní problémy v dané oblasti a možnosti aplikací výsledků výzkumu. Dále byl čtenář seznámen s cíli a výstupy, které byly naplánovány v rámci tohoto projektu. Na závěr jsme popsali výsledky, které byly dosaženy během prvního roku a půl řešení projektu.

Poděkování: Tento článek vznikl za podpory projektu "VaV inteligentních komponent pokročilých technologií pro metropolitní oblast Plzeňského kraje (InteCom)"reg. č.: CZ.02.1.01/0.0/0.0/17_048/0007267 finacovaného z EFRR.

Literatura

1. Brychcín, T.: Linear transformations for cross-lingual semantic textual similarity. Knowledge-Based Systems (2019). https://doi.org/10.1016/j.knosys.2019.06.027
2. Brychcín, T., Taylor, S., Svoboda, L.: Cross-lingual word analogies using linear transformations between semantic spaces. Expert Systems with Applications **135**, 287–295 (2019). https://doi.org/10.1016/j.eswa.2019.06.021
3. Kim, Y.: Convolutional neural networks for sentence classification. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar (October 25-29 2014)

4. Lenc, L., Martínek, J., Král, P.: Tools for semi-automatic preparation of training data for ocr. In: MacIntyre, J., Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds.) Artificial Intelligence Applications and Innovations. pp. 351–361. Springer International Publishing, Cham (24-26 May 2019). https://doi.org/10.1007/978-3-030-19823-7_29
5. Martínek, J., Král, P., Lenc, L., Cerisara, C.: Multi-lingual dialogue act recognition with deep learning methods. In: Interspeech. Graz, Austria (15-19 September 2019)
6. McDonald, R., Petrov, S., Hall, K.: Multi-source transfer of delexicalized dependency parsers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). pp. 62–72. Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
7. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
8. Přibáň, P., Martínek, J.: UWB at IEST 2018, emotion prediction in tweets with bidirectional long short-term memory neural network. p. 224–230. Association for Computational Linguistics, Brusel, Belgie (2018), https://www.aclweb.org/anthology/W18-6232
9. Svoboda, L., Brychcín, T.: Improving word meaning representations using wikipedia categories. Neural Network World **28**(6), 523–534 (2018). https://doi.org/10.14311/NNW.2018.28.029
10. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research **37**, 141–188 (2010)
11. Zanzotto, F.M., Korkontzelos, I., Fallucchi, F., Manandhar, S.: Estimating linear models for compositional distributional semantics. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 1263–1271. Association for Computational Linguistics (2010)

Annotation:

The goal of this paper is to familiarize the reader with the main tasks of our research group in the project " Research and development in Intelligent Components of Advanced Technologies for the Pilsen metropolitan area" and their wider context. It consists mainly in text semantic analysis and its use in related tasks in the field of natural language processing (NLP). The article will also present relevant issues in the area and the possibilities of applying the research results. Finally, we present the objectives and outputs planned within this project.