Západočeská univerzita v Plzni
Fakulta aplikovaných věd

Vyhledávání a automatická sumarizace textů v multilinguálním prostředí

Závěrečná zpráva o řešení projektu FRVŠ 1326/2005/G1

Ing. Josef Steinberger

Ing. Michal Toman

Doc. Ing. Karel Ježek, CSc.

Obsah

1 Úvod

V současné době se častěji objevuje nutnost uchovat a počítačově zpracovávat dokumenty, které jsou sice 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 buď 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. 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. Za předpokladu, že uživatel zná několik jazyků, je vhodné umožnit jedním dotazem vyhledat více relevantních dokumentů.

V rámci projektu jsme navrhli metody multilinguálního vyhledávání obohacené o automatickou sumarizaci vyhledaných textů. Sumarizace umožňuje lepší a rychlejší orientaci uživatele ve vyhledaných dokumentech. Dále jsme vyvinuli systém MUSE (Multilingual Search and Extraction), ve kterém jsme implementovali a otestovali nově navržené metody. Jádrem vyhledávání je thesaurus EuroWordNet [11] a sumarizátor je založen na latentní sémantické analýze [3]. V předkládané práci jsme se zaměřili na zpracování anglického a českého jazyka, nicméně princip zpracování zůstává stejný i pro ostatní jazyky poskytované tezaurem EWN. U některých jazyků je kompletnost tezauru nižší než v případě angličtiny, předpokládáme však, že bude EWN postupně doplňován.

V další kapitole lze nalézt cíle řešení projektu s odkazy na popis jejich vlastního řešení. V třetí kapitole popíšeme systém MUSE, jeho architekturu a všechny jeho moduly. Dále uvedeme výsledky testování s popisem kolekcí textů a nakonec nastíníme otevřené problémy. V příloze lze nalézt výši finančních prostředků a zdůvodnění na co byly použity, změny řešení proti projektu a údaje o jejich schválení. Poslední přílohou je popis elektronických příloh.

2 Cíle řešení projektu

Předkládaný projekt navázal na dosavadní výsledky výzkumu v oblasti získávání znalostí z textových databází. Cílem projektu bylo vyvinout systém umožňující vícejazyčné vyhledávání v rozsáhlých textových databázích, případně digitálních knihovnách. Pro lepší orientaci ve vyhledaných dokumentech měl systém umožňovat jejich sumarizaci. V systému měly být použity moderní metody pro dolování dat.

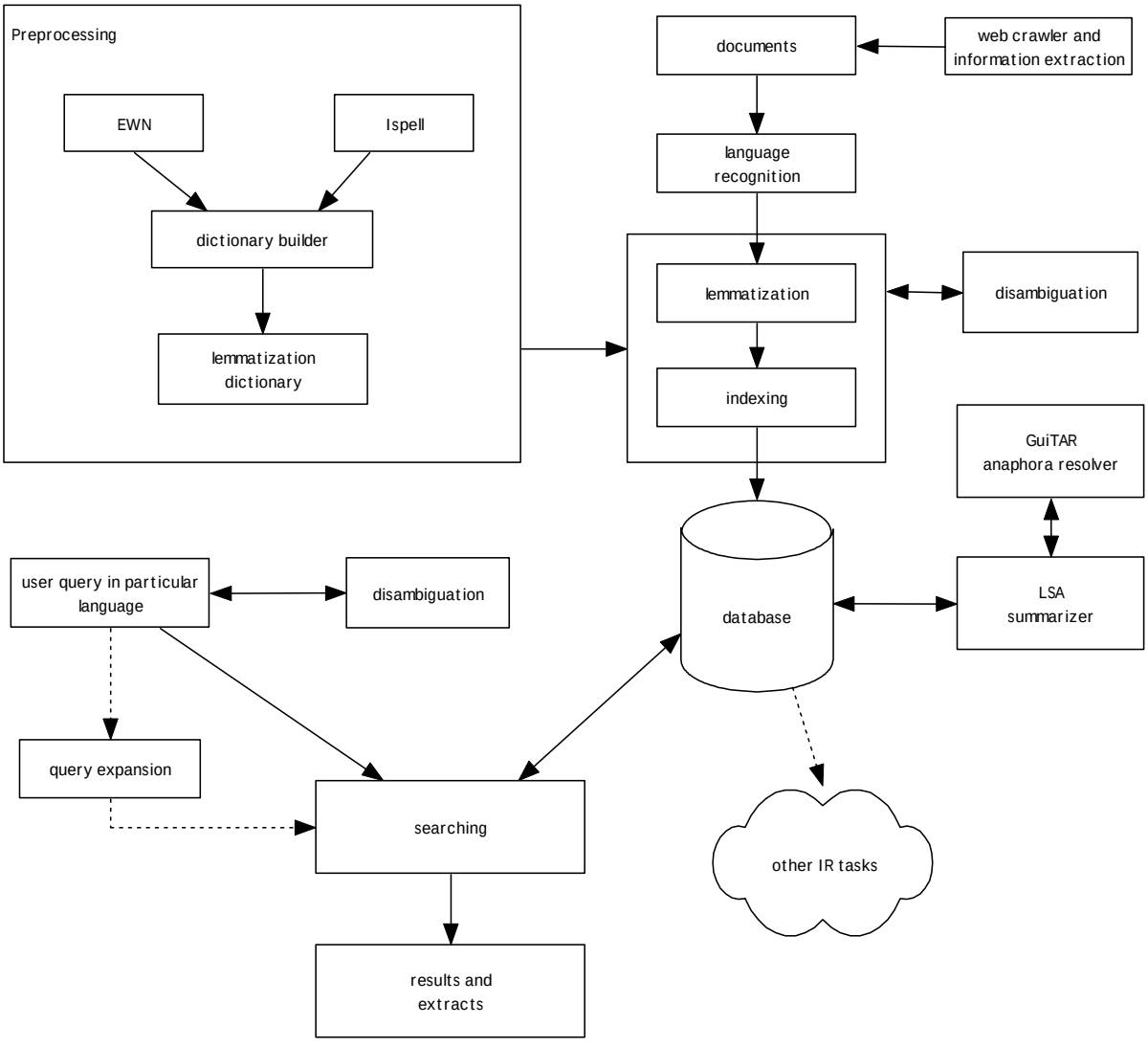
Hlavními cíli předkládaného projektu byly:

