

Projekt M-CAST: vývoj vícejazyčného systému agregace informací

Vilém Sklenák, Petr Strossa

Katedra informačního a znalostního inženýrství, FIS, Vysoká škola ekonomická v Praze,
nám. W. Churchilla 4, 130 67, Praha 3
{sklenak, kizips}@vse.cz

Abstrakt: Cílem projektu M-CAST je vyvinout vícejazyčný systém umožňující integrovat a prohledávat rozsáhlé kolekce textů (a multimédií). Je popsána struktura navrhovaného systému, dosavadní práce na jeho implementaci a zejména na jeho rozšíření o český jazykový modul.

Klíčová slova: vyhledávání informací, vícejazyčné vyhledávání, lingvistika, ontologie

1 Úvod

Cílem projektu M-CAST¹ (Multilingual Content Aggregation System based on TRUST Search Engine, projekt programu eContent #22 249) je vyvinout vícejazyčný systém, který bude umožňovat tvůrcům obsahu integrovat a prohledávat rozsáhlé kolekce textů (a multimédií), jako jsou internetové knihovny, informační zdroje nakladatelství a tiskových agentur nebo databáze vědeckých informací.

2 O systému M-CAST obecně

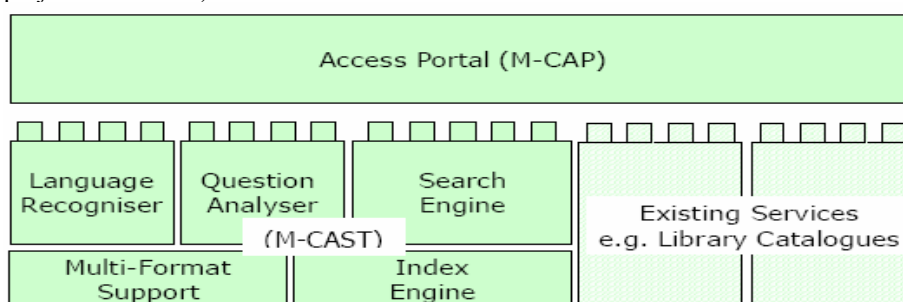
Systém M-CAST umožní tvorbu digitálních knihoven prostřednictvím agregace dat dostupných v různých formátech a z různých zdrojů. Systém bude testován dvěma knihovnami, pro které budou vytvořeny portály agregující multimediální informace (M-CAP, Multimedia Content Aggregation Portals) na základě existujících portálů a systémů. Tyto portály budou schopny hledat odpovědi na dotazy v přirozeném jazyce v rozsáhlých kolekcích vícejazyčných dat. Prezentační vrstva portálu bude multimediální, díky čemuž bude možné zpřístupnit digitalizované kopie starých tisků, právních dokumentů, notových záznamů, obrázků a videa, i když indexovány budou pouze jejich textové popisy.

Systém M-CAST bude založen na výsledcích projektu TRUST – Multilingual Semantic and Cognitive Search Engine for Text Retrieval Using Semantic Technologies² (Vícejazyčný sémantický a kognitivní mechanismus vyhledávání textů využívající sémantické technologie, IST-1999-56416), který byl financován z Pátého

¹ <http://www.m-cast.infovide.pl>

² <http://www.trustsemantics.com>

rámcového programu EU pro rozvoj vědy a výzkumu. Vyhledávací stroj TRUST umí v současnosti vyhledávat ve čtyřech jazycích (francouzštině, italštině, polštině a portugalštině [2]). V rámci projektu bude transformován z jedinouživatelské aplikace pro PC na serverovou aplikaci pro operační systém UNIX nebo Windows. Jazykové zdroje [1] systému TRUST budou aktualizovány a v systému nadále využity. K dosud využívané jazykové ontologii (taxonomii) bude vytvořena alternativa založená na standardním mezinárodním desetinném třídění (MDT), celosvětově používaném v knihovních systémech. Systém bude obohacen o dva další jazyky: angličtinu (bude řešena italským partnerem projektu TRUST) a češtinu (bude doplněna v rámci projektu M-CAST).



