Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Bakalářská práce

Odhad morfologických značek pro neznámá slova

Plzeň 2012 Milan Nikl

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 8. srpna 2012

Milan Nikl

Abstract

This thesis describes assigning morphological tags to unknown words based on relations between those words and words known (to the computer). It is focused on the largest differences between human and computer perception of the Czech and Slovak language in general as well. This thesis also explains what the biggest problems of guessing morphological tags are, how to eliminate them and at least diminish the influence of possible errors on the tag assigning process.

Obsah

1 Úvod

V dnešní době je interakce člověka s počítačem naprosto běžnou součástí lidského života. Stále více pak nabývá na významu strojové zpracování textu a řeči (většina současných systému zpracovává mluvený projev tak, že jej nejprve převede do psané podoby a pak analyzuje tímto způsobem vytvořený text). Při další práci s textem je většinou třeba rozdělit věty v daném sdělení na jednotlivá slova, která je dále nutné nějakým způsobem kategorizovat. Každému tvaru slova je třeba přidělit určitý symbol, který by jej jednoznačně charakterizoval. Takový symbol, který by přesně popisoval daný tvar slova (nejen) z hlediska jeho mluvnických kategorií, se nazývá morfologická značka. Struktura a použití těchto značek je popsána v kapitole 3 na straně 5.

Při zpracování libovolného textu počítač využívá určitou slovní zásobu, se kterou porovnává slova vyskytující se v analyzovaném textu. Pokud bychom chtěli pomocí morfologických značek označit libovolný text, je pravděpodobné, že dříve nebo později narazí počítač na slovo, které jeho slovní zásoba neobsahuje.

Právě na přidělování morfologických značek takovým, tj. předem neznámým, slovům se tato práce zaměřuje. Konkrétně se zabývá morfologickým značkováním slov českého a slovenského jazyka. Na následujících stranách je popsáno, jak probíhá odhad morfologických značek u neznámých slov v počítači, jak se tento proces liší od podobného rozhodování u člověka, a jaké problémy s sebou tato činnost přináší.

2 Morfologie

Na tomto místě bych rád podotkl, že tato práce je zaměřena především na odhad morfologických značek pro češtinu. Proto se použité příklady a postupy budou týkat především českého jazyka. Morfologické značkování se samozřejmě používá i u jiných jazyků, a jak je popsáno dále v textu, u mnohých jazyků je často jednodušší. Na tomto faktu se podílí zejména rozmanitost a bohatost jazyka českého, který obsahuje mnoho tvarů a výjimek, se kterými si počítač jednoduše nedokáže poradit. Zatímco pro člověka, který češtinu aktivně používá, nepředstavují nejrůznější nuance českého jazyka nic neobvyklého. Jak je v textu této práce popsáno na několika příkladech, strojové zpracování češtiny je velmi obtížné (i díky výše zmíněným faktům), což znamená, že účinnost samotného morfologického značkování nemůže být stoprocentní. Nicméně text práce rovněž obsahuje úvahy a postupy, kterými jsem se snažil pravděpodobnost úspěchu maximalizovat.

2.1 Morfologie

Tvarosloví (morfologie) je nauka o tvarech slov (jmen a sloves).

Takto zní oficiální definice dle [Šmilauer(1973)]. Ač je morfologie věda jistě velmi zajímavá a naučná, je také velice obsáhlá (zejména co se českého jazyka týče). Proto bych se rád zaměřil jen na několik faktů, které jsou podstatné pro tuto práci - tedy pro odhad morfologických značek u slov. Pro lepší pochopení problematiky uvádím v následujících odstavcích základní pojmy týkající se členění slov z hlediska tvarosloví.

2.2 Členění slov

Členíme-li postupně slovo (slovní tvar) na menší a menší bezprostřední složky, dospějeme k nejmenším útvarům, které mají ještě povahu znakovou, tj. jsou charakterizovány po stránce výrazové i významové. Takové nejmenší části mající svůj výraz i význam, vyčleněné na základě opakování (v různých kombinacích) v jiných slovních tvarech, nazýváme morfy. [Dokulil(1986)]

2.2.1 Typy morfů

Jak dále uvádí [Dokulil(1986)], z hlediska jejich funkce rozlišujeme v češtině dva typy morfů:

1. **Kořeny** jsou zřídkakdy samostatné, většinou však nesamostatné, morfy vyjadřující elementární (tj. nesložené) lexikální významy.

2. **Afixy** jsou vždy jen nesamostatné morfy vyjadřující elementární nebo složené „gramatické“, tj. zobecněné významy; jsou jednak slovotvorné, blíže určující (determinující) kořen, jednak tvaroslovné (tvarotvorné), sloužící ohýbání (flexi) slova, tj. skloňování a časování.

   Podle postavení v slovním tvaru vzhledem ke kořenu se rozlišují:

   (a) **prefixy** neboli předpony, mající místo před kořenem (a modifikující význam tohoto kořene, resp. celého slovního tvaru bez tohoto prefixu).

   (b) **sufixy** neboli přípony, mající místo za kořenem, a to bezprostředně nebo za jiným sufixem determinujícím kořen (a pozměňující jeho význam).
   Zatímco většina předpon patří mezi slovotvorné (tj. pomocí těchto předpon se vytvářejí nová slova), mezi příponami se nachází velké množství tvaroslovných afixů (tvaroslovné přípony - pádové, osobní a infinitní - stojící na konci slovního tvaru nazýváme koncovky). Což znamená, že se u slova mění přípona se změnou pádu či osoby, nikoli jeho samotný význam. Při odhadování morfologických značek využíváme právě toho, že slova s odpovídajícím sufixem mají většinou i stejnou značku.

   (c) **postfixy** jsou zvlášní, pouze slovotvorné afixy, připojované až za gramatický sufix, koncovku, tedy k úplnému slovnímu tvaru. (Tvoří tedy přímý protějšek prefixů, které se rovněž spojují s celým slovním tvarem.)

2.3 Morfologické alternace

V poslední částí této kapitoly věnované morfologii bych se rád zaměřil na jev zvaný *Morfologická alternace*. I když se jedná v podstatě o okrajovou tematiku, na účinnost odhadu morfologických značek má nezanedbatelný vliv. Pro objasnění uvádím nejprve několik definic:

Fonémy, jak uvádí [Dokulil(1986)] jsou jazykové jednotky, ze kterých se skládají znakové jednotky jazykového systému. Jsou to znakové jednotky tvořené podle ustálených forem (šablon).

Dále se zde uvádí, že základní znakové jednotky se skládají ze dvou složek. Jedna je nositelem významu (tzv. složka označující), druhá je složka významová (označovaná). V rovině jazykového systému se složka označující skládá z fonémů, teprve komunikačním aktem vzniká z abstraktního útvaru fonémového akustická konkretizace, bez níž by se komunikace nemohla uskutečnit.

