# Disambiguace několika českých slov pomocí Brillova taggeru

*Přemysl Šrubař*

## Specifikace úlohy

Cílem je disambiguace (zjednoznačnění) několika vybraných českých slov pomocí *Brillova taggeru*. Jedná se o tato slova a jejich značky:

| | |
|---|---|
| se | k3, k7 |
| je | k5, k3 |
| vedoucí | k1gM, k1gF, k2gM, k2gF |
| Jana | k1gM, k1gF |

## Příprava dat

Data byla získána z korpusu *Desam*. Pro každé vybrané slovo (se, je, vedoucí, Jana) v kontextu pěti řetězců (pět před i po vybraném slově), pouze značky (bez lematu).

Pro použití v *Brillově taggeru* bylo třeba získaný korpus upravit:

- Doplnit konkrétní značku za prázdné značky. Například tečka nebo čárka za větou je v korpusu *Desam* označena pouze lomítkem bez značky. Takovéto řetězce byly označeny značkou *ke*.
- Odstranit odsazení a prázdné řádky.
- Odstranit ostré závorky, kterými jsou označena vybraná slova. Např: <*Jana/k1gMnSc2*>.
- Vynechat řádky, které obsahují značky jako <*x*>, <*z*> apod.
- Doplnit značku ke slovům, která nejsou označkována. To bývají převážně slovní spojení jako například *Tyranosaurus Rex/k1gMnSc1*. Těmto slovům byla dána značka *ke*.
- Výskyt více lomítek za sebou přepsat na znak |.

Pro zjednodušení byl navíc u značek ponechán pouze druh slova a rod (pro rozlišení slova *Jana*).

Všechny tyto úpravy obstarává skript desam2brill.prl

## Brillův tagger

Přeložený *tagger* (*win32*): Tagger Win32.zip.

Naučená data: SeJeJanaVedouci.zip.

Značkování pomocí Brillova *taggeru* probíhá zhruba takto:

- Všem slovům se přiřadí jejich nejčastější značka.
- Pomocí pravidel pro značkování neznámých slov se přiřadí značka i neznámým slovům
- Aplikují se kontextová pravidla – opravují se chyby.

Naučená data jsou v těchto souborech:

| *Soubor* | *Popis* | *Příklad* |
|---|---|---|
| LR | Lexical-rules - Pravidla pro značkování neznámých slov.<br>Naučený *tagger* obsahuje těchto pravidel 1207 (pro angličtinu zhruba 140). | **o char k5 5328.27817522851**<br>Pokud se ve slově vyskytuje písmenko *o,* označkuj slovo *k5* |
| | | **k1 a fchar k1gF 1257.97859349202**<br>Změň značku *k1* na *k1gF,* pokud se ve slově vyskytuje *a*. |
| | | **ní hassuf 2 k1gN 987.397887274155**<br>Má-li slovo suffix *ní*, označkuj slovo jako *k1gN*. |
| | | **k5 v fgoodright k1gF 31.3566037735849**<br>Změň značku *k5* na *k1gF,* pokud se napravo vykytuje slovo *v*. |
| CLR | Contextual-rules – Kontextová pravidla.<br>Naučený *tagger* obsahuje těchto pravidel 448 (pro angličtinu zhruba 250). | **k2gF k2gI NEXT1OR2TAG k1gI**<br>Změň *k2gF* na *k2gI*, pokud má jedno ze dvou následujících slov *k1gI*. |
| | | **k2gI k2gN NEXTTAG k1gN**<br>Změň *k2gI* na *k2gN*, pokud má následující slovo značku *k1gN*. |
| | | **k3gI k3gM PREV1OR2OR3TAG k1gM**<br>Změň *k3gI* na *k3gM*, pokud má jedno ze tří předchozích slov *k1gI*. |
| | | **k3 k7 PREV1OR2WD se**<br>Změň *k3* na *k7,* pokud je jedno ze dvou předchozích slov *se*. |
| BGL | Viz dále. | |
| TL | Viz dále. | |

Ve všech souborech je použité kódování češtiny *Win CP 1250*.
Ve vstupním textu musí být všechny řetězce odděleny mezerou, včetně interpunkce. Například:

*Po zápase řekl: "Cítím se dobře."*

musí být:

*Po zápase řekl : " Cítím se dobře . "*

Také by měla být každá věta na samostatném řádku, ale není to nutné.
Učení *taggeru* trvalo přibližně 66 hodin. Z toho 56 hodin trvalo učení modulu pro značkování neznámých slov, které je napsáno v *perlu*. Učení kontextových pravidel je napsáno v jazyku *C*.