- Analýza stávajícího stavu řešené problematiky.
- Navržení a vývoj metod pro podporu vyhledávání ve vícejazyčných textových databázích (viz popis jednotlivých modulů – 3.2.1-3.2.6).
- Vývoj moderních metod automatické sumarizace s respektováním multilinguality databáze (viz 3.2.7).
- Implementace navrženého systému vyhledávání a sumarizace (viz 3).
- Experimentální ověření navržených metod na anglických a českých textech (viz 4, 3.2.7 + přiložené publikace).
- Rozšíření problematiky na oblast WWW (viz 4.4).

3 Systém MUSE

MUSE [10] se skládá z několika relativně samostatných modulů (viz obrázek 1), které jsou dále popsány. V závěru kapitoly lze najít ukázku grafického uživatelského prostředí systému MUSE.

3.1 Architektura



Obrázek 1: Architektura systému MUSE

3.2 Moduly

3.2.1 Rozpoznání jazyka

V multilinguáním prostředí, kde jsou jednotlivé dokumenty psány různými jazyky, vzniká nutnost vytvoření modulu rozpoznávající jazyk. V některých jazycích (např. češtině) je navíc nutné rozpoznat kódování jazyka (např. Windows 1250, ISO 8859-2). Modul rozpoznávající jazyk je tudíž zřejmá součást výpočetního systému, který se zabývá zpracováním multilinguálních textů.

K rozpoznání jazyka jsme použili dva rozdílné přístupy [2]. První z nich je založený na zjišťování frekvence písmen ve slovech. Každé písmeno se v textu vyskytuje s určitou (obecně rozdílnou) pravděpodobností výskytu. Tyto frekvence jsme použili jak pro určení jazyka, tak pro určení kódování. V případě rozpoznávání kódování se v kódových stránkách liší pouze několik písmen, přesto je stále možné podle nich kódování rozpoznat.

Druhý přístup je založený na použití tzv. stoplistu (seznam slov, která nemají konkrétní sémantickou informaci). Tato slova jsou unikátní pro každý jazyk a představují dobré vodítko pro řešení našeho problému. Slova obsažená ve stoplistu jsou například: a. an, the, of, in v případě angličtiny. Předběžné výsledky ukazují že metoda založená na stop-slovech poskytuje lepší výsledky než frekvenční metoda.

	138	141	142	154	157	158	169	171	174	185	187	190
CP1250	Š	Ť	Ž	š	ť	ž	©	«	®	ą	»	ľ
ISO-8859-2	□	□	□	□	□	□	Š	Ť	Ž	š	ť	ž

Tabulka 1: Ukázka odlišností znaků v jednotlivých kódováních češtiny.

