Klasifikace multilinguálních korpusů s využitím tezauru EuroWordNet

Michal Toman[1] a Karel Ježek[1]

[1] Katedra Informatiky a výpočetní techniky, Západočeská univerzita,
Univerzitní 22, 30614 Plzeň
{mtoman, jezek_ka}@kiv.zcu.cz
http://www.kiv.zcu.cz

Abstrakt. Klasifikace textů je jednou z aktuálních oblastí výzkumu. Tento článek popisuje výsledky experimentů klasifikace textových korpusů s využitím tezauru EuroWordNet (EWN), porovnává různé přístupy a vyvozuje závěry pro možná vylepšení klasifikační úlohy. Součástí práce je popis algoritmů a metod použitých v navrženém klasifikačním systému. Cílem experimentů bylo ověřit vliv multilinguality textových korpusů na kvalitu klasifikace a navrhnout vhodné využití vícejazykového tezauru za účelem zlepšení, či zobecnění možností klasifikace multilinguálních textových korpusů. V článku je diskutována problematika lemmatizace a indexace s respektováním vícejazyčného prostředí. Popsáno je též následné navázání lemmat na tezaurus EWN. Testy byly vykonány na korpusech Reuters a ČTK a poukazují na fakt, že při použití vhodných klasifikačních algoritmů lze provádět klasifikaci dokumentů zcela nezávislou na jazyku. Dále je ve článku prokázáno, že výsledky jsou velmi závislé na volbě klasifikačního algoritmu. Jako perspektivní klasifikační metody modifikovatelné pro vícejazykové prostředí se ukazují např. algoritmy NBCI, Itemsets, TF/IDF.

Klíčová slova: klasifikace, kolekce dokumentů, tezaurus, EuroWordNet, mutilinguální korpus, lemmatizace, přirozený jazyk, analýza dokumentů, Bayesův teorém.

1. Multilinguální korpusy

V současné době se častěji objevuje nutnost uchovat a počítačově zpracovávat dokumenty, které jsou uloženy v jedné knihovně, ale jsou napsány v různých jazycích. Dříve se tento aspekt spíše zanedbával. Mnohé systémy pro zpracování textů předpokládají jednojazyčné prostředí a svou funkci tomu mají uzpůsobenou. Možnost uložení vícejazyčných dokumentů bud vůbec neřeší, nebo pouze okrajově. Považujeme-li Internet, konkrétně webové stránky, za velký elektronický archiv, je zřejmé, že obsažené informace jsou obecně v různých jazycích. S postupující integrací jednotlivých států a rozšiřováním Evropské unie se dostává respektování vícejazyčnosti do popředí

zájmu. Typickým příkladem aplikace multinguálního systému může být prohledávání webových stránek, vědeckých článků, zákonů, předpisů a podobně. Lze také rozšířit stávající vyhledávací systémy tak, aby lépe umožňovaly vyhledávání ve vícejazykovém prostředí. Předpokládáme, že vytvářený systém by našel uplatnění v rozsáhlejších digitálních knihovnách, kde se vyskytují dokumenty v různých jazycích, případně ve státní správě, která bude stále častěji přicházet do styku s cizojazyčnými dokumenty. V neposlední řadě může být zakomponovaný jako součást námi řešeného systému pro podporu vědeckých pracovníků.

Stávající řešení dokumentografických systémů, až na výjimky reprezentované pokusnými systémy (MILK, AutIndex [5], …), nemají problematiku multilinguálního prostředí jako svůj primární cíl. Ve většině systémů se provede rozpoznáním jazyka a následně oddělené zpracování dokumentů, což nemusí dostačovat.

Za předpokladu, že uživatel zná několik jazyků, je vhodné umožnit jedním dotazem vyhledat všechny relevantní dokumenty. S použitím navrhovaného modelu se může například provádět kategorizace dokumentů i v případě, že jsou stávající třídy definovány pouze pro jediný jazyk.

Problém nastává v případě indexování vícejazyčného korpusu, kdy ekvivalentní překlady téhož slova mají v rozdílných jazycích různé indexy (např. slovo strom je indexováno jinak než anglický překlad tree). Výhodnější je indexovat ekvivalentní překlady jedním indexem.