Agregační systém M-CAST bude ústředním prvkem agregčního portálu M-CAP a bude vyvinut podle zásad metodologie návrhu systémů pro správu obsahu založených na znalostech (Knowledge-based Content Management Application Design Methodology). Tuto metodologii vytvořila firma Infovide, S.A. v rámci jiného projektu financovaného z Pátého rámcového programu EU – ICONS³ – inteligentní systém pro správu obsahu (Intelligent Content Management System, IST-2001-32429).

Portály budou nasazeny a testovány ve dvou veřejných knihovnách: v Polské internetové knihovně⁴, jejímž provozovatelem je Kopernikova knihovna v Toruni, a v Národní knihovně České republiky v Praze⁵. Cílem je zpřístupnit vícejazyčné digitální informační zdroje pro zodpovídání dotazů v přirozeném jazyce. Výsledkem hledání budou fragmenty textu, které obsahují odpovědi na dotaz.

Výsledkem projektu bude komerční produkt pro management vícejazyčných znalostí. Očekává se komerční využití výsledků v knihovnách (vyhledávání informací, předmětové katalogy), správě sbírek (akvizice a agregace dat, statistiky výpůjček, vyřazování z fondu), bibliografických databázích, informačních službách (selektivní šíření informací, personalizace, objevování znalostí) a v sémantických datových sítích.

Pro potřeby projektu bylo vytvořeno konsorcium těchto subjektů:

- Infovide S.A., Varšava, Polsko (koordinátor projektu)
- Polská internetová knihovna, Toruń, Polsko
- TiP sp. z o.o., Katowice, Polsko
- Synapse Développement SARL, Toulouse, Francie

³ <http://www.icons.rodan.pl>

⁴ <http://www.pbi.edu.pl>

⁵ <http://www.nkp.cz>

- Priberam Informática Lda., Lisabon, Portugalsko
- Expert System S.p.A., Modena, Itálie
- Národní knihovna České republiky, Praha, Česká republika
- Vysoká škola ekonomická v Praze, Česká republika

3 Čeština v systému M-CAST

Lingvistické moduly dosavadního systému TRUST pro jednotlivé jazyky byly řešeny do značné míry nezávisle – obecně sdílely jen základní funkční schéma, které by bylo možné vyjádřit v těchto bodech:

- Texty mají být indexovány jednotlivými slovy, resp. (v případě homonymních / polysémních slov) jejich jednotlivými významy, dále idiomy a jmennými frázemi obsaženými v příslušných slovnících, vlastními jmény obsaženými v příslušných slovnících, pojmenovanými entitami rozpoznávanými podle určitých obecných pravidel, koncepty jazykové ontologie (taxonomie) a jmény domén podle speciálního seznamu.
- Indexovaný text tedy musí být podroben zejména morfologické analýze a lematizaci (což je úkol různě náročný podle konkrétního jazyka); vhodná (avšak nikoli absolutně nezbytná) je i určitá omezená syntaktická analýza – alespoň v takové míře, do jaké se o syntaktické kategorie opírají pravidla rozpoznávání významů homonym a polysémních slov: např. některé významy sloves může být vhodné rozlišovat podle toho, zda je či není přítomen předmět nebo příslovečné určení jistého typu.
- Doporučuje se používání derivačního slovníku (tj. slovníku popisujícího vzájemné odvozování slov různých slovních druhů – např. slova *testovat*, *testování*, *testovací*, *testovaný*, *testující* všechna odkazují na *test*) a tezauru nebo alespoň slovníku synonym.
- Specifickým prvkem každého jazykového modulu je analyzátor typů dotazů (v současnosti je popsáno asi 80 základních typů faktografických dotazů, které se mohou poněkud lišit způsobem vyhledávání potenciálních odpovědí v textech). Rovněž s tímto bodem souvisí potřeba určité omezené syntaktické analýzy. Úlohy dotazování a indexování textů v daném jazyce ovšem mohou být řešeny odděleně a nezávisle – pro daný jazyk není principiálně nutné řešit obě zároveň.