Naučený *tagger* si můžete vyzkoušet přes webové rozhraní.

# Proces učení

Soubory nutné k učení (pod *Win32*): Learn_Win32.zip.

Většina utilit, které je třeba k učení, jsou napsána v jazyku *Perl.* Je také nutno použít některé základní programy *Unixu* : *cat, dos2unix, (g)awk, ls, mv, perl, rm, sort, tee, (unix2dos).*

I některé utility napsané v jazyku *C* vyvolávají systémové příkazy. Dokonce i *tagger.exe* vyvolává programy *cat* a *tee,* nastěstí ne při samotném značkování (bez použití volitelných přepínačů).

Pro učení pod *Win32* je proto vhodné nainstalovat *Cygwin* - některé programy *unixu* portované pro *Win32*. Pozor na program *sort*, který je součástí novějších *windows,* ale není kompatibilní s unixovským *sort* (přepínače *-rn)*. Bývá většinou v cestě.

Pokud nemáte nebo nechce instalovat *Cygwin,* stačí rozbalit archív CygwinCropped.zip do stejného adresáře jako archív *Learn_Win32.zip* a přidat soubor *RegCygwin.reg* do registrů (obsahuje pouze tři záznamy o mapování adresářů */usr/bin*, */usr/lib* a */*, které se namapují na aktuální adresář).

Učení *Brillova taggeru* se skládá ze dvou hlavních částí: hledání pravidel pro značkování

neznámých slov a hledání pravidel podle kontextu. K zahájení učení (všech částí) je možné použít skript

Learn.bat *corpus1 [[[[corpus2] corpus3 ] .. ] corpus9]*

Kde *corpus1 .. corpus9* jsou korpusové soubory ve formátu *desam*. Je-li zadáno více souboru, jsou nejdříve spojeny dohromady.

Dále podrobněji:

## *Učení modulu pro značkování neznámých slov:*

Před zahájením samotného učení je třeba nejdříve udělat několik úprav, případně vytvořit některé soubory:

| ***Soubor*** | ***Skript*** | ***Popis*** | ***Příklad*** |
|---|---|---|---|
| Tagged | CrtTagged.bat | Převede korpus (v souboru *input*) z formátu *desam* na formát požadovaný *taggerem*. Upraví i konce řádků z /CR/LF na /LF. | |
| Tagged1<br>Tagged2 | Split.bat | Náhodně rozdělí soubor *tagged* na *tagged1* a *tagged2* (po řádcích). | |
| Utagged<br>Utagged1<br>Utagged2 | CrtUntagged.bat | Odstraní všechny značky v souborech *tagged*, *tagged1* a *tagged2*. Výsledek je v souborech s prefixem *U*. | |
| BWL | CrtBwl.bat | Ze souboru *Utagged* vytvoří seznam všech slov a setřídí je podle četnosti. | se<br>,<br>.<br>v<br>na |
| SWL | CrtSwl.bat | Ze souboru *Tagged1* (první polovina korpusu) vytvoří seznam, obsahující slovo a jeho nejčastější značku (a počet, pouze pro informaci). | , kE 18430<br>se k3 16862<br>. kE 15686<br>je k5 10256 |
| BGL | CrtBGL.bat | Ze souboru *Utagged* vytvoří seznam slovních dvojic. Tento soubor je výstupní, používá se při samotném značkování. | výfukovými plyny<br>své podpisy<br>zeslabení negativního<br>stoprocentní jistotou |

Všechny tyto skrpity lze dohromady vyvolat pomocí learnUnknownInit.bat.

Samotné učení se spustí souborem learnUnknown.bat. Vznikne tak soubor *LR*.

