Vliv normalizace slov na klasifikaci textů

Michal Toman[1], Roman Tesař[1], Karel Ježek[1]

[1]Katedra informatiky a výpočetní techniky, FAV, ZČU v Plzni,
Univerzitní 22,
306 14, Plzeň
{mtoman, romant, jezek_ka}@kiv.zcu.cz

Abstrakt. Příspěvek porovnává vliv různých normalizačních metod na klasifikační úlohu. Část článku je věnována popisu naší lemmatizační metody založené na použití tezauru EWN. Prezentujeme srovnání výsledků získaných EWN metodou a ostatními normalizačními metodami. Zkoumána je také celková míra ovlivnění výsledků klasifikace textu jeho předzpracováním – normalizací slov a odstraněním stop-slov.

Klíčová slova: normalizace slov, lemmatizace, stemming, klasifikace, EuroWordNet

1 Úvod

V tomto článku jsme se zaměřili především na vliv normalizace slov na klasifikační úlohu. Zkoumáme míru ovlivnění výsledků nejen v anglickém jazykovém prostředí, čímž navazujeme na práci [3], ale zabýváme se také specifiky českého jazyka.

Pro anglický a český jazyk jsme vyvinuli vlastní lemmatizační metodu založenou na využití tezauru EuroWordNet. Ve spojení s modulem indexace je metoda schopná převádět jednotlivé tvary slov do interní jazykově nezávislé formy. Míru ovlivnění klasifikace normalizací slov popisujeme v kapitole 3, kde také diskutujeme klady a zápory EWN metody a uvádíme porovnání s ostatními normalizačními metodami.

2 EWN lemmatizace

Tezaurus EWN [2] se skládá z množin synonym (synsetů) a vztahů mezi nimi. Každý synset má přiřazený unikátní index, označovaný jako ILI (InterLingual Index), který je shodný pro všechny jazykové mutace tezauru. Tvorba lemmatizačního slovníku pro námi navrženou metodu je založena na využití slovníků Ispell [5] a tezauru EWN. EWN metoda vyžaduje lemmatizační slovník obsahující dvojice slov *(w, l)*, kde w představuje tvar slova v textu a l je jemu odpovídající lemma. Ispell poskytuje slovník kořenů slov r_j, afixů a omezující pravidla pro generování tvarů slov. Ispellem generujeme množinu slov W_j obsahující slova w_{ij} vytvořená z kořenu r_j. Předpokládáme, že množina slovních tvarů W_j obsahuje také základní tvar slova w_{lj}. Procházením slov w_{ij} množiny W_j hledáme slovo w_{lj}, které koresponduje s některým lemmatem l tezauru EWN. Při nalezení shody ztotožníme slovo w_{lj} s lemmatem l a svážeme množinu W_j se synsetem, ve kterém bylo nalezeno l.

Výše popsaným postupem dochází k nejednoznačnosti při přiřazování synsetu k víceznačným slovům. Řešením je disambiguace slov, kterou se zabýváme v práci [10], a proto tuto problematiku nebudeme dále v článku rozebírat.

Jedním z možných vylepšení výkonu lemmatizátoru je fuzzy porovnávání tvarů slov v dokumentu a lemmat ve slovníku. Pro další vylepšení metody jsme použili mofologický analyzátor [4]. Jelikož je princip metody EWN jazykově nezávislý, bylo možné vytvořit lemmatizační slovník i pro angličtinu.