Při realizaci českého lingvistického modulu bylo rozhodnuto v maximální míře využít existující programové kódy pro polštinu, které vyvinula firma TIP, a nově sestavovat pokud možno jen slovníky a datové tabulky ve formátech požadovaných těmito programovými kódy. Dosavadní výsledky potvrzují schůdnost takového postupu. Hypotetická alternativa spočívající v získání a zakomponování jednotlivých modulů, o kterých je známo, že již někde existují (např. pro morfologickou analýzu češtiny), byla zamítnuta jako pravděpodobně celkově komplikovanější.

V současnosti probíhající práce se soustřeďuje zejména na morfologický analyzátor češtiny. Základem v tomto směru je prozatím formalizovaný popis přibližně 300 vzorů skloňování podstatných jmen, 50 vzorů skloňování a stupňování přídavných jmen (zahrnujících i odvozování příslovcí) a 150 vzorů časování sloves. (Každý vzor

je definován určitou – v závislosti na slovním druhu – posloupností ohýbacích koncovek; pro jeden tvar může být specifikováno více alternativních koncovek; v rámci různých vzorů mohou navíc různé tvary vyžadovat různé typy kmenových změn. Principem je, že každé paradigma ohýbání slova odlišující se od jiných je považováno za vzor, i kdyby se jím řídilo jen jediné slovo, nerozlišujeme tedy “pravidelné” a “nepravidelné” způsoby ohýbání slov. Přitom může být shledáno zajímavým, že podle pokusného přiřazení zmíněných vzorů v pracovní databázi celkem 120 000 podstatných jmen, zkompileované z různých volně dostupných zdrojů, existují dva vzory – konkrétně “kost” a “stavení” – pokrývající dohromady 70 % všech podstatných jmen; 11 nejfrekventovanějších vzorů pokrývá skloňování 90 % všech podstatných jmen; pomocí 56 vzorů bychom dokázali popsat skloňování 99 % všech českých podstatných jmen; zhruba polovina z uvedených přibližně 300 vzorů pro podstatná jména popisuje skloňování pouze 1–3 slov, 79 vzorů bylo přiřazeno jen jedinému podstatnému jménu. Podobné výsledky jsme získali i pro přídavná jména a slovesa.)

Zároveň aktuálně probíhají přípravné práce na derivačním slovníku a na slovníku synonym a antonym.

4 Závěr

Projekt M-CAST je plánován jako dvouletý, vlastní práce začaly na jaře 2005. Kromě zdokonalení /doplnění zdrojů pro jednotlivé jazyky byly zahájeny práce na vytvoření ontologie na bázi MDT, podle plánu by měl být na počátku roku 2006 k dispozici funkční prototyp systému. Projekt bude vrcholit koncem roku 2006.

Tento příspěvek vznikl díky podpoře projektu eContent EDC no. 22249 M-CAST. Autoři dále děkují pracovníkům firmy TiP, a zejména Piotru Fuglewiczovi, za pomoc při vývoji české lingvistické podpory pro systém M-CAST.

Reference

1. Amaral, C.; Figueira, H.; Mendes, A.; Mendes, P.; Pinto, C. A Workbench for Developing Natural Language Processing Tools. URL: <http://www.priberam.pt/docs/WorkbenchNLP.pdf>
2. Amaral, C.; Laurent, D.; Martins, A.; Mendes, A.; Pinto, C. Design and Implementation of a Semantic Search Engine for Portuguese. URL: <http://www.priberam.pt/docs/LREC2004.pdf>

Annotation:

Project M-CAST: Development of Multilingual Content Aggregation System

The goal of M-CAST project is to develop a multilingual system for integration and searching in large collections of texts (and multimedia). The structure of the proposed system, the work done on its implementation and especially on its Czech language module is described.