Rozhodli jsme se vytvořit model vícejazykového systému, který by poskytoval stejně kvalitní výstupy jako stávající systémy, ale s dokonalým respektováním vícejazyčného prostředí. Dlouhodobějším cílem je vytvoření systému, který bude poskytovat jedním dotazem výsledek, jenž není závislý na jazyce dotazu, ani jazyce vyhledaných dokumentů.

Jako pokusnou úlohu na otestování našeho přístupu jsme zvolili klasifikaci dokumentů. V tuto chvíli je testovací kolekce složena z českých dokumentů (tiskové zprávy ČTK) a anglických dokumentů (tiskové zprávy Reuters). Do budoucna se počítá s rozšířením i na další evropské jazyky. Kolekce vznikla výběrem shodných tříd z obou tiskových agentur dovolujícím vyzkoušet klasifikaci dokumentů. Cílem bylo ověřit vhodnost použití tezauru EuroWordNet (EWN) jako jádra zpracování vícejazyčných korpusů.

2. Tezaurus EWN

Jádro navrhovaného systému je aplikace tezauru EuroWordNet (EWN [6]) jako referenčního slovníku pro provázání slov jednotlivých jazyků. EWN je tezaurus, který sobě odpovídajícím skupinám synonym (synsetům) přiřazuje shodné indexy. To následně umožňuje jednotnou indexaci pro různé jazyky.

EWN je vícejazyčná databáze pro některé evropské jazyky (angličtina, dánština, italština, španělština, němčina, francouzština, čeština, estonština). Je strukturovaná podobně jako původní Wordnet vytvořený Princetonskou univerzitou. EWN obsahuje množiny synonym a vzájemné vztahy mezi nimi. Jednotlivé množiny synonym (synsety) jsou navíc spojené pomocí tzv. ILI (inter-lingual-index) tak, že shodný synset

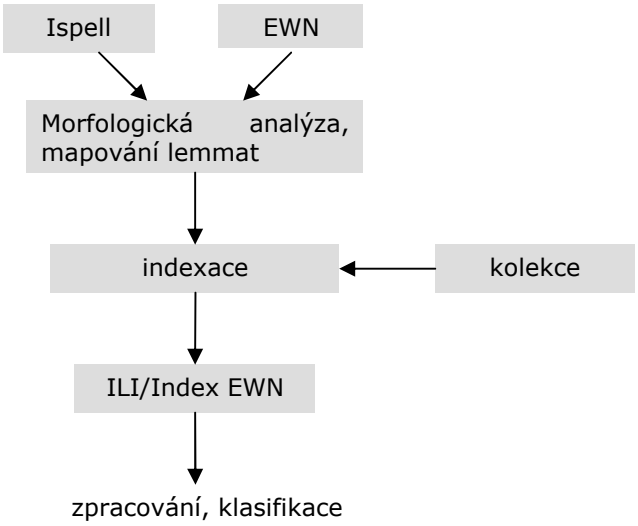
jednoho jazyka má tentýž index v jiném jazyce. Díky tomu bylo možné považovat tento index za ekvivalentní indexu, jenž je použitý při klasické indexaci jednojazyčného korpusu.

Určitou nevýhodou tezauru EWN je jeho přílišná jemnost. Pro téměř totožný výraz je často definován rozdílný index. V takovém případě může docházet k nekorektní indexaci a výsledky klasifikační metody mohou být velice špatné. Jedná se o problém, který je řešitelný např. shlukováním podobných synsetů. Řešení jsou diskutována dále. Další nevýhoda EWN souvisí s nestejnou rozpracovaností jednotlivých wordnetů, takže některé synsety nemají v určitých jazycích odpovídající ekvivalent.