Morfologickou alternací (neboli střídáním fonémů) se pak rozumí fonologicky nepodmíněná zákonitá záměna téhož morfému při tvoření slov a tvarů.

Přičemž výskyt toho či onoho alomorfu nemůže být objasněn fonologickými zákony současnho jazyka, nýbrž pouze historicky. Z hlediska současného jazyka je podmíněn pouze sousedními morfy v struktuře slovního tvaru.

Příkladem morfologických alternací jsou následující dvojice slov: *ruka - ručka, hoch - hošík, kniha - knižní* ale třeba i změny typu: *myslit - myšlení, den - dny, úžit - úžím.*

Mezi nejčastější alternace patří například dloužení u samohlásek či měkčení u souhlásek. Celkem [Dokulil(1986)] uvádí čtyři hlavní typy alternací s celkem sedmadvaceti podtypy. Nicméně pro odhad morfologických značek jsou podstatné pouze takové alternace, kde dochází například ke změně kořenu slova (se změnou pádu či čísla u podstatných jmen, se změnou osoby u sloves či při stupňování jmen přídavných). Nebo takové alterace, kde pro shodné lemma existuje více variant, jejichž psané podoby se velice výrazně liší. Najít společnou část u takových slov je často velice nesnadné, v některých případech dokonce nemožné, a výsledky takové snahy mohou být velice zavádějící. Bližší vysvětlení možných problémů, které mohou nastat, společně s několika příklady je uvedeno až v kapitole věnované samotnému odhadu morfologických značek.

3 Morfologické značky

3.1 Definice

Morfologické značky jsou součástí výsledku (výstupem) morfologické analýzy, která pracuje s izolovanými slovními tvary, tedy bez ohledu na jejich kontext. Druhou částí výsledku je tzv. lemma, které identifikuje příslušnou lexikální jednotku ve smyslu jednoznačné identifikace slovníkového hesla. Jak uvádí [Hajič(2010)]

3.2 Čeština - systém pozičních značek

Na tomto místě bych rád popsal celý systém morfologického značkování tak, jak je používán pro práci s Českým národním korpusem vytvořeným na Filozofické fakultě Univerzity Karlovy. Je nutno podotknout, že se rozhodně nejedná o jediný existující systém značkování. Každý jazyk používá značky podléhající jeho konkrétním potřebám s ohledem na tvorbu slov v daném jazyce a na určování mluvnických kategorií (a dalších aspektů) u jednotlivých slovních tvarů. Tím, že čeština patří mezi flektivní jazyky a jako taková je velmi rozmanitá, je dána i poněkud složitější struktura samotných značek. V následujících odstavcích bych rád popsal systém značkování alespoň na takové úrovni, aby mu porozuměl i člověk, který se s ním nikdy nesetkal. Detailní popis celého systému morfologického značkování je pak možno najít v [Hajič(2004)]

3.2.1 Struktura značky

Každá značka je řetězcem 16 znaků (Výjimkou jsou korpusy vytvořené před rokem 2000, kde značku tvoří pouze 15 pozic. Vývoj značení češtiny popisuje [Jelínek(2008)]). Značka je konstruována tak, aby každá pozice odpovídala jedné morfologické kategorii podle víceméně tradičního lingvistického pojetí. Každé hodnotě v dané kategorii odpovídá jeden znak, převážně písmeno velké abecedy (např. `P` pro plurál, neboli množné číslo), výjimečně i jiný znak (např. `f` pro infinitiv, nebo '`,`' pro podřadicí spojky). Hodnota, která nedává smysl

(např. pád u sloves), je reprezentována znakem '`-`' (pomlčka).

Následuje popis jednotlivých pozic značky tak, jak následují bezprostředně za sebou:

1. **Slovní druh** - toto označení odpovídá víceméně tomu, jak slovní druhy rozlišuje česká gramatika (jak jsou slovní druhy například vyučovány na školách v hodinách českého jazyka). Nicméně pro potřeby další analýzy jazyka může být značení upraveno. Zejména pak v těch případech, kde se od sebe liší určení slovního druhu u téhož slova v závislosti na zvolené gramatice či slovníku.
   Příklad: `A` - přídavné jméno (adjektivum), `V` - sloveso (verbum), `Z` - interpunkce

2. **Detailní určení slovního druhu** - slouží pro přesnější rozlišení jednotlivých slovních druhů. Od značky v této pozici se většinou odvíjejí i značky následující. Jelikož se jedná pouze o upřesnění, je také možno ze značky v této pozici odvodit symbol v pozici předcházející - samotný slovní druh.
   Příklad: `D` - ukazovací zájméno (ten, onen, ...), `r` - číslovky řadové, `f` - slovesný tvar - infinitiv

3. **Jmenný rod** - pokud se určuje, jinak `-`
   Příklad: `M` - mužský (maskulinum), `F` - ženský (femininum), `N` - střední (neutrum)

4. **Číslo** - určuje-li se
   Příklad: `S` - jednotné (singulár), `P` - množné (plurál)

5. **Pád** - určuje-li se
   Příklad: `1` - nominativ (1. pád), `2` - genitiv (2. pád), ...

6. **Přivlastňovací rod** - přičemž rody střední a mužský neživotný se nikdy nevyskytují samostatně. Rod mužský životný se může vyskytovat pouze u přivlastňovacích adjektiv (přídavných jmen).
   Příklad: `F` - rod ženský (feminunum), `M` - rod mužský životný (maskulinum animatum), `X` - libovolný rod

7. **Přivlastňovací číslo** - určuje-li se
   Příklad: `S` - jednotné (singulár), `P` - množné (plurál)

8. **Osoba** - tak, jak ji určujeme u sloves bez závislosti na čísle.
   Příklad: `1` - 1. osoba, `2` - 2. osoba, `3` - 3. osoba

9. **Čas** - jak jej určujeme u sloves
   Příklad: `F` - futurum (budoucí čas), `P` - présens (přítomný čas), `R` - minulý čas

10. **Stupeň** - určujeme u přídavných jmen
    Příklad: `1` - 1. stupeň, `2` - 2. stupeň, `3` - 3. stupeň

11. **Negace** - podle toho, zda je ve slově obsažena předpona „ne-“
    Příklad: `A` - afirmativ (pozitivní tvar), `N` - negativ (obsahuje předponu „ne-“)

12. **Aktivum / Pasivum** - určuje-li se
    Příklad: `A` - aktivum, `P` - pasivum