## *Učení kontextových pravidel*

Učení kontextových pravidel musí být spuštěno až po učení modulu pro neznámé slova, protože využívá některé soubory vzniklé při učení tohoto modulu.

| *Soubor* | *Skript* | *Popis* | *Příklad* |
|---|---|---|---|
| TL | CrtTL.bat | Ze vstupního korpusu *tagged1* (první polovina) vytvoří seznam, obsahující slovo a seznam všech jeho možných značek. První uvedená značka je ta nejčastější, pořadí ostatních není určené. | Jana k1gM k1gF<br>Hradci k1gI<br>podobným k2gI k2gN<br>antika k1gF |
| FL | CrtFl.bat | Stejně jako *TL,* ale pro kompletní korpus *tagged*. Tento soubor je výstupní, používá se při samotném značkování. | |
| DTC | CrtDtc.bat | Aplikuje *tagger,* na soubor *Utagged2*. Vznikne tak korpus pro ověřování při učení (testovací množina). | |

Všechny tyto skrpity lze dohromady vyvolat pomocí learnContextInit.bat.

Samotné učení se spustí souborem learnContext.bat. Vznikne tak soubor *CLR*.

Všechny (čtyři) výstupní soubory, které jsou třeba ke značkování, je možné zkopírovat do zvoleného adresáře pomocí copyOutput.bat. Dočasné soubory vzniklé při učení maže skript clearTmp.bat.

# Kompilace modulů v *C* pod *Win32*

K překladu bylo použito GNU *Dev-C++* od *Bloodshed software*, knihovna *Libgw32c* (kvůli funkci *getopt*). *Dev-c++* nepodporuje více projektů v jednom řešení (*solutions*), proto je pro každý projekt vytvořen samostatný adresář. Pro použití *taggeru* s českými značkami a pro úspěšné přeložení bylo třeba udělat následující změny v původním kódu:

### Start-state-tagger

- Ošetřen nesprávný počet parametrů (*null-exception*).
- Změněn algoritmus na počáteční značkování neznámých slov. Původně byla všechna slova začínající malým písmenem označena tagem *NN* a začínající velkým *NNP*. Nyní se všechna slova nejdříve označí tagem *k1*. (Obdobně byl upraven i skript *unknown-lexical-learn.prl*}

### Final-state-tagger

- Ošetřen nesprávný počet parametrů (*null-exception*).

### Tagger

- Přidán hlavičkový soubor *getopt.h* a knihovny *libgw32c.a*, *libole32*.a, *libuuid.a*. Kvůli funkci *getopt*, která není součástí *Dec-C++*. Tyto knihovny jsou distribuovány pod GNU licencí.
- Přidána přípona *.exe* k názvům modulů *start-state-tagger* a *final-state-tagger*.

### Contextual-rule-learn

- Ošetřen nesprávný počet parametrů (*null-exception*).
- Funkce *tmpnam* nahrazena, protože vytvářela názvy začínající lomítkem a programy z *cygwin*u hledají takovéto soubory v aktuálním adresáři. Pomocné soubory se jmenují napevno *tmp1*, *tmp2*.
- Upravena cesta k programu *rm* z absolutní (*/bin/rm*) na relativní (v aktuálním adresáři).

# Dosažená přesnost

| *Slovo* | *Přesnost (%), získaná přiřazením nejčastější značky* | *Přesnost taggeru* |
|---|---|---|
| Libovolné | | 95.06 |
| Jana, vedoucí, je, se | 90.097 | 95.6 |
| Jana | 74.5 | 89.775 |
| vedoucí | 39.131 | 60.624 |
| je | 89.851 | 95.843 |
| se | 90.541 | 95.595 |

Slovo *vedoucí* je v korpusu nejméně zastoupené (cca 50 vět) a navíc se klasifikuje do čtyř skupin. Proto má nejmenší úspěšnost.

Úspěšnost *taggeru* pro libovolné slovo se zdá být větší, než pro některá zaměřená slova. To je nejspíše způsobeno tím, že se mezi libovolné slova započítávají i oddělovače slov (tečka, čárka), uvozovky a podobně, které mají vysoký výskyt (viz *BWL*) a 100% úspěšnost.

Tento dokument je ve formě html pdf.