3. Moduly systému

Experiment využití EWN pro klasifikaci vícejazyčných dokumentů byl rozdělen do několika fází. Nejdříve bylo nutné vytvořit kvalitní lemmatizační slovník. Stávající algoritmická a slovníková lemmatizace se musela upravit, aby bylo možné lemmata propojit s tezaurem EWN. Dosud námi používané metody převáděly slova na lemmata, která nebyla totožná se základními tvary slov. Následně nebylo možné vytvořit korektní vazbu na EWN. Pro splnění této podmínky bylo nutné vytvořit nové lemmatizační slovníky.

Lemmatizovaná slova jsou následně namapována na synsety EWN a pomocí jednoduché transformace je zjištěn index, který je zpracován při klasifikaci [obr. 1]. Funkce jednotlivých komponent jsou rozebrány dále.



Obr. 1 Návaznost komponent klasifikačního systému

3.1 Lemmatizační modul

K vytvoření lemmatizačního slovníku jsme zvolili extrakci tvarů slov z programu Ispell [7]. Lemmatizačním slovníkem rozumíme alfabeticky uspořádanou množinu slov a jim odpovídající lemmata. Ispell je interaktivní program pro kontrolu pravopisu, který podporuje většinu evropských jazyků. Primárním účelem programu je procházet texty, kontrolovat pravopis a případně navrhovat opravy nerozpoznaných slov.

Základní myšlenkou bylo vzít kmeny slov, které jsou uloženy ve slovníku Ispellu a z těch pomocí Ispellu odvodit všechny existující tvary. Kmen slova byl považován za základní tvar, měl by se tedy vyskytovat v tezauru. Tento přístup fungoval dokonale u anglického jazyka, ovšem selhával u češtiny, která disponuje daleko větší flexí.

Základní problém spočíval v tom, že kmen slova se nemusí shodovat se základním tvarem. Příkladem může být slovo lano, jehož kořen uvedený v Ispellu je lan a přípony mohou být například -o -em -ech apod. Tedy základní tvar slova (lano) se neshoduje s kořenem (lan).

Při řešení výše uvedeného problému jsme vycházeli z předpokladu, že základní tvar slova je v množině všech možných tvarů slova, které jsme získali z Ispellu. Proto jsme vzali slovník Ispellem vygenerovaných slov, vytvořili jsme podmnožiny tvarů slov odpovídající jednomu kmeni, resp. základnímu tvaru slova, a pro každou množinu jsme hledali odpovídající lemma v EWN.

Algoritmus lze popsat:

- Pro každou množinu kmene slova proveď:
  - Pro každý tvar z množiny proveď:
    - Hledej odpovídající tvar v EWN
      - Pokud se tvar vyskytuje, tvar je základní pro všechna slova z aktuální množiny
      - Pokud se zde tvar nevyskytuje, pokračuj dalším tvarem

V případě, že se v množině tvarů nenalezne ani jeden tvar slova, který lze navázat na EWN, tak se prohlásí kmen slova za základní tvar. Tato možnost nenastává příliš často.

Dalšího vylepšení bylo dosaženo využitím morfologického analyzátoru. Jako ukázka výsledku morfologického analyzátoru mohou sloužit slova být a je, která se typicky indexují různými indexy, protože nemají shodný kmen slova. Po aplikaci analyzátoru jsou slova převedena do korektního základního tvaru, tedy být. Podobná vlastnost je důležitá také při stupňování přídavných jmen.

Nevýhodou modelu je velikost takto vytvořeného slovníku. V případě uchování slov v seznamu dvojic {slovo, základní tvar} dosáhne velikost téměř 100 MB pro češtinu. Tu lze však považovat za určitý extrém, jelikož podobnou flexi má jen málo jazyků. Pro angličtinu je velikost slovníku pouze 3 MB.

Do lemmatizátoru vstupují slova a výstupem jsou indexy obsažené v EWN (ILR indexy). Každý index se skládá z označení (např. eng20 znamenající slovo zařazené anglickým týmem, EWN verze 2.0), vlastního unikátního čísla (např. 06900919) a

jednopísmenné zkratky označující slovní druh (v – sloveso, n – podstatné jméno, a – přídavné jméno, apod.).

Příklad výstupu lemmatizátoru a indexace do EWN:

Původní tvar:
„Vlna chladu si vyžádala 100 mrtvých."