13. **Neurčená pozice** - obsahuje znak `-`

14. **Neurčená pozice** - obsahuje znak `-`

15. **Varianta, stylový příznak, ...** - například podle toho, jak je dané slovo používané
    Příklad: `4` - velmi archaický nebo knižní tvar, `6` a `7` - hovorový tvar, `8` - zkratka

16. **Slovesný vid** - tato pozice byla doplněna oproti starším verzním slovníků na základě slovníku morfologické analýzy. Proto není v některých korpusech k dispozici.
    Příklad: `P` - perfektum (dokonavé sloveso), `I` - imperfektum (nedokonavé sloveso), `B` - obouvidné sloveso

3.2.2 Příklady značek

Pro lepší představu uvádím několik slov, která se mohou vyskytovat v textu, a podobu k nim náležících značek:

Oldřich	`NNMS1-----A----`	**a**	`J^------------`
zdravotní	`AAIS4----1A----`	**zatím**	`Db------------`
452	`C=------------`	**OECD**	`NNFXX-----A---8`
obdrží	`VB-S---3P-AA---`	**%**	`Z:------------`

Samozřejmě jednomu tvaru slova je možno přiřadit více značek, přičemž ta správná se odvíjí od kontextu dané věty. Nicméně tato konkrétní slova mají sloužit pouze jako ukázka morfologického značkování, takže si jejich

vytržení z kontextu můžeme dovolit. Další informace k pozičnímu systému morfologického značení lze najít na [Hajič(2010)].

3.3 Slovenština

Zatímco při značkování češtiny je použit tzv. poziční systém, ve kterém mají všechny značky shodnou délku (15 pozic), Slovenský národní korpus používá systém značek proměnlivé délky. Ač je počet pozic ve značce variabilní, jejich pořadí je pevně dané. A stejně jako u značení češtiny je na prvním místě informace o slovním druhu daného tvaru (respektive ke slovní třídě, nebot' je rozlišována i interpunkce a další znaky).

V následujících odstavcích je popsána struktura značek tak, jak je používá *Jazykovedný ústav L'udovíta Štúra Slovenskej akadémie vied v Bratislave.* (viz. [SAV(2004)]) Přičemž nejprve jsou popsány značky nejjednodušší, následně značky pro jednotlivé slovní druhy.

3.3.1 Značky o jedné pozici

Uvádím výčet základních typů. Značku o jedné pozici mají dále např. neurčitelné slovní druhy (`Q`) či neslovní elementy (`#`).

Znak	**Hodnota** (Slovní druh)	**Příklad**
`J`	Citoslovce (interjekcia)	*bác, plesk, ahoj*
`R`	Zvratné zájmeno (reflexívum)	*sa, si*
`Y`	Kondicionálový morfém	*by*
`W`	Značky a zkratky (abreviácie)	*km, kg, SND, NaOH*
`Z`	Interpunkce	*., !,), +*
`O`	Číslice	*158, 230, 3 (razy)*

3.3.2 Značky používající až dvě pozice

U spojek a částic se používá druhá pozice, pakliže mají podmiňovací tvar či význam. Tato vlastnost (tzv. kondicionálnost) se vyjadřuje znakem `Y`.

Znak	**Hodnota** (Slovní druh)	**Příklad**
`O`	Spojka	*ale, alebo, pretože*
`OY`	Spojka s podmiňovací podobou	*aby, keby, žeby*

Znak	**Hodnota** (Slovní druh)	**Příklad**
`T`	Částice	*nuž, bodaj, azda*
`TY`	Částice s podmiňovací podobou	*kiežby, žeby*

Dále se dvě pozice používají u příslovcí, zde pak druhá značka určuje stupeň daného příslovce (ne všechna příslovce jdou samozřejmě stupňovat, taková pak mají opět značku jen o jedné pozici).

Znak	**Hodnota** (Slovní druh)	**Příklad**
`D`	Příslovce	*milo, prázdno, pravidelne*
`Dx`	1. stupeň (pozitív)	*draho, vzácne*
`Dy`	2. stupeň (komparatív)	*drahšie, vzácnejšie*
`Dz`	3. stupeň (superlatív)	*najdrahšie, najvzácnejšie*

3.3.3 Značky používající více pozic

Podstatná jména

Značky pro podstatná jména jsou vždy pětimístné. První pozice logicky obsahuje znak pro substantivum - `S`. Druhá pozice upřesňuje druh podstatného jména (tzv. paradigma), třetí pozice značí rod podstatného jména (ženský, střední, mužský životný a mužský neživotný). Čtvrtá pozice značí číslo (jednotné, množné), na páté pozici je pak číslo pádu.

Přídavná jména

Značky pro přídavná jména mají šest pozic, přičemž první pozice obsahuje vždy `A` (adjektivum), druhá opět upřesňuje druh přídavného jména. Třetí pozice značí shodu (kongruenci) s podstatným jménem v rodě, čtvrtá v čísle, pátá v pádu. Konečně šestá pozice pak řeší stupeň přídavného jména. Kupodivu nikoli odpovídajícím číslem, ale znaky `x` pro pozitiv, `y` pro komparativ a `z` pro superlativ.

Zájmena

Délka značky pro zájmena je opět šest pozic, přičemž složení značky odpovídá tomu pro značku přídavného jména. Na první pozici `P` - pronominum, druhá upřesňuje druh zájmena, třetí rod (resp. shodu v rodě), čtvrtá číslo, pátá pád. Šestá pozice je pak vyhrazena pro příznak tzv. aglutinovanosti, u zájmen, která do takové skupiny patří, je zde pak znak `g`.

Číslovky

Pro číslovky je používána pětimístná značka. Na první pozici `N` - numeralia, druhá pozice pro upřesnění druhu číslovky, třetí pro rod (shodu v rodě), čtvrtá pro číslo, pátá pro pád.

Slovesa

Nejdelší značka je použita pro slovesa. Na první pozici znak `V` - verbum. Druhá pozice je věnována tzv. slovesné formě, rozlišuje takové kategorie jako *infinitiv, prezent, imperativ, přechodník, futurum, příčestí minulé*. Třetí znak určuje slovesný vid, čtvrtý číslo. Pátá pozice značí osobu, ale opět poněkud netradičně značenou (`a` pro první, `b` pro druhou, `c` pro třetí). Na šesté pozici se nachází znak pro shodu v rodě, přičemž jsou rozlišeny i rod všeobecný (znak `h`) a případný neurčitelný nebo neurčený rod (zejména u vícenásobného podmětu, znak `o`). Poslední pozice rozlišuje kladný (afirmace, znak `+`) a záporný (negace, znak `-`) tvar slovesa.