Pro každé kódování a jazyk je definován binární vektor s jedničkami na místech písmen, která jsou charakteristická pro daný jazyk a kódování. Při zpracování je vytvořen binární vektor odpovídající analyzovanému textu, který je následně porovnán se vzory jednotlivých kódování a jazyků. Pro porovnání podobnosti vektorů je použita Hammingova vzdálenost – kolik bitů se v binárních řetězcích liší. Čím nižší je Hammingova vzdálenost, tím pravděpodobnější je daný jazyk a kódování. Vzdálenost je nuné znormovat podle počtu charakteristických písmen v jednotlivých jazycích (např. 28 pro češtinu, 34 pro slovenštinu, 7 pro němčinu).

3.2.2 Lemmatizační modul

K vytvoření lemmatizačního slovníku jsme zvolili extrakci tvarů slov z programu Ispell [12]. 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 použít 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.

Dalšího vylepšení bylo dosaženo využitím morfologického analyzátoru [13]. 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.

3.2.3 Thesaurus EWN

Jako základní stavební kamen pro řešení problému vícejazyčnosti dokumentů jsme zvolili tezaurus EuroWordNet [11]. Tezaurus zajišťuje provázání termů jednotlivých jazyků a umožňuje pracovat se synonymi jazyka.

Tezaurus EuroWordNet (EWN) obsahuje slova uspořádaná do množin synonym (tzv. synsetů). V každém synsetu jsou slova podobného významu a mají přiřazenou unikátní značku (index), která je shodná ve všech jazycích EuroWordNetu. Tezaurus obsahuje následující evropské jazyky: angličtina, dánština, italština, španělština, němčina, francouzština, čeština, estonština. EWN je strukturovaný podobně jako původní Wordnet vytvořený Princetonskou univerzitou.

Jelikož jsou značky pro jednotlivá slova shodné v různých jazycích, lze vhodně navrženým systémem provádět také křížové vyhledávání – tzn. trénování provést na kolekci v jednom jazyce a rozlišovat významy slov v jiném jazyce. Taková vlastnost je výhodná především v případě jazyků, pro které nemáme dostatek trénovacích dat.

V současné době nejsou jednotlivé wordnety stejně rozsáhlé. Některé synsety nemají v určitých jazycích odpovídající ekvivalent, což lze chápat jako předmět dalšího doplnění.

3.2.4 Provázání lemmatizačních slovníků s 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.

3.2.5 Disambiguace

Rozlišení významů slova je nezbytným krokem pro většinu aplikací zpracovávajících přirozený jazyk (Natural Language Processing, NLP). Jedná se o klíčovou úlohu pro správné porozumění sdělení, uplatňuje se v komunikaci člověk-počítač. Jako příklad lze uvést automatický překlad, kde se disambiguace využívá pro nalezení správné interpretace víceznačného slova. Mějme anglické slovo *bank*, jenž lze přeložit mimo jiné jako *břeh* nebo *banka*. Správný překlad vyplývá z kontextu, ve kterém je slovo použito a je zřejmé, že překlady nelze zaměňovat.

Disambiguaci v této zprávě chápeme jako klasifikaci víceznačného slova do tříd, které představují vždy jeden význam slova. Možné významy jsou typicky vyjmenovány ve slovníku, kde mohou být uvedeny i doplňující atributy pomáhající disambiguaci (např. synonyma, vztahy k ostatním slovům, slovní druh, apod.). Zařazení slova do třídy je ovlivněno jeho kontextem, případně další informací získanou ze slovníku, tezauru, encyklopedie, či jiného lexikálního zdroje. Často je jako zdroj informací použitý tezaurus EuroWordNet (EWN).

Z výsledků analýzy textů jsme zjistili, že téměř 20% slov je víceznačných. To poukazuje na důležitost disambiguace při zpracování přirozeného textu prakticky ve všech oblastech NLP.

Disambiguační metody lze dělit podle způsobu trénování. V případě, že máme k dispozici označkovaný korpus, mluvíme o metodách s učitelem. V označkovaném korpusu má každé víceznačné slovo přidruženou značku, která určuje jeho význam. Korpus se ve většině případů značkuje ručně.

Druhou skupinou disambiguačních algoritmů jsou metody bez učitele. Není nutná žádná apriorní informace o významech jednotlivých slov, tedy odpadá nutnost značkování trénovacího korpusu. Takovou disambiguaci lze považovat za úlohu shlukovaní víceznačných slov podle významů.