Výstup po lemmatizaci a indexaci:
„eng20-06900919-n eng20-04448750-n eng20-02526983-v eng20-13048967-n eng20-00100393-a"

Jednodušší je situace u anglického jazyka, kde lze použít pro lemmatizaci i algoritmickou metodu – např. Porterův algoritmus. Také vytvoření slovníků s použitím Ispellu poskytuje korektní výsledky i s „naivním" přístupem, jenž se u jazyků se složitějším tvaroslovím nedá uplatnit.

Velkou výhodou takto vytvářených lemmatizačních slovníků je využití již ověřených částí (Ispell, EWN), které jsou navíc dostupné již téměř ve všech evropských jazycích. Obecně lze předpokládat kvalitu takto vytvářených slovníků mezi kvalitou anglického (velmi jednoduché tvarosloví) a českého (složité tvarosloví). Přesto považujeme za vhodné provést pro každý nově přidávaný jazyk určité úpravy algoritmu lemmatizace takové, aby respektovaly specifika flexe daného jazyka. Příkladem může být němčina, kde by bylo nutné věnovat zvýšenou pozornost slovům s odlučitelnými předponami.

3.2 Mapování lemmat na synsety EWN

Získané slovníky je nutné namapovat na synsety EWN. Cílem je vyhledat k jednotlivým základním tvarům odpovídající slova v EWN a odvozeným tvarům slova přiřadit index EWN. V jazyce se ovšem vyskytují víceznačná slova, která jsou stejně zapsána, ale mají jiný význam, tudíž jsou zahrnuta v několika synsetech. K rozhodnutí správného významu je nutné využít disambiguaci. Ve slovnících není dostatečná znalost souvislosti daného slova s nějakým významem – jedná se o oddělená slova bez kontextu. Úloze disambiguace se nelze vyhnout a bude se provádět při zpracování textového korpusu. Výsledkem mapování jsou slovníky, kdy každému tvaru slova odpovídá jeden index EWN.

3.3 Morfologická analýza

Do fáze vytváření lemmatizačního slovníku byl zakomponován morfologický analyzátor vytvořený Janem Hajičem [8]. Jedná se o univerzální nástroj pro morfologickou analýzu textu. Uplatnil se především při zpracování stupňovaných přídavných jmen a nepravidelných sloves.

4. Použité klasifikační metody

Využití EWN pro zpracování multilinguálních kolekcí jsme se rozhodli ověřit na případě klasifikační úlohy, se kterou máme zkušenosti a vytvořené nástroje pro testování přesnosti a úplnosti klasifikace. Jako testovací korpus jsme zvolili české a anglické texty, konkrétně tiskové zprávy ČTK a Reuters, které jsme používali na testování i dříve. Testy byly prováděny na korpusu 82000 českých a 25000 anglických dokumentů. Zprávy spadaly do jedné z 5 tříd, které byly v obou korpusech obsahově podobné. Jednalo se o počasí (4 %), sport (30 %), politiku (58 %), zemědělství (3 %) a zdravotnictví (5 %). Cílem bylo ověřit, zda a jak moc multilinguální prostředí ovlivní výsledky klasifikace. Volili jsme různý stupeň zapracování EWN do klasifikační úlohy. Nejdříve byl uvažován referenční stav, kdy EWN není použit vůbec, dále bylo použito EWN na jednojazyčný korpus a nakonec se testovala křížová klasifikace (trénovací korpus český, testování na anglických datech).

4.1 NBCI

Tato metoda byla vytvořena naší katedrou a jedná se o kombinaci metody Naive Bayes s Itemsety [9]. Z každého dokumentu jsou vybrány charakteristické itemsety a ty jsou dále zpracovávány metodou Naive Bayes. Výhodou je vyšší rychlost a v případě multilinguálního korpusu netrpí metoda problémem malého průniku termů v různých jazycích jako prosté použití Naive Bayes.

4.2 TFIDF

Jedná se o metodu založenou na sledování frekvence termů v dokumentech (viz [10]). Každému termu je přiřazena váha, která je tím vyšší, čím se daný term vyskytuje v dokumentu více a naopak nižší, čím se term vyskytuje častěji v jiných dokumentech.