3.3.4 Speciální příznaky

U jakéhokoli slova může být přidán ke značce ještě speciální příznak, přičemž tento příznak je vždy složený ze dvou znaků a je připojen až za značku samotnou. Obecně jsou rozlišovány dva druhy příznaků:

Znak	**Hodnota** (Slovní druh)	**Příklad**
`:r`	Vlastní jméno	*Marián, Vašáryová, Tatry*
`:q`	Chybný zápis	*čučoreidka, pridátete*

3.3.5 Příklady značek

Jak je zřejmé, je struktura slovenských značek také systematická, ale poněkud jiným způsobem, pokud provedeme porovnání s českým značením. U českého způsobu značení je pozice vždy využita, pokud jde u daného tvaru slova určit, nebo obsahuje znak '-'. Dalo by se tedy říci, že stavba morfologické značky je mnohem pevnější.

Peter	`SSms1:r`	**a**	`O`
nad	`Eu7`	**najkratšie**	`Dz`
452	`0`	**OECD**	`W`
spievajúceho	`Gkns2x`	%	`Z`

3.4 Lemmatizace

Jak je zmíněno v popisu morfologického značkování, jakýmsi protikladem značky je tzv. *lemma*. Lemma je základní tvar slova, používaný např. ve slovnících. Jelikož je český jazyk velmi rozmanitý, zejména co se týká ohebných slovních druhů, pro jedno slovo existuje většinou několik tvarů (například v závislosti na pádu a čísle u podstatných jmen, nebo na osobě a času u sloves). Výhoda lemmatu je pak právě v tom, že všechny tvary téhož slova mají společné lemma, čehož se dá patřičně využít při odhadu morfologických značek (samotný proces je popsán v kapitole 5 na straně 15). Lemma tak slouží jako unifikované spojení daného tvaru slova s jeho významem. U některých slov jsou pak v lemmatu uvedeny ještě speciální značky nebo vysvětlivky pro lepší rozlišení. Pro lepší představu uvedu opět několik příkladů, jak může lemma vypadat:

Lemma	**Možné varianty slova**
`který`	která, kterého, kterém...
`uvést`	uvedl, uvede, uvedena...
`los-1_^(zvíře)`	losem, losa, losy...
`mít`	měla, má, mají...
`květen`	května, květnu, květnem...
`hezký`	hezkou, hezčím, nejhezčími...

4 Skryté Markovovy modely

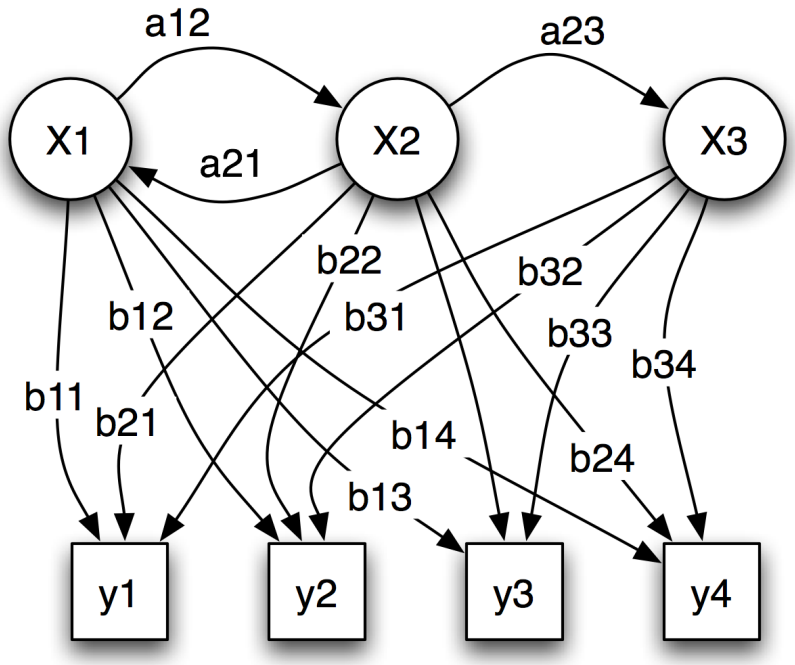
Zatímco v předchozí kapitole byly popsány především jazykové zákonitosti češtiny a vlastnosti systému morfologického značkování, tato kapitola má za úkol popsat matematický postup, který je přímo využíván (nejen) při odhadu morfologických značek. Skrytý Markovův model (zkr. HMM[1], autorem A. A. Markov, ruský matematik) je statistický aparát určený k popisu chování vybraného procesu či celého systému. Markovovy modely byly představeny v r. 1912, upraveny a poprvé prakticky využity byly v 60. letech, ale významněji používány jsou až od konce 80. let. Uplatnění nalézají v nejrůznějších rozpoznávacích, predikčních a dalších podobných systémech. Markovovy modely použité při rozpoznávání řeči jsou podrobně popsány v [Rabiner(1989)].

4.1 Funkce HMM

Princip činnosti HMM spočívá v tom, že se snaží nějaký problém převést do formy konečného automatu. Přičemž pro některé problémy lze použít několik automatových reprezentací a i výběr správného automatu (modelu) je klíčový pro správné řešení úlohy.

Automat vytvořený pro popis dané úlohy obsahuje N stavů $(S_1, S_2, \cdots S_N)$, přičemž v určitých okamžicích (v každém diskrétním čase) $t = 1, 2, \cdots T$ dochází k přechodu z aktuálního stavu do stavu nového či setrvání v současném stavu. Stav, ve kterém se automat nachází v daném čase t, se značí q_t. V každém stavu je pak generováno určité pozorování (observation) O (ve stavu q_t je to pak pozorování O_t). Výsledkem takového pozorování je pak jeden z výstupních symbolů $v_1, v_2, \cdots v_M$. Jak může skrytý Markovův model vypadat, znázorňuje následující obrázek (zde použito alternativní značení X_1, X_2, X_3 jsou stavy modelu, $y_1, \cdots y_4$ pak výstupní symboly HMM):

[1] z angl. Hidden Markov Model



Obrázek 4.1: Příklad skrytého Markovova modelu.

Celý model je pak možno (společně s počtem stavů N a počtem výstupních symbolů M) jednoznačně popsat třemi parametry:

1. Maticí přechodových pravděpodobností $A = \{a_{ij}\}$, kde:

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), \qquad \sum_{k=1}^{N} a_{ij} = 1, \qquad 1 \leq i, j \leq N.$$

   Přičemž platí, že pokud přechod ze stavu S_i do stavu S_j není možný, je $a_{ij} = 0$.