Rozhodli jsme se zaměřit především na metody s učitelem. Cílem bylo vytvořit disambiguátor, který dokáže víceznačným slovům přiřadit značku (index) uvedenou v tezauru EWN. V případě použití metod bez učitele není možné takového výsledku dosáhnout, jelikož není zřejmé, jaké jsou vazby mezi shluky získanými při disambiguaci a jejich indexy v EWN.

Základní metodu bayesovské disambiguace jsme se snažili modifikovat tak, aby byly minimalizovány její nedostatky. Tomuto tématu se věnuje publikace [9]. Především jsme zvýhodnili některá slova z kontextu disambiguovaného a naopak potlačili taková, která nepřináší žádnou informaci použitelnou pro rozlišení významu víceznačných slov.

3.2.6 Vyhledávací modul

Vyhledávací systém se zabývá mimo jiné reprezentací, uložením, organizací a přístupem k datovým položkám. Tradiční vyhledávací systémy se typicky snaží dokumentu přiřadit charakterizující indexové termy, tzn. každému dokumentu zvolí množinu klíčových slov. To předpokládá že sémantika dokumentu může být vyjádřena přirozeným způsobem množinou indexů. V případě nahrazení dokumentu několika indexy dojde k přílišnému zjednodušení, protože se mnoho sémantické informace ztratí. Navíc je pro uživatele obtížné formulovat v takovém systému korektně dotaz, což má často za následek nerelevantní odpovědi systému.

V literatuře se uvádí tři klasické vyhledávací modely – booleovský, vektorový a pravděpodobnostní. V booleovském modelu jsou termy dokumentů a dotazů reprezentovány jako množiny indexů. Ve vektorovém modelu jsou zaznamenány jako vektory v konečném n-dimenzionálním prostoru. V pravděpodobnostním modelu tvoří kostru dokumenty a dotazy založené na teorii pravděpodobnosti. Samozřejmě bylo navrženo mnoho alternativních vyhledávacích modelů, které více, či méně z již zmíněných modelů vycházejí. Jedná se například o fuzzy a rozšířený booleovský model. Dále můžeme rozlišovat modely založené na latentním sémantickém indexování, případně založené na využití neuronových sítí. Ve všech těchto modelech vyjadřuje uživatel své požadavky pomocí dotazu daného termy.

Pro realizaci navrhovaného systému jsme zvolili modifikovaný vektorový model, kde byl jako hodnotící algoritmus použitý TF-IDF. Dokumenty jsou rozlišeny podle jazyka, naindexovány a uloženy do databáze. Následně může uživatel provádět dotazy v libovolném jazyku podporovaném systémem. Vyhledané výsledky mohou být omezeny na jeden nebo více

jazyků. Ke každému textu je vyhledán také automaticky vytvořený abstrakt, který umožňuje jednodušší orientaci ve výsledcích.

Součástí vyhledávacího systému je také modul rozšíření dotazu, který může uživatelský dotaz rozšířit, aby vyhledávací algoritmus poskytoval více relevantních odpovědí. Pro rozšíření dotazu je použitý tezaurus EWN, který obsahuje sémantické vztahy mezi jednotlivými synsety. Pro rozšíření jsou použity následující vztahy slov s významem:

1. nadřazený,
2. podřazený,
3. podobný.

3.2.7 Sumarizační modul

Metody automatické sumarizace

Automatická sumarizace textů je vědecká disciplína, která v dnešní době poutá stále větší pozornost. Obrovské množství elektronických informací se musí redukovat, aby se s nimi uživatel mohl efektivněji vypořádat. Existuje mnoho sumarizačních metod od extraktivních[1] až po generativní[2]. Náš přístup je extraktivní a dále patří do kategorie metod založených na identifikaci nejvýznamnějších termů dokumentu. Nejdůležitější informace o těchto termech je následně extrahována. Tyto metody lze dále rozdělit na lexikální a koreferenční. Lexikální přístup používá podobnost slov a jiné lexikální vztahy k identifikaci centrálních termů v textu. Koreferenční přístupy identifikují tyto termy pomocí systému pro rezoluci anafor, který je schopen vyhledat anafory[3], sestavit z nich anaforické řetězce, které potom představují centrální termy textu. Náš sumarizační systém využívá jak lexikální, tak i anaforické informace. Jeho jádro tvoří latentní sémantická analýza (LSA).