Pro klasifikaci dokumentů je použita kosinová míra:

$$Sim(Q, Di) = \frac{\sum_{k=1}^{n} (q_k \cdot w_{ik})}{\sqrt{\left(\sum_{k=1}^{n} {w_{ik}}^2 \cdot \sum_{k=1}^{n} {q_k}^2 \right)}}$$

, kde Q je vektor vah q_k termů nově zařazovaného dokumentu a D_i je vektor vah w_{ik} náležejících termům ve třídě C_i.

4.3 Itemsety

Pro hledání častých skupin termů je využita modifikace Apriori algoritmu, která slouží k nalezení položek, které se v dokumentech jedné třídy často vyskytují. Itemset je vybrán jako charakteristický pro určitou třídu, pokud váhový faktor překročil zvolenou mez. V klasifikační fázi je dokument zařazen do třídy za předpokladu, že C_j (množina itemsetů reprezentující třídu T_j) obsahuje itemsety s dostatečnou váhou asociace dokumentu D s třídou T_j podle vzorce:

$$W_{Tj}^{D} = \sum_{i=1}^{|Cj|} wf_{|\Pi_i|} \times w_{\Pi_i} \text{ , kde } (\prod_i \in C_j) \wedge (\prod_i \subseteq D) \text{ pro j = 1, 2, 3, … , L}$$

Váha asociace je zde určena součtem součinů vah w_{Π_i} s váhovými faktory $wf_{|\Pi_i|}$ pro všechny itemsety dané třídy, jež se vyskytují v právě klasifikovaném dokumentu. Více je metoda diskutována v [1] a [2].

4.4 Naive Bayes

Metoda je založena na Bayesově teorému [3]. Klasifikace je prováděna na základě následujícího vzorce:

$$v_{NB} = \arg\max_{v \in V} P(v_j) \prod_i P(a_i \mid v_j),$$

kde a_i jsou termy nově zařazovaného dokumentu a v_j jsou slova zkoumané třídy, do které se snažíme dokument zařadit.

Jak bude vidět z výsledků testů, je přesnost metody velice nízká. Tak nízká přesnost je způsobena malým překrytím množin lemmat obou korpusů. V takovém případě není splněn předpoklad, že testovací a trénovaní data obsahují statisticky shodná data. Metoda Naive Bayes poskytuje v případě křížového testu systematicky chybný výsledek. Tento problém lze eliminovat shlukováním podobných synsetů, což je postup popsaný v kapitole 6.

5. Výsledky testů

Analýzou textů jsme dospěli k názoru, že obě části korpusu se liší ve své celkové skladbě. Zatímco ČTK je poměrně obecný zdroj informací, Reuters je zaměřený na burzovní zprávy a například v kategorii počasí není výjimkou rozsáhlé hodnocení ekonomických dopadů živelné katastrofy, což v českém korpusu není možné najít. Tato vlastnost právě způsobila špatné výsledky u klasifikační metody Naive Bayes.

Testy byly prováděny pomocí dříve vytvořených klasifikačních metod a bylo použito několik druhů předzpracování vstupních dat. Nejdříve jsme testovali referenční

případ (viz tab. 1, řádka 1, 2, 3, 4; sloupec *monolinguální*), kdy jsou zprávy v korpusech klasifikovány odděleně (jako dva jednojazyčné korpusy).

Dalším krokem bylo aplikování lemmatizace založené na využití EWN na jednojazyčný korpus (viz tab. 1, řádka 1, 2, 3, 4; sloupec *multilinguální*), čímž jsme ověřili korektní fungování lemmatizátoru a indexace. Jak je vidět z tabulky, jsou výsledky srovnatelné s referenčními.