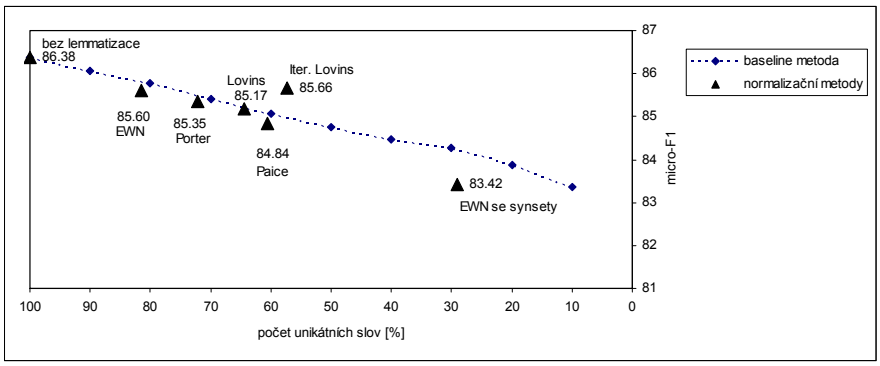
Pro zjištění vlivu indexace slov pomocí synsetů v porovnání s tradiční indexací lemmat jsme vytvořili modifikaci EWN metody označenou jako EWN bez použití synsetů. Každému slovu ze synsetu je v tomto případě přiřazován odlišný index, podobně jako u tradiční lemmatizace.

3 Experimenty

V experimentech sledujeme míru ovlivnění výsledků klasifikace normalizací slov a odstraněním stop-slov na dvou korpusech. Pro testy byl vybrán klasifikátor multinomial Naive Bayes [7]. Kvalitu klasifikace porovnáváme pomocí standardních metrik, jmenovitě mírou micro-F1, macro-F1, přesností a úplností. Statistickou významnost výsledků použitím t-testu a dále používáme techniku 4-cross fold validace [1]. Oba korpusy jsou nejdříve předzpracovány rozdílnými normalizačními technikami a následně klasifikovány.

Pro experimenty jsme použili dva textové korpusy - anglické dokumenty vybrané z Reuters Corpus Volume 1 a české dokumenty agentury ČTK. Testovali jsme 6 normalizačních algoritmů – Lovins, Iterated Lovins [6], Paice[8], Porter's stemmer[9], EWN metodu s použitím a bez použití synsetů. Pro český jazyk byla do testů zahrnuta také algoritmická lemmatizace.

Jelikož předběžné výsledky naznačily korelaci mezi počtem unikátních slov v korpusu a kvalitou klasifikace, rozhodli jsme se pro srovnání vytvořit uměle redukované testovací korpusy. Jako základ posloužil korpus bez předzpracování, ze kterého jsme postupně odebírali vždy 10 % nejméně četných slov. Charakteristika je v grafu zobrazena tečkovanou čarou a je označena jako „baseline".

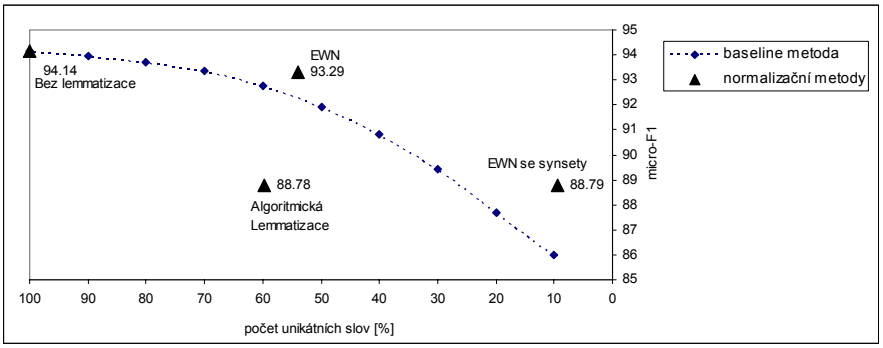


Graf 1. Počet unikátních slov v korpusu a přesnost klasifikace pro anglický korpus.

Přesnost klasifikace je úzce spojena s množstvím slov vstupujících do klasifikátoru. Výjimku tvoří algoritmus Iterated Lovins, kde je míra micro-F1 nad hodnotou baseline přístupu.

Tabulka 2. Výsledky klasifikace anglického korpusu.