2. Posloupností pravděpodobnostních rozdělení $B = \{b_j(v_k)\}$, popisující pravděpodobnost pozorování výstupního symbolu v_k ve stavu S_j:

$$b_j(v_k) = P(O_t = v_k | q_t = S_j), \qquad 1 \leq i \leq N, \qquad 1 \leq k \leq M.$$

3. Množinou $\pi = \{\pi_i\}$, kde π_i je pravděpodobnost, že S_i bude počáteční stav systému:
$$\pi_i = P(q_1 = S_i) \qquad 1 \leq i \leq N.$$

Pokud známe matice pravděpodobností A, B a vektor pravděpodobností π (a parametry N a M), známe i kompletní specifikaci daného HMM. Výsledná model můžeme zapsat vztahem:

$$\lambda = (A, B, \pi)$$

4.2 Použití při odhadu značek

Při určování morfologických značek jsou skryté Markovovy modely použity na výběr vhodné značky pro každé slovo na základě kontextu (tzv. morfologické zjednoznačnění). Přidělované značky jsou jednotlivými stavy HMM. Slova, která chceme označkovat, jsou pozorováními (observation) takového modelu. Cílem programu je k posloupnosti slov přiřadit nejpravděpodobnější posloupnost stavů (značek). jsou Podrobnějším popisem se pak zabývá [Brychcín(2010)]

5 Odhad morfologických značek

Tato kapitola je věnována samotným metodám odhadu morfologických značek (dále *značka, tag*), myšlenkovým postupům, které za těmito metodami stojí, a případným problémům, které se mohou během procesu vyskytnout.

5.1 Základní úkol programu

Jak je patrné z předchozích kapitol této práce, základní snahou vytvořeného programu je přiřadit odpovídající značku každému slovu z určitého textu. Samotné přiřazení značky se pak skládá ze dvou dílčích problémů:

1. Morfologická analýza - přiřazení všech morfologických kombinací danému slovu.
2. Zjednoznačnění - výběr vhodné značky na základě kontextu.

Mnou vytvořený program se zaměřuje pouze na první část odhadu. Program tedy přiřadí slovu všechny možné morfologické značky, které by k němu mohly patřit. Část pro výběr odpovídající značky na základě kontextu je již hotová v laboratoři LIKS[1].

Fakt, že jsou vraceny všechny možné značky ke zkoumanému slovu bez ohledu na kontext pak znamená, že vracíme pouze ty značky, které jsou k dispozici u daného tvaru slova. Takže např. u slova *kalendář* je výsledkem značka jak pro 1., tak pro 4. pád podstatného jména, stejně tak u tvaru *židli* může být platná značka pro 3. i 6. pád podstatného jména. Výběr té vhodné značky je pak uskutečněn pomocí HMM.

5.2 Přidělované značky

Je nutné poznamenat, že značky přidělované slovům musí odpovídat značkám doopravdy existujícím, již známým. Testovaným slovům tedy můžeme

[1]Laboratoř Inteligentních Komunikačních Systémů

přidělit značkou pouze na základě podobnosti se slovy, která už známe - se slovy obsaženými ve vstupním korpusu (ať už PDT, SNK, či jiném).

Při samotném odhadu značek narážíme na problém chápání takové podobnosti mezi slovy z hlediska počítače. Při strojovém zpracování je totiž třeba vycházet z dat exaktně určených a podložených matematickými výpočty. Zatímco člověk se základní znalostí pravidel pro vytváření a odvozování slov v češtině by mohl uvažovat následujícím způsobem:

Příklad: slovo *nejkrásnější*

- určitě jde o přídavné jméno
- přípona *-ější* značí, že jde o 2. nebo 3. stupeň
- předpona *nej-* jednoznačně určuje 3. stupeň

⋮

Příklad: slovo *neznámý*

- určitě jde o přídavné jméno
- koncovka *-ý* značí, že se jedná o rod mužský
- předpona *ne-* znamená, že jde o záporný tvar jména

⋮

I když by se mohlo zdát, že u některých slov by se značka dala určit pomocí série několika jednoduchých pravidel, jako je tomu na výše uvedených příkladech, není to bohužel možné. Samotné převedení takových pravidel do podoby zdrojového kódu sice lze uskutečnit, nicméně množství pravidel, která by bylo nutné vytvořit, je obrovské. Stejně tak by bylo potřeba mít dokonalé lingvistické znalosti o daném jazyce (např. češtině).

V okamžiku, kdy značku přidělíme, musí být známá pravděpodobnost, se kterou se může přidělená značka vyskytovat u daného slova. Při výpočtu této pravděpodobnosti vycházíme z četností výskytů podobných kombinací *Slovo + značka* ve výchozím korpusu. Pokud bychom „vytvářeli“ značky ručně, nebylo by možné pravděpodobnost spočítat tak, aby odpovídala hodnotě statisticky vypočtené. Rovněž je třeba poznamenat, že cílem práce bylo vytvořit pokud možno univerzální nástroj. A pokud bychom měli psát rozsáhlý soubor

pravidel pro odvozování značek pro každý jazyk (nebo i každý nový systém značkování) zvlášť, požadavek na univerzálnost by určitě splněn nebyl.

Po přečtení předchozího odstavce by bylo možno se domnívat, že by se dal vytvořit systém aspoň několika pravidel, která by mohla hledání odpovídající značky alespoň částečně usnadnit. Ale kromě výše popsaných problémů s výpočtem pravděpodobnosti páru *Slovo + značka* by mohlo jakékoli ručně vytvořené pravidlo vést k neočekávaným komplikacím. U tak rozmanitého jazyka, jakým čeština nesporně je, se vždy najde určitý jazykový jev (slovo či nějaký jeho tvar), se kterým nelze dopředu počítat. Stejně tak nelze všechny možné výjimky ošetřit. Pro lepší porozumění uvedu několik názorných příkladů:

- Víme, že slovo, pro které hledáme značku, je přídavné jméno (např. podle typické koncovky *-ý,-á,-é* pro tvrdá přídavná jména). Dále víme, že začíná na *nej-*. Podle známých pravidel by se tedy mohlo jednat o 3. stupeň přídavného jména.
  Nicméně čeština obsahuje slova jako: *nejedlý, nejapný, nejasný, ...*
- Příklad opačný. Slovo začíná na *ne-*, ale neobsahuje předponu *nej-*. Vcelku logická úvaha by byla: a) odtrhnout předponu *ne-*, b) zkusit najít slovo bez této předpony, c) pokud slovo najdeme, změnit značku v 11. pozici na *N* (negativní forma slova).
  Zatímco takový postup by mohl fungovat pro přídavná jména *nehezký, hezký*, problém by mohl nastat u slova *nesen*. Domnělý kladný tvar *sen* je rovněž platné slovo, nepochybně i s odpovídající značkou. Zatímco v prvním případě se ale jedná o trpný rod slovesa nést, v druhém jde o výraz pro vidinu nebo zdání. Došlo by tak k vytvoření naprosto nesmyslné značky. Další taková kombinace je třeba *nerudná* (nevrlá, mrzutá) a *Rudná* (název několika obcí v ČR).