Ve třetím kroku jsme provedli klasifikaci vícejazyčné kolekce. Nejdříve byl uvažován případ, kdy v testovacích i trénovacích datech jsou oba jazyky zastoupeny v náhodném poměru (viz tab. 1, řádka 5, 6, 10). To simuluje situaci, kdy jsou například v digitální knihovně již zatříděné texty různých jazyků a probíhá automatická klasifikace nově příchozích dokumentů. Jinak řečeno, ekvivalentní překlady mají v různých jazycích jiné indexy. Pro případ klasifikační úlohy je takový přistup možný a poskytuje smysluplné výsledky. Pro klasifikační úlohu bylo dosaženo výsledků srovnatelných s referenčními. Test jsme opakovali na korpusech lemmatizovaných klasickou metodou (slovníkovou) i pomocí metody s využitím EWN.

Posledním testem byla křížová klasifikace (viz tab. 1, řádka 7, 8, 9), kdy v archivu předpokládáme zaklasifikovaná data jednoho jazyka a snažíme se zatřídit data jiného jazyka, tj. trénovací jazyk klasifikátoru je jiný než testovací. V tomto případě velmi záleží na velikosti shody obou korpusů a některé metody jsou na tento faktor velice citlivé (Naive Bayes).

Typickou přesnost a úplnost klasifikace lze odhadovat v rozmezí třetího a čtvrtého testu, tedy mezi 80 – 95%.

			monolinguální		**multilinguální**	
	metoda	**data**	**Přesnost [%]**	**Úplnost [%]**	**Přesnost [%]**	**Úplnost [%]**
1	**NBCI**	**cz**	90.53	93.83	91.28	93.41
2	**NB**	**cz**	95.36	95.57	92.44	93.15
3	**NBCI**	**eng**	95.11	95.47	96.04	96.20
4	**NB**	**eng**	96.85	96.91	94.79	95.17
5	**NBCI**	**cz+eng**	86.75	92.06	86.05	89.52
6	**NB**	**cz+eng**	95.25	95.46	92.04	92.83
7	**NBCI křížově**	**cz+eng**	-	-	80.93	89.42
8	**NB křížově**	**cz+eng**	-	-	3.42	3.42
9	**Itemsets křížově**	**cz+eng**	-	-	73.78	81.49
10	**Itemsets**	**cz+eng**	75.76	81.91	78.65	84.90
11	**TFIDF**	**cz+eng**	93.37	93.37	92.79	92.79

Tab. 1 Výsledky klasifikačních testů

6. Možnosti zdokonalení

V této části zmíníme dvě možnosti, které na základě zatím předběžných experimentů slibují zlepšit výsledky klasifikace. První možností je začlenění disambiguace (zjednoznačnění) slov do procesu klasifikace, druhou je úprava seskupení synsetů v EWN.

6.1 Disambiguace

Poměrně samostatnou úlohou, kterou je nutné řešit při indexaci, je disambiguace [4]. Po provedení lemmatizace mají některá slova shodné základní tvary, ale jejich význam je odlišný. Jedním z příkladů může být slovo kohout, kde není bez znalosti kontextu zřejmé, zda se jedná o kohout – uzávěr, nebo kohout – živočich. Pro rozhodnutí významu je nutná znalost kontextu, ve kterém se slovo nachází. Pro disambiguaci existuje několik algoritmů.

Jako výchozí metodu jsme zvolili disambiguaci s učitelem, konkrétně Bayesovskou disambiguaci. Jedná se o poměrně jednoduchou metodu, která poskytuje kvalitní výsledky s přesností přes 90 %. Jistou nevýhodou metody je veliká závislost na trénovacím korpusu. Pro pokusy s disambiguátorem jsme pro trénování zvolili Brownův korpus, který je volně k dispozici na stránkách EWN. Jedná se o označkovaný korpus anglických textů, kterým jsou již přiřazeny indexy synsetů EWN. Daný korpus jsme použili pro natrénování bayesovského disambiguátoru a testovali jsme přesnost na tomtéž korpusu i na vybraných případech vět z jiných kolekcí. V bayesově disambiguátoru jsme provedli několik vylepšení – použití kontextového okénka a jeho modifikace, využití syntaktických vztahů mezi slovy a použití váhové funkce pro zvýhodnění blízkých slov víceznačnému slovu. Vylepšení zvýšila přesnost disambiguace od 1 do 3 %.