Sumarizace založená na LSA

LSA [3] je technika pro extrakci skrytých dimenzí sémantických reprezentací termů, vět nebo dokumentů na základě jejich výskytu v kontextu. Byla již aplikována na řadu problémů zpracování přirozené řeči, např. vyhledávání nebo segmentace textu. V poslední době bylo navrženo použití LSA pro sumarizaci [1]. Tato čistě lexikální metoda je startovacím bodem naší práce. Základem je získání latentní reprezentace dokumentu ve dvou krocích. Nejprve se vytvoří matice termů proti větám. Každý prvek matice udává zda se určitý term vyskytuje v určité větě. Dalším krokem je aplikace singulární dekompozice (SVD = Singular Value

[1] Jejich cílem je vybrat nejvýznamnější věty v sumarizovaném textu a výsledkem je tedy extrakt.
[2] Generují nové věty a výsledkem je abstrakt.
[3] Anafora je slovo nebo fráze, která se odkazuje zpět na nějaké slovo, frázi nebo myšlenku v textu.

Decomposition) na předem vytvořenou matici. Detailní matematický proces lze nalézt v [3][5][6][7]. Důležitou vlastností SVD je schopnost zachytit vzájemné vztahy mezi termy, takže termy a věty jsou shlukovány na základě sémantiky do témat. Gong a Liu (2002) [1] navrhly metodu, která vybere do extraktu pro nejvýznamnější témata textu vždy jednu nevýznamnější větu. To však přináší řadu problémů. Hlavním problémem je, že se musí určit takový počet témat, jako je počet vět, které chceme do extraktu zařadit. Výsledkem tohoto požadavku je, že extrakt potom může obsahovat věty o tématech, které nejsou důležité. V naší modifikované metodě [6] jsme změnili kritérium pro výběr vět. Myšlenkou je vybrat věty s největší váhou přes všechny významná témata. Navržený algoritmus však potřebuje metodu pro zjištění počtu významných témat. Bližší podrobnosti o této metodě lze nalézt v [5].

Užití anaforické informace k upřesnění LSA témat

Metoda popsaná v předchozím odstavci určuje hlavní témata dokumentu na bázi nejjednodušší představy termů – slov. V [7] jsme navrhli jak integrovat informace o anaforách do procesu získávání LSA témat. Pro automatické získání anafor (a jejich referentů) z textu jsme použili systém GuiTAR [4], který byl vyvinut na univerzitě v Essexu (Anglie). Dále jsme navrhli metodu jak skombinovat získané informace o anaforách se základními lexikálními. Z anafor se nejprve vytvoří anaforické řetězce, které reprezentují určité entity zmiňované v textu. Reprezentace věty potom nespecifikuje pouze, zda obsahuje určitá slova, ale navíc zda obsahuje zmínku o určité entitě (tj. zda obsahuje anaforický řetězec). Tyto větné reprezentace se pak použijí pro vytvoření matice pro SVD. Detailnější popis lze získat v [5][6].