Výjimku při určování značek tvoří pouze interpunkční, matematická a jiná znaménka (značka `Z:----------`) a číslo (resp. číslovka) psané číslicemi (značka `C=----------`). Tyto dvě skupiny jsou velice snadno odhalitené díky ASCII[2] hodnotě svých znaků. Vyhledávat zástupce těchto dvou skupin mezi ostatními slovy korpusu je tak vcelku zbytečně, i když je to samozřejmě možné. Zatímco u interpunkce by bylo hledání poměrně jednoduché, neboť se stejná znaménka hojně opakují v jakémkoli textu, pravděpodobnost nalezení dvou stejných čísel už by byla znatelně nižší.

[2]American Standard Code for Information Interchange, česky „americký standardní kód pro výměnu informací“

5.3 Hledání podobných slov

V úvodu této kapitoly jsem zmínil, že vhodná značka pro slovo je vybírána na základě jeho podobnosti se slovy obsaženými ve slovníku. Nyní bych tento velmi zjednodušený popis rád rozvedl. Kapitola 2.2.1 uvádí několik typů morfů (částí slov). Zatímco kořen slova vypovídá zejména o významu slova a prefix slouží především k odvozování nových slov, je to hlavně sufix, který přímo koresponduje s mluvnickými kategoriemi daného tvaru slova.

5.3.1 Koncovky a přípony

Ne nadarmo se označují koncovky podstatných jmen jako pádové. Tvar přídavného jména pak koresponduje se jménem podstatným, podle kterého se rovněž řídí koncovka. Ač čeština rozlišuje u podstatných jmen pouze 14 základních vzorů, věnuje jejich skloňování [Havránek(1960)] 45 stran (což značí, že jednoduché zdaleka není).

Skloňování přídavných jmen, zájmen a číslovek podléhá pravidlům o něco jednodušším, než je tomu u jmen podstatných. To znamená, že koncovky nejsou natolik rozmanité, a že jejich různých tvarů se vyskytuje adekvátně menší množství. U sloves je situace opět o něco složitější, neboť čeština u nich rozlišuje mluvnické kategorie: osoba (první, druhá třetí), číslo (jednotné, množné, způsob (oznamovací, rozkazovací, podmiňovací), čas (přítomný, minulý, budoucí), rod (činný, trpný), vid (dokonavá, nedokonavá slovesa), třída (pět tříd, celkem 14 slovesných vzorů). Pokud ještě přidáme tvary sloves jako jsou přechodníky, jedná se o velmi rozmanitou paletu slovních tvarů.

Dá se říci, že většina slov z kategorie ohebných slovních druhů podléhá pravidlům pro skloňování či časování (příp. stupňování). To znamená, že existuje relativně dobrá šance na jejich rozpoznání dle koncovky. Toho je samozřejmě využito při odhadu morfologických značek.

5.3.2 Neohebné slovní druhy

U neohebných slovních druhů (příslovce, předložky, spojky, částice, citoslovce) je situace poněkud jiného rázu. U neohebných slovních druhů nemůžeme počítat se snadnou identifikací tvaru slova podle koncovky. Na druhou

stranu se jedná o slova, která se v textu vyskytují poměrně hojně, což platí především pro spojky a předložky, takže s nalezením odpovídající značky většinou není problém. Relativně častý je rovněž výskyt většiny příslovců.

U částic platí to, že ač se nevyskytují v textu zcela běžně, není jich mnoho. [Havránek(1960)] uvádí jako nejběžnější částice *at', necht', kéž, což, copak*. Tento fakt opět znamená, že nalezení odpovídající značky by mělo být poměrně jednoduché, protože v rozsáhlém korpusu by se tyto částice měly vyskytovat. [Havránek(1960)] dále uvádí, že charakteru částic mohou v určitých případech nabývat vybrané spojky či způsobová příslovce, nicméně zde už se jedná o velmi okrajové případy.

Citoslovce jsou pak slova velice charakteristická, u nichž je jakákoli identifikace nesnadná, musíme se proto spoléhat na jejich výskyt v korpusu, ze kterého čerpáme morfologické značky.

5.3.3 Zkratky a cizí slova

Opět specifická skupina slov, která se v textu (zejména odborném, či v žurnalistice) vyskytují poměrně hojně. Bohužel u nich nelze zavést žádné univerzální pravidlo pro určení morfologické značky. Pro úplnost jenom uvedu několik příkladů zkratek:

- **zkratková slova** z nichž některá podléhají skloňování např. *Čedok*, pak budou s největší pravděpodobností označena jako podstatná jména na základě koncovky.
- **zkratky čistě grafické** mohou být speciální znaky abecedy např. *V* pro Volt (bude nejspíše považováno za předložku), *M* pro mega, pak pro ně platí v podstatě totéž co pro citoslovce (záleží, zda se vyskytují v korpusu, ze kterého čerpáme).
- zbylé zkratky a zkratková slova, u nichž opět záleží pouze na tom, zda už jsme se se zkratkou setkali ve zdrojovém korpusu, či nikoli.

Obecně pak platí, že zatímco člověk zkratku pozná na první pohled (aniž by musel nutně znát i její význam), počítač s ní zachází jako s kterýmkoli jiným slovem. Stejně tak je tomu i v případě cizích slov, která většinou nepodléhají pravidlům deklinace (skloňování) tak jako česká podstatná jména.

Oproti tomu zdomácnělá podstatná jména a slovesa často používají české koncovky společně s cizím tvarem kořenu slova.

5.3.4 Koncovka x Zakončení

V předchozích několika odstavcích se zmiňuji o rozpoznání podobných slov podle stejných koncovek či přípon. Je ale třeba podotknout, že vnímání koncovek a přípon je vlastní pouze lidem znalým pravidel pro dělení slov v češtině. Zatímco oni vnímají slovo jako prostředek jazyka složený z několika morfologických elementů, počítač „vnímá“ slovo jen jako posloupnost znaků.

Takže zatímco pro člověka není problém rozdělit například podstatné jméno na kořen, příponu a koncovku, a následně porovnat koncovku s koncovkou jiného slova, u počítače to není tak snadné. Strojové pojetí se musí spokojit s odtržením několika posledních znaků. Takovýto podřetězec slova je možno nazvat *zakončení*. V následujících kapitolách se tak mohou pojmy *zakončení* a koncovka zaměňovat, neboť bude kladen důraz na jejich vnímání především z pohledu počítačového zpracování, nikoli jazyka.