V případě použití disambiguace na věty z jiného než trénovacího korpusu se kvalita významně zhorší, což je způsobeno nedostatečným rozsahem trénovacího korpusu a tím, že neobsahuje veškerá víceznačná slova. Přesnost disambiguace se pak pohybovala pouze okolo 50 %.

Do budoucna uvažujeme o použití paralelních korpusů pro disambiguaci s využití tezauru EWN. Vycházíme z předpokladu, že slova, která jsou víceznačná v jednom jazyce, jsou jednoznačná v jiném. Srovnáním kolokací v různých jazycích můžeme víceznačná slova disambiguovat.

6.2 Shlukování synsetů v EWN

Při testování klasifikátoru a návrhu indexátoru jsme zjistili, že jemnost členění jednotlivých významů slov v EWN je příliš veliká. Tudíž disambiguace pracuje s omezeným trénovacím korpusem na velkém počtu tříd, což snižuje její přesnost.

Druhý problém související s jemností EWN spočívá v navazování jednotlivých jazyků na jednotný index ILR. Přílišná jemnost vede k situaci, kdy se slovo podobného významu převádí na index, který nemá v jiném jazyku přesný ekvivalent, což činí problémy zejména při křížové klasifikaci. Vlastnost se projevila především při klasifikaci metodou Naive Bayes, kde byl velice malý průnik shodných lemmat jedné třídy v české a anglické části korpusu.

Pro eliminaci tohoto nedostatku je nutné vybrané synsety seskupit do větších celků a těmto shlukům přiřadit nové indexy. Tím bude množství klasifikačních tříd menší při zachování významu jednotlivých slov. Shluky lze vytvořit například na základě podobnosti kontextu, ve kterém se slova nacházejí.

Jinou možností je modifikovat klasifikační metody tak, že bude pro jedno slovo definováno více indexů s váhou. Tzn. slovo bude reprezentováno množinou slov (indexů). Váha slova s příbuzným významem se bude snižovat se vzdáleností od slova, které odpovídá lemmatu. Vzdálenosti lze určit průchodem v grafu EWN přes slova nadřazeného významu.

7. Závěr

Použitím tezauru EWN na klasifikační úloze bylo dokázalo, že se jedná o použitelnou metodu, která poskytuje slibné výsledky. Přestože je přesnost klasifikace vícejazyčných korpusů pochopitelně o něco nižší než při klasifikaci jednojazyčných korpusů, lze považovat metodu za kvalitní s přesností pohybující se okolo 90%. Po zahrnutí navrhovaných vylepšení lze očekávat další zvýšení přesnosti. V následujících měsících hodláme zahrnout navrženou klasifikaci do systému vyhledávání v multilinguálních korpusech.

Reference

1. Hynek J., Ježek K. Automatic dokument classification using Itemsets Metod, its modification and evalution. Sborník konference Datacon 2001 Brno, Mária Bieliková (Ed.), ISBN 80-227-1597-2
2. Hynek J., Jezek K. Dokument Classification Using Itemsets. Sbornik konference MOSIS 2000, ISBN: 80-85988-45-3
3. Manning, C. D.., Hinrich S., Foundations of Statistical Natural Language Processing, The MIT Press, 2000
4. Brown, P.F., S. A. Della Pietra, V. J. Della Pietra a R. L. Mercer, Word-Sense Disambiguation Using Statistical Methods, Berkeley, 1991
5. Maas D., Ripplinger B., The AutIndex System, http://www.iai.uni-sb.de/docs/autindex2.pdf
6. EuroWordNet, http://www.illc.uva.nl/EuroWordNet/
7. Ispell, http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html
8. Hajic, J., Morfologický analyzátor, http://quest.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Morphology/index.html
9. Kučera, M., Ježek, K., Hynek, J., Kategorizace textů metodou NBCI, Znalosti 2004, ISBN 80-248-0456-5
10. Joachims, T., A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, International Conference on Machine Learning (ICML), 1997

Příspěvek vznikl za částečné podpory výzkumného záměru MSM 235200005.