Výsledky testování sumarizace

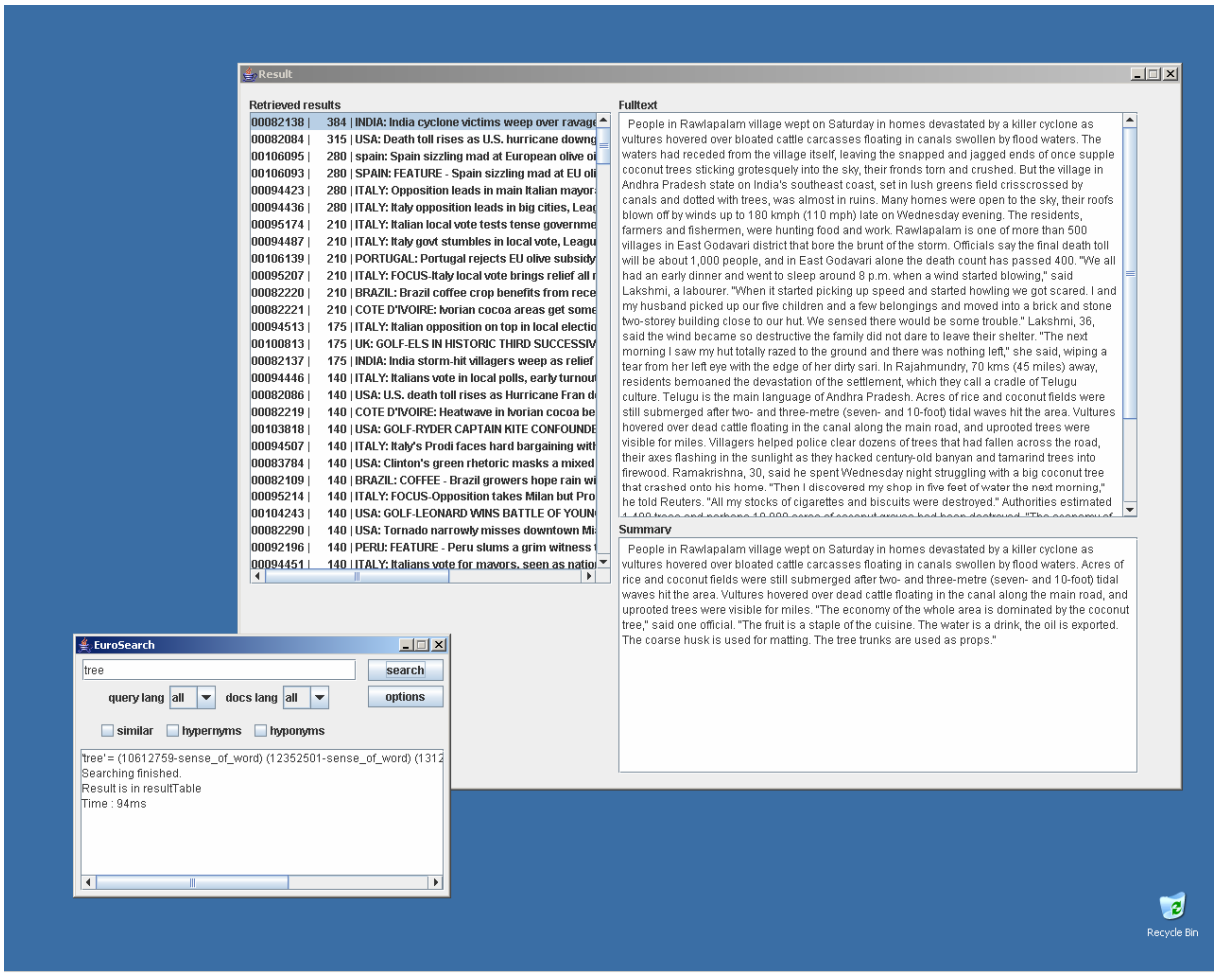
Pro testování jsme použili korpus DUC 2002 [14], který obsahuje 567 dokumentů z různých zdrojů. Pro každý dokument jsou zde k dispozici 2 abstrakty o velikosti 100 slov, vytvořené anotátory. Korpus navíc obsahuje výsledné extrakty/abstrakty 13 sumarizačních systémů, které se zúčastnili konference DUC 2002[4]. To umožňuje porovnat naše extrakty s jinými sumarizačními systémy. Jako porovnávací metrika se běžně používá ROUGE [5], který měří podobnost extraktu s referenčními abstrakty na základě n-gramů. Výsledky porovnání ukázaly, že naše modifikovaná sumarizační metoda, která používá pouze lexikální informace,

[4] DUC (Document Understanding Conference) slouží k porovnání sumarizačních systémů a k prezentaci výsledku výzkumu v oblasti sumarizace.

dává lepší výsledky než původní metoda (Gong a Liu). Přidáním anafor dále úspěšnost stoupla. Celkově se náš systém umístil na třetím místě ze 17, což je slibný výsledek. Podrobnosti i se statistickými významnostmi lze nalézt v [5].

3.3 GUI systému

Grafické uživatelské prostředí se skládá ze dvou oken. V prvním (obrázek 2 – vlevo dole) uživatel zadává dotaz a v druhém se zobrazují výsledky vyhledávání (levá část okna), plný text (vpravo nahoře) a extrakt (vpravo dole).



Obrázek 2: GUI systému MUSE

4 Testování

4.1 Testovací korpus

Pro testování systému jsme vybrali přes 100 tisíc vhodných novinových článků, které byly získány z dvou tiskových agentur. Pro české texty byla použita databáze tiskových zpráv ČTK s celkovým počtem 82000 dokumentů, jako zdroj anglických textů jsme použili články tiskové agentury Reuters (Reuters Collection Volume 1), kde jsme vybrali 24000 nejdelších textů.

Lemmatizační slovníky čítají 3330000 slov v prípadě českého jazyka a 119000 slov u anglického jazyka. Po provedení lemmatizace jsou termy namapovány do 115000 synsetů v případě angličtiny a 27000 synsetů pro češtinu.