Fakt, že počítač neumí rozlišit mezi koncovkou, příponou či kořenem slova, je z jistého hlediska velmi omezující. Při automatickém zpracování není možné věnovat se každému slovu zvlášť, rozebrat ho, odtrhnout koncovku a tu pak vyhledávat mezi slovy známými. Je třeba proces změnit na vyhledávání zakončení proměnlivé délky, neboť koncovku může tvořit jedna hláska (resp. znak), stejně jako skupina tří i více hlásek (znaků).

Ač má takový postup mnoho nevýhod (hledáme skupiny znaků, které jsou ve skutečnosti kratší či naopak delší než vlastní koncovka), přináší s sebou i několik výhod, například při vyhledávání značek pro jednoslabičná slova, podstatná jména rodu mužského neživotného v 1. či 4. pádu (obecně slova končící souhláskou). Tato slova totiž obvykle žádnou koncovku nemají, takže podle ní vyhledávat nemůžeme. Na rozdíl od hledání dle koncovky tak hledání podle zakončení představuje prostředek, kterým se opravdu dá najít slovo se stejnou značkou. Nejčastěji k tomuto jevu dochází u slov, která se v češtině rýmují.

Příklad:

Hledáme značku ke slovu *mrak*

Koncovka je u tohoto slova nulová, avšak při hledání zakončení *-ak* je nalezeno slovo *vlak*. Jde sice o slova naprosto nesouvisející, nicméně si lze všimnout, že jsou obě rodu mužského neživotného, sdílejí vzor *hrad* zda se jedná o 1. či 4. pád slova sice nejsme schopni rozlišit. Každopádně lze ale usuzovat, že pro takový tvar slova *mrak* bude možné použít stejnou morfologickou značku jako pro slovo *vlak*.

5.4 Metody odhadu značek

5.4.1 Známá a neznámá slova

Při hledání vhodné značky pro slovo mohou obecně nastat celkem tři případy:

1. Hledané slovo (resp. daný tvar) je obsaženo ve zdrojovém korpusu. Pak je snadno nalezena odpovídající značka. (Nicméně bez rozlišení kontextu může jít jen o jednu z několika správných značek.)
2. Hledaný tvar slova se ve slovníku sice nevyskytuje, ale na základě zakončení jsme schopni přidělit jednu či více značek, které by mohly být správné.
3. Hledaný tvar slova není ani obsažen v korpusu, ani se k němu nepodařilo najít značku na základě zakončení.

Je zřejmé, že slova z 1. kategorie jsou případ ideální a nejjednodušší. I když nemusíme znát všechny možné značky pro dané slovo, minimálně jednu značku pro slovo nalezneme bez větší námahy.

Ale jsou to právě slova ze 2. a 3. kategorie, která označujeme jako **neznámá**. A právě na nalezení značek pro tato slova (zejména slova ze 2. kategorie, protože u slov ze 3. kategorie neexistuje žádný spolehlivý postup pro nalezení značky) se program zaměřuje. V následujících odstavcích budou popsány jednotlivé typy hledání značek dle zakončení rovněž i postup pro slova ze 3. kategorie.

5.4.2 Tvorba slovníku

Prvním krokem při hledání morfologických značek je vždy vytvoření tzv. slovníku. Ten slouží v podstatě jako databáze slov, mezi kterými hledáme a se kterými následně pracujeme. V podstatě se jedná o původní korpus (PDT či SNK) přetvořený do takové podoby, která umožňuje pohodlné vkládání slov a prohledávání.

Díky tomu, že používané korpusy obsahují jak formu slova se značkou, tak i lemma odpovídající danému slovu, můžeme vytvořit další formu slovníku, ve které spojíme různé tvary slova, které mají společné lemma. To nám umožňuje lepší rozlišení hledaných tvarů slov, neboť tímto způsobem je možné na základě společného lemmatu extrahovat různá zakončení téhož slova a k nim náležející značky.

5.4.3 Slova známá

Jak již bylo zmíněno v předcházejících odstavcích, nejjednodušší určení značky je u těch slov, která jsou obsažena ve slovníku. První částí testovacího cyklu je hledání slov právě tam. Tímto způsobem nalezneme značky zejména ke slovům hojně používaným, tzn. zejména ke spojkám, předložkám, nejčastěji používaným podstatným a přídavným jménům, běžně se vyskytujícím tvarům sloves.

Nelze ovšem vyloučit, že k danému tvaru slova neexistuje ještě nějaká další platná značka. Proto rozlišujeme dvě kategorie známých slov podle jejich četnosti ve výchozích datech nalezených značek:

1. Daný tvar slova se vyskytuje ve slovníku často (např. 1000×), pokaždé se stejnou značkou. Pak máme důvod se domnívat, že správná bude právě tato značka, a nemá velký význam zkoušet takovému slovu přiřadit další značky.

2. Daný tvar slova se vyskytuje ve vstupních datech velmi řídce. Pravděpodobnost, že by k němu mohla existovat i jiná možná značka je tedy poměrně vysoká, proto zkoušíme přidat další značky na základě jeho zakončení.

5.4.4 Využití N-gramů

Při odhadu morfologických značek se zaměřujeme na výskyt charakteristických skupin hlásek ve slovech. U českého jazyka, zejména pak u řeči, si lze snadno všimnout, že některé skupiny hlásek (zobecněně vzato slabiky) se vyskytují ve slovech společně daleko častěji než jiné. Vzhledem k tomu, že většina slov vzniká v češtině odvozováním, je výskyt bigramů daleko častější než je tomu u trigramů. Při vytváření nových slov totiž často dochází k tomu jevu, že se za tzv. slovotvornou příponu naváže ještě další postfix, čímž vzniká právě bigram. Oproti tomu trigramy jsou doménou typickou spíše pro jazyky, ve kterých vzniká většina nových slov skládáním (např. němčina, švédština)

Typickým příkladem bigramů jsou slova v nichž se vyskytují tzv. přípony živé a produktivní, jak je nazývá [Havránek(1960)]. Mezi tyto přípony patří:

- přípona *-tel* či *-el* u podstatných jmen rodu mužského a následně ve slovech: *ředitelství, učitelský, kazatelna, ...*
- přípona *-ost* či *-st* u podstatných jmen rodu ženského a následně ve slovech: *radostně, ctnostný, blbostmi, ...*

Hledání bigramů je další oblast, ve které nám paradoxně prospívá to, že počítač nerozlišuje přípony a koncovky tak, jak to vnímá česká jazykověda. Díky tomu je možné najít daleko větší počet charakteristických posloupností znaků, které by jinak zůstaly bez povšimnutí (např. *pian-ist-ka*). Obecně pak hledáme dvojice podřetězců na koncích slov délky: 3+4 znaky, 3+3 znaky, 3+2 znaky, 2+3 znaky. Hledání podřetězců kratších už nemá zcela takový význam. Ne, že by se takové bigramy v češtině nevyskytovaly, nicméně takovou skupinu slov budeme vnímat jako jedno společné zakončení. Naopak hledání delších skupin nemá větší význam z důvodu jejich řídkého počtu v češtině (i slovenštině).

