DETEKCE DUPLICITNÍCH ČLÁNKŮ PRO SYSTÉM JAZYKOVÉHO MODELOVÁNÍ Z WEBU

Jan Vavruška[1]

1 ÚVOD

Jednou z výzkumných oblastí katedry kybernetiky (dále jen KKY) na FAV ZČU jsou systémy automatického rozpoznávání mluvené řeči (dále jen ASR). Mezi tyto patří i diktovací systém „*Mega Word*“, jehož princip je založený na statistickém přístupu k rozpoznávání řeči. Jeho základními stavebními kameny tedy jsou pravděpodobnostní modelování řečníka (akustický model) a jazyka (jazykový model).

Jazykový model se snaží ocenit pravděpodobnosti posloupností slov ve větě a pro jeho sestavení je potřeba zpracování rozsáhlých textů z dané tématické oblasti. Proto byl na KKY založen projekt na vytvoření systému pro jazykové modelování z webu (dále jen JMZW), jehož cílem bude automatická aktualizace a adaptace slovníků systémů ASR prostřednictvím článků z nejrůznějších tématických oblastí, pravidelně stahovaných z internetu.

Mnoho článků publikovaných na internetu jsou částečnými a nebo úplnými duplikáty článků jiných. Tento fakt je významný jak z hlediska úspory dat v databázi systému JMZW, ale i např. proto, že vícekrát publikovaný článek se nejspíše týká nějakého důležitého tématu. Proto je součástí systému i modul pro detekci duplicitních a téměř duplicitních článků.

2 DETEKCE DUPLICITNÍCH ČLÁNKŮ

Detekce duplicitních článků je založena na metodě tzv. *šindelování*. Pro náš případ ji lze definovat tak, že článek je reprezentován souborem tri-gramových šindelů, tedy všech možných trojic po sobě jdoucích slov z daného článku. Při porovnávání dvou dokumentů se z jejich souborů šindelů počítá vzájemná podobnost. Bylo vyvinuto množství nejrůznějších metrik podobnosti, v našem případě byla použita jednoduchá symetrická metrika podobnosti, definovaná následovně:

Mějme dva články A, B a jejich odpovídající soubory šindelů $S(A)$, $S(B)$, potom podobnost těchto článků (*resemblance*)je definována jako

$$res(A,B) = \frac{S(A) \cap S(B)}{S(A) \cup S(B)}. \qquad (1)$$

Hodnota je z intervalu <0,1>, přičemž 0 znamená odlišné články a 1 články zcela shodné.

Je třeba tedy zvolit minimální práh podobnosti dvou dokumentů, které budou označeny za duplikáty. Jako optimální hodnotu prahu jsem poněkud heuristicky zvolil $p = 0{,}45$. Při tom jsem uvažoval délky dokumentů v poměru k jejich shodným částem (úvahy naznačuje tab. 1, pro zjednodušení převedeno na jednotky).

[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

délka(A)	*délka(B)*	*A∩B*	*A∪B*	*res(A,B)*
0.3	1	0.3	1	0.3
0.45	**1**	**0.45**	**1**	**0.45**
0.5	**1**	**0.5**	**1**	**0.5**
1	1	0.5	1.5	0.33
1	1	0.6	1.4	0.43
1	**1.25**	**0.7**	**1.55**	**0.45**
1	**1**	**0.65**	**1.35**	**0.48**

Tab. 1: Podobnost článků v závislosti na jejich délce a shodných částech.

Metrika je poměrně závislá na změnu délky dokumentů - s jejich rostoucí délkou v poměru se shodnou částí její hodnota klesá. Pro náš účel to však není újmou, protože pokud se články délkou a tedy i obsahem informací významně liší, bylo by větší chybou je označit za duplikáty než je ponechat jako jedinečné.

4 IMPLEMENTACE

Systém JMZW včetně modulu detekce duplikátů je implementován v jazyce *Python* ve spolupráci s databází *MySQL*. K tomu byla využita univerzální databázová knihovna *Voiar* (zkr. *Voice Archiv*, autorem je Ing. J. Švec z KKY), používající framework *SQLAlchemy*.

Duplikáty jsou v databázi implementovány seznamem. Protože metrika je symetrická, platí tranzitivita mezi nalezenými duplikáty. Seznam je uspořádán sestupně podle délky dokumentu neboť jistě je menším zlem pokud je kratší článek duplikátem delšího (i přesto, že je třeba starší), než naopak. Zvláště pak, pokud je kratší článek úplnou podmnožinou delšího.

3 ZÁVĚR

Oprávněnost detekce duplicit v systému JMZW ukázaly už první pokusy s malou referenční databází s celkem 2888 články (zdroj České Noviny a Anopress), z nichž plných 1399 bylo duplikováno (převážně zcela identickými články).

Výhodou metody *šindelování* je malá citlivost na změnu pořadí, výmaz nebo přidání slova. Jeho největší nevýhodou jsou značné nároky na úložný prostor. Tento problém je vyřešen tím, že pracovní soubory šindelů nezůstávají v databázi natrvalo, ale pouze během vlastní detekce.

Kvůli reprezentaci duplikátů seznamem zatím není zcela vyřešen problém, kdy článek *A* je úplnou podmnožinou článku *B*. Za duplikát bude označen jen v případě, že jeho délka vzhledem k B je větší, než práh podobnosti (viz např. první tři řádky v tab. 1).

Budoucí práce spočívá zejména v „ostrých“ pokusech s hlavní databází systému JMZW (obsahující miliony článků), řešení zmíněného problému duplikátů - podmnožin a také způsob, jak bude s vlastními duplikáty naloženo.

Poděkování:

Práce byla podpořena grantem Západočeské univerzity, projekt č. SGS-2010-054 - "Inteligentní metody strojového vnímání a porozumění".

LITERATURA

Dufková, K., 2008. *Dynamická detekce plagiátů*. Diplom. práce MFF UK, Praha.

Hauzírek, M., 2007. *Možnosti automatické detekce plagiátů.* Diplom. práce FIS VŠE, Praha