4.2 Přesnost vyhledávání

Přesnost vyhledávání jsme ověřovali na prvních 30 nejvíce relevantních dokumentech pro dotaz zadaný uživatelem a jeho rozšíření.

dotaz	bez rozšíření		rozšíření o podobné			rozšíření o podřazené			rozšíření o nadřazené			celkové rozšíření			přesnost bez rozšíření (%)	přesnost s celkovým rozšířením (%)
	celk.	rel.	celk.	rel.	nové	celk.	rel.	nové	celk.	rel.	nové	celk.	rel.	nové		
formula * one * champion	88	27	88	27	0	88	27	0	465	26	4	465	26	4	90,0	86,7
terorismus * útok	265	29	265	29	0	265	29	0	300	29	2	300	29	2	96,7	96,7
white * house * president	2393	29	2657	28	0	2393	29	0	5880	23	3	6116	23	2	96,7	76,7
povodeň * škody	126	29	126	29	0	126	29	0	126	29	0	126	29	0	96,7	96,7
cigarettes * health	366	25	366	25	0	366	25	0	393	25	2	393	25	2	83,3	83,3
rozpočet * schodek	2102	30	2102	30	0	2102	30	0	2174	30	4	2174	30	4	100,0	100,0
plane * crash	221	29	221	29	0	221	29	0	2306	29	8	2306	29	8	96,7	96,7

Tabulka 2: Přesnost vyhledávání a vliv rozšíření dotazu na přesnost.

4.3 Srovnání s Googlem

Pro ověření výsledků vyhledávání jsme zvolili srovnání s již existujícími vyhledávacími systémy. Jsou to Google Desktop Search (všeobecne uznávaný jako jeden z nejlepších

vyhledávacích systémů) a Copernic Desktop Search. Cílem testu bylo srovnání výsledků systému MUSE s/bez rozšíření dotazu a existujících komerčních systémů.

dotaz	MUSE			MUSE s rozšířením dotazu			Google desktop search		Copernic desktop search	společné nejohodnocenější MUSE a GoogleDS (%)	
	spol. 30	spol. 10	celk.	spol. 30	spol. 10	celk.		celk.	celk.	10	30
formula * one	25	9	351	24	9	2075	30	282	395	90	83,33
national * park	9	3	508	9	3	1198	30	266	390	30	30,00
religion * war	20	7	73	20	7	74	30	84	140	70	66,67
water * plant	11	7	73	6	4	1489	30	45	134	70	36,67
hockey * championship	20	7	82	20	7	85	30	61	81	70	66,67
traffic * jam	18	6	64	16	6	165	28	28	86	60	64,29
heart * surgery	16	7	563	17	7	703	30	550	572	70	53,33
weather * weekend	19	10	140	19	10	158	30	134	142	100	63,33

Tabulka 3: Porovnání MUSE a komerčních systémů.

4.4 Diskuse

Jak vyplývá z tabulek, dosavadní výsledky jsou slibné s přesností vyhledaných dokumentů přes 80%. Při rozšíření dotazu se podle očekávání mírně zhorší přesnost, ale vzroste úplnost vyhledávání. Systém jsme také vyzkoušeli na kolekci webových dokumentů. Dosáhli jsme podobných výsledků jako na kolekcích ČTK a Reuters. Nasazení systému do prostředí webu navíc znamená stahování a předzpracování dokumentů. Tyto moduly jsou již v systému rovněž implementovány.

5 Závěr a otevřené problémy

Náš systém poskytuje výsledky srovnatelné s obecně uznávaným systémem Google DS, který je považován za jeden z nejlepších vyhledávacích strojů. Avšak MUSE má několik výhod ve srovnání s Googlem. Za prvé, náš systém respektuje vícejazyčnost prostředí. Pokud vložíme dotaz v angličtině, Google nemůže najít relevantní dokumenty napsané v jiných jazycích. Naopak MUSE vyhledá dokumenty ve všech požadovaných jazycích. Za druhé, synonyma (např. *car* / *auto*) jsou považována za rovnocenná při vyhledávání. Dále lze dotaz rozšiřovat o související synsety. Například jeden z nejrelevantnějších dokumentů pro dotaz „*strom*“ pojednával o „*kalamitě borového lesa*“, přestože se v něm nevyskytovalo slovo dotazu ani jednou. A nakonec, částí systému je automatický sumarizátor.

Na částech projektu se podíleli studenti v rámci svých diplomových prací.