Prakticky není možné nashromáždit seznam všech možných bigramů, které by se v češtině mohly vyskytovat. Ale pro potřeby odhadu morfologických značek stačí několik desítek nejčastěji používaných. Najdeme-li pak takový bigram ve zkoumaném slově, dokážeme už velmi snadno přiřadit patřičnou značku. Přitom to, že se ve slově vyskytuje známá forma bigramu je pro počítač daleko směrodatnější údaj než výskyt obyčejného zakončení.

5.4.5 Zakončení získaná využitím lemmat

Fakt, že známe u všech slov obsažených v korpusu i lemma (základní tvar) můžeme využít při hledání zakončení zkoumaného slova. Základem je nalezení všech možných tvarů se shodným lemmatem, které se v korpusu vyskytují. V okamžiku, kdy nalezneme dva nebo více tvarů slova se společným lemmatem, můžeme jednoznačně spojit zakončení daného slovního tvaru se značkou, která je tomutu tvaru přidělena v korpusu. Pokud má testované slovo stejné zakončení jako některý slovní tvar z této množiny, existuje jistá pravděpodobnost, že bude mít rovněž stejnou značku.

K samotnému nalezení potřebných zakončení slov používáme proces hledání nejdelšího společného prefixu (dále LCP[3]) u daných slovních tvarů. Přičemž musí platit, že se jedná o řetězec znaků začínající prvním písmenem obou slovních tvarů. Pro úplnost dodám, že více než samotné znaky LCP nás zajímá především jeho délka. Zbytek slova za touto částí (znaky za pozicí odpovídající délce LCP) pak označíme jako ono hledané zakončení. Proces hledání LCP probíhá tak, že se postupně porovnávají jednotlivé znaky slova v pořadí od prvního k poslednímu. Tento postup je znázorněn na následujícím příkladu:

Lemma: *lahodný*, Tvary: *lahodnou, lahodnějšího*

l	a	h	o	d	n	o	u
l	a	h	o	d	n	ě	jšího
✓	✓	✓	✓	✓	✓	×	dále nezkoumáme

Společná část	**Zakončení**
lanodn	-ou
lahodn	-ějšího

Zakončením *-ou, -ějšího* tedy můžeme přidělit odpovídající značky. Pokud postupně projdeme všechna lemmata obsažená v daném korpusu, můžeme následně vytvořit seznam takto získaných zakončení (s patřičnou značkou a informací o četnosti výskytu v korpusu). Při setkání s neznámým slovem pak vyhledáme jeho zakončení právě v tomto seznamu a na základě toho mu můžeme přidělit značku (resp. značky).

Algoritmus hledání LCP funguje u většiny zkoumaných slov vcelku spolehlivě, nicméně právě zde se projevuje jev zvaný *morfologická alternace* (po-

[3]z angl. Longest common prefix

psaný v kapitole věnované morfologii). Právě změna hlásek ve vnitřní části slova může způsobit nemalé potíže. Tento problém je patrný třeba na následujícím příkladu:

Lemma: *vrah*, tvary: *vrahem, vrahovi*

v	r	a	h	e	m
v	r	a	h	o	vi
✓	✓	✓	✓	×	dále nezkoumáme

Ale u tvarů množného čísla: *vrazích, vrahy*

v	r	a	z	ích
v	r	a	h	y
✓	✓	✓	×	dále nezkoumáme

Pokud máme k dispozici pouze poslední dva tvary, není možné záměnu hlásek odhalit. Jako LCP pak bude zvolen řetězec *vra*, jako zakončení pak řetězce *zích, y*, kterým budou rovněž přiřazeny odpovídající značky. Vzhledem k tomu, že zkoumáme zakončení z hlediska strojového zpracování, nikoli jako koncovky z hlediska jazyka, může se stát, že během prohledávání celého korpusu narazíme na další výskyty shodných zakončení.

Bohužel, stejně tak může dojít k tomu, že hledání LCP povede k tomu, že získáme velké množství různých zakončení. Při zkoumání četnosti výskytu zakončení nám samozřejmě vyjdou čísla nižší, než kdybychom odtrhli samotné koncovky (v tomto případě *ích, y*).

O něco pozitivnější situace nastává v případě, kdy máme k dispozici více tvarů slova. Pakliže jako první dva tvary zkoumáme ty, ve kterých k morfologické alternaci nedochází (např. *vrahem, vrahy*, LCP těchto slov bude mít délku „klasickou" (v našem případě 4). Pokud jsou i ostatní tvary slova dostatečně dlouhé, můžeme odtrhnout podřetězec této stanovené délky bez ohledu na samotný tvar slova. Nicméně tento postup může vést pro změnu k odtržení části skutečné koncovky slova, čímž získáme zakončení o něco kratší, než by bylo vhodné. Jak je vidět, nelze určit postup jednoznačně ideální.

Při využívání lemmat se logicky nabízí i takový postup, kde by byl porovnáván použitý slovní tvar se samotným lemmatem. Nicméně je třeba mít na paměti, že lemma kromě základního tvaru slova často obsahuje vysvětlivky a různé doplňující údaje (např. u slova *přijetí* je uvedeno lemma `přijetí-2_^(např._návrh)_(*5mout-2)`). Stejně tak proti tomuto plánu hraje fakt, že lemma jakožto základní tvar je svým způsobem „vytrženo z reality".

To znamená jednak skutečnost, že se v textu může vyskytovat mnohem řidčeji než různé od něj odvozené tvary. Dále je u lemmat relativně častým jevem to, že z hlediska hlásek (resp. znaků) nemá s vyskloňovanými tvary lemma nic společného. Příkladem toho, že lemma slova a nejrůznější slovní tvary spolu souvisejí pouze po stránce významové, jsou zejména zájmena. Příkladem může být hojně používané zájmeno *já*, které se ale kromě 1. pádu vyskytuje ve tvarech: *mě/mne, mně/mi, mnou*.

5.4.6 Nejčastější zakončení

Dalším, a zřejmě nejjednodušším postupem, jak vyhledat zakončení zkoumaného slova, je vytvořit databázi nejfrekventovanějších zakončení slov obsažených v korpusu. Pro zjednodušení chápeme v tomto bodě pojem různá zakončení jako různé páry *zakončení + značka*, nikoli jako samotné řetězce znaků.

Vzhledem k tomu, že korpusy obsahují 1,2 - 1,5 milionů hesel, mohlo by se zdát, