SYSTÉM JAZYKOVÉHO MODELOVÁNÍ Z WEBU – ARCHITEKTURA A MODUL DEKAPITALIZACE

Jan Vavruška[1]

1 ÚVOD

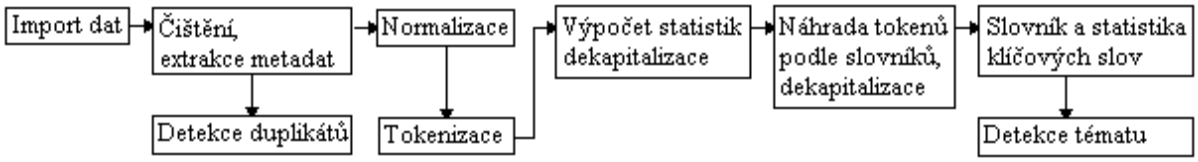
Systémy automatického rozpoznávání mluvené řeči jsou založeny převážně na statistickém přístupu, tj. pravděpodobnostním modelování řečníka (akustický model) a jazyka (jazykový model). Jazykové modely se snaží ocenit pravděpodobnosti po sobě jdoucích N-tic slov. Pro jejich sestavení je třeba zpracování rozsáhlých textů, vztahujících se k dané tématické oblasti, ze které chceme mluvenou řeč rozpoznávat. Významný zdroj takových textů dnes představuje internet. Proto byl týmem pracovníků Katedry kybernetiky vytvořen systém jazykového modelování z webu (dále jen JMZW), jehož cílem je automaticky (tj. s minimálními zásahy člověka) shromažďovat a zpracovávat textová data z českého internetu.

V první části abstraktu je popsána obecná architektura a běhové prostředí systému JMZW. Další část je pak věnována jednomu konkrétnímu modulu zpracování dat - dekapitalizaci.

2 ARCHITEKTURA A BĚHOVÉ PROSTŘEDÍ SYSTÉMU JMZW

Jádro celého systému tvoří MySQL popř. SQLite databáze zpracovávaných článků, nad kterou operují jednotlivé algoritmy. Jednomu záznamu (článku) odpovídá dále nedělitelný objekt databáze (Item) s množstvím atributů. Ke každému Itemu jsou ukládány mezivýsledky algoritmů. Ty mohou tvořit položky (Recordy) různých typů – databázové (TextRecord), data v souborovém systému (StoredRecord) a datové objekty prg. jazyka Python (PyRecord).

Základem pro tvorbu algoritmů se stala knihovna Voiar (Voice Archive), vyvinutá na Katedře kybernetiky. Ta plně využívá frameworku SQLAlchemy, který podporuje databázově nezávislé mapování záznamů a relací na datové objekty v Pythonu při současném zachování integrity dat (včetně dat v souborovém systému). Využitím této knihovny jsme schopni splnit základní požadavky kladené na implementaci systému, jako např.: možnost definovat databázově nezávislé algoritmy pro různé výpočetní operace, s podobnou strukturou a s možností jejich vzájemného volání či snadné konfigurace (*modularita*), *rozšiřitelnost* databázového schématu podle aktuálních požadavků, *škálovatelnost* jak v objemu zpracovávaných dat (od desítek po miliony článků) tak v možnosti paralelního běhu algoritmů (od jednoho procesoru po cluster více počítačů). Dále i nezávislost běhu systému na operačním systému a v různých módech činnosti (interaktivní, neinteraktivní, ladící, apod.).



Obr. 1: Architektura zpracování dat v systému JMZW

Na obr. 1 vidíme architekturu systému se základními moduly. Jejich význam je následující:

[1] Jan Vavruška, student navazujícího studijního programu Aplikované vědy a informatika, obor Kybernetika a řídící technika, specializace Umělá inteligence, e-mail: sandokan@students.zcu.cz

Import dat – periodický monitoring RSS kanálů vybraných zpravod. serverů, stažení článků
Čištění, extrakce metadat – extrakce čistého textu a metadat z HTML (nadpis, autor, datum,..)
Normalizace – převod neortografických symbolů (číslice, zkratky) do plných slovních tvarů
Tokenizace – oddělení „diktovacích jednotek" (slov, čísel, interp.,..) tak jak jsou diktovány
Statistiky dekapitalizace – výpočet parametrů pro vyhodnocení výrazu (1) v násled. kapitole
Náhrada podle slovníků, dekapital. – nahrazení výrazů jinými podle nahrazovacích slovníků
Slovník, klíčová slova – slovník (rejstřík) celé databáze, přiřazení klíčových slov článkům
Detekce tématu – přiřazení článku k určitému tématu pro účel tématického jazyk. modelování
Detekce duplikátů – pravidelná detekce duplicit pro články stažené za určité časové období

3 DEKAPITALIZACE

Dekapitalizace je zahrnuta v modulu nahrazování podle slovníků. Jejím cílem je nahradit slova s velkým písmenem, která se vyskytují na začátku věty či odstavce a nejsou vlastními jmény, za slova s písmenem malým.

Do seznamu kandidátů k dekapitalizaci se nejprve uloží všechna slova na začátku vět či odstavců a s velkým písmenem. Z něj jsou odebráni ti, kteří splňují následující podmínky:

- Nevyhovuje výrazu:

$$\frac{C'(w)}{C(w)} \geq p, \tag{1}$$

  kde $C(w)$ je celkový počet výskytů slova w v databázi, $C'(w)$ je celkový počet slov w v databázi, která jsou kandidátem na dekapitalizaci a p je zvolený práh.

- Dekapitalizovaný kandidát se ani v jednom případě nevyskytuje ve velkém slovníku českého jazyka (3 mil. běžných českých slov) a současně v celé databázi JMZW a současně se nejedná o slovo s příklonkou „-li", např. „bude-li".
- Kandidát se vyskytuje jako „správný tvar slova" ve speciálním nahrazovacím slovníku.

4 ZÁVĚR

Budoucí práce spočívá zejména v modelových experimentech (s automatickým titulkováním vybraných pořadů ČT) pro vyhodnocení přínosu systému pro jazykové modelování. V neposlední řadě je ve vývoji zpracování titulků audio/video záznamů z rozhlasu a televize.

Poděkování:
Práce byla podpořena grantem Západočeské Univerzity, projekt č. SGS-2010-054 a Ministerstva průmyslu a obchodu, projekt č. MPO FR-TI1/486.

LITERATURA

Švec, J., 2010. *Knihovna Voiar (Voice Archive)*. Katedra kybernetiky, Západočeská univerzita v Plzni.

Švec, J., Skorkovská, L., Vavruška, J., Ircing, P., Lehečka, J., Pražák, A., Kanis, J., Hoidekr, J., Pressl, D., Stanislav, P., Soutner, D., 2010. *Výzkumná zpráva projektu Jazykové modelování z webu*, Výzkumná zpráva interního grantu Západočeské univerzity v Plzni č. SGS-2010-054

Švec, J., Hoidekr, J., Soutner, D., Vavruška, J., 2011. Web Text Data Mining for Building Lagre Scale Language Modelling Corpus, *TSD 2011,* Pilsen.