Reference

[1] Y. Gong a X. Liu.: Generic text summarization using relevance measure and latent semantic analysis. V *Proceedings of ACM SIGIR*, New Orleans, USA, 2002.

[2] Karel Ježek a Michal Toman: Documents Categorization in Multilingual Environment. V *Proceedings of ELPUB'05*, Leuven, Belgie, 2005.

[3] T. K. Landauer a S. T. Dumais: A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

[4] M. Poesio a M. A. Kabadjov: A General-Purpose, offthe-shelf Anaphora Resolution Module Implementation and Preliminary Evaluation. V *Proceedings of LREC*. Lisabon, Portugalsko, 2004.

[5] Massimo Poesio, Josef Steinberger, Mijail A. Kabadjov a Karel Ježek: An Approach to Summarization Combining Anaphoric and Lexical Knowledge within the LSA Framework. Podáno na *EACL'06*, Trento, Itálie, 2006.

[6] Josef Steinberger a Karel Ježek: Text Summarization and Singular Value Decomposition. V *Proceedings of ADVIS '04, Springer Verlag* , Izmir, Turecko, 2004.

[7] Josef Steinberger, Mijail A. Kabadjov, Massimo Poesio a Olivia Sanchez-Graillet: Improving LSA-based Summarization with Anaphora Resolution. V *Proceedings of EMNLP'05*, Vancouver, Kanada, 2005.

[8] Michal Toman a Karel Ježek: Klasifikace multilinguálních korpusů s využitím tezauru EuroWordNet. V *Proceedings of ITAT 2003*, Slovensko.

[9] Michal Toman a Karel Ježek: Modifikace bayesovského disambiguátoru. V *Znalosti 2005*, Stará Lesná, Slovensko, 2005.

[10] Michal Toman, Josef Steinberger a Karel Ježek: Searching and Summarizing in Multilingual Environment. Podáno na *ELPUB'06*, Bansko, Bulharsko, 2006.

[11] http://www.illc.uva.nl/EuroWordNet/

[12] http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html

[13] http://quest.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Morphology/index.html

[14] http://www-nlpir.nist.gov/projects/duc/data.html

Přílohy

1. Použití finančních prostředků

Stipendia MŠMT	
Ing. Steinberger	20.000,-
Ing. Toman	20.000,-
Zahraniční cestovné	
Vložné 3/48/05 – konference ELPUB'05, Leuven, Belgie	8.082,50
Letenka Praha – Vancouver a zpět (konference HLT/EMNLP 2005)	24.885,-
Ubytování – Vancouver	3.032,50
Ostatní	
Kniha – The Oxford Handbook of Computation Linguistics	5.046,80
Materiál, CD, DVD	874,20
Poštovné (knihy)	79,-
==========	==========
Celkem	82.000,-

2. Změny řešení proti projektu

V původním projektu se předpokládalo použití celé částky na zahraniční cestovné pro jednu evropskou konferenci. Z důvodu přijetí příspěvku na prestižní konferenci HLT/EMNLP (Vancouver, Kanada) se částka rozdělila na 2 konference – ELPUB'05 (Leuven, Belgie) a HLT/EMNLP.

24. srpna byla žádost schválena.

3. Elektronické přílohy

Publikace vzniklé v rámci tohoto projektu:

Michal Toman a Karel Ježek: Modifikace bayesovského disambiguátoru. V *Znalosti 2005*, Stará Lesná, Slovensko, 2005. [*znalosti2005.pdf*]

Karel Ježek a Michal Toman: Documents Categorization in Multilingual Environment. V *Proceedings of ELPUB'05*, Leuven, Belgie, 2005. [*elpub2005.pdf*]

Josef Steinberger, Mijail A. Kabadjov, Massimo Poesio a Olivia Sanchez-Graillet: Improving LSA-based Summarization with Anaphora Resolution. V *Proceedings of EMNLP'05*, Vancouver, Kanada, 2005. [*emnlp2005.pdf*]

Massimo Poesio, Josef Steinberger, Mijail A. Kabadjov a Karel Ježek: An Approach to Summarization Combining Anaphoric and Lexical Knowledge within the LSA Framework. Podáno na *EACL'06*, Trento, Itálie, 2006. [*eacl2005.pdf*]

Michal Toman, Josef Steinberger a Karel Ježek: Searching and Summarizing in Multilingual Environment. Přijato na *ELPUB'06*, Bansko, Bulharsko, 2006. [*elpub2006.pdf*]