	P	R	micro-F1	macro-F1	Počet unikátních slov v korpusu
Bez lemmatizace	82.55	90.60	86.38	84.26	50494
EWN-Lem.	81.35	90.33	85.60	83.49	41151
EWN-L. (se synsety)	77.36	90.51	83.42	80.59	14625
Iterated Lovins	80.80	91.14	85.66	83.58	28896
Lovins	80.57	90.35	85.17	83.03	32517
Paice	80.01	90.30	84.84	82.70	30540
Porter's Stem.	80.66	90.62	85.35	83.17	36421



Graf 2. Počet unikátních slov v korpusu a přesnost klasifikace pro český korpus.

Jako pozitivní lze hodnotit výkon EWN metody používající synsety, která při více než desetinásobném zmenšení korpusu produkuje výsledky bez významného zhoršení přesnosti klasifikace. Jak je zřejmé z grafu 2, slovníkové lemmatizační metody jsou v případě morfologicky složitých jazyků vhodnější. Obě EWN metody poskytují na českém korpusu dobré výsledky s přesností větší než baseline metoda.

Tabulka 3. Výsledky klasifikace českého korpusu.

	P	R	Micro-F1	Macro-F1	Počet unikátních slov v korpusu
Bez lemmatizace	93.79	94.49	94.14	74.21	130428
EWN-Lem.	90.91	95.80	93.29	75.20	70289
EWN-L. (se synsety)	81.35	97.72	88.79	70.71	12224
Algoritmická L.	89.20	88.41	88.78	56.00	78051

4 Závěr

Jako nejlepší přístup pro zpracování textových dokumentů se jeví odstraňování stop-slov bez použití normalizačních metod. Pokles přesnosti klasifikace v případě aplikování normalizace je zřejmý a v některých případech statisticky významný. Na druhou stranu umožňuje normalizace redukci korpusu a dimenze dokumentů, což je přínosné, preferujeme-li zpracování velkého objemu dat nebo rychlost.

Metoda EWN se jeví jako vhodná. Obě konfigurace – s použitím a bez použití synsetů – produkují slibné výsledky na testovacím korpusu ČTK. Očekáváme, že rostoucí kvalita tezauru EWN bude vylepšovat také výsledky lemmatizační metody.

Tento výzkum byl částečně podpořen Národním programem výzkumu II, projekt 2C06009 (COT-SEWing).

Reference

1. Dietterich, T.G. 1998. Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, pp. 1895-1923.
2. EuroWordNet. http://www.illc.uva.nl/EuroWordNet/
3. Goncalves T., Quaresma P.: The impact of NLP techniques in the multilabel classification problem. *Intelligent Information Processing and Web Mining 2004*, Advances in Soft Computing, pp. 424-428, Zakopane, 2004. Springer-Verlag.
4. Hajic, J.: Morphological Analyzer. http://quest.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Morphology
5. Ispell. http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html
6. Lovins, J. B.: Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics 11*, 1968, pp. 22-31.
7. McCallum A., K. Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI/ICML-98*, AAAI Press, pp. 41-48.
8. Paice, Chris D.: Another Stemmer. *SIGIR Forum 24 (3)*, 1990, pp. 56-61.
9. Porter, M. F.: An Algorithm for Suffix Stripping. *Program 14*, 1980, pp. 130-137
10. Toman M., Jezek K.: Modifikace bayesovského disambigutoru, *ZNALOSTI2005*, pp.306-313, Stará Lesná, Slovensko 2005, ISBN 80-248-0755-6

Annotation:

Impact of Word Normalization on Text Classification

In this paper we focus our attention on the comparison of various lemmatization and stemming algorithms, which are often used in natural language processing (NLP). We present a novel lemmatization algorithm that utilizes the multilingual thesaurus Eurowordnet (EWN). We describe the algorithm in detail and compare it with other widely used algorithms for word normalization on two different corpora. We also discuss the influence of the word normalization on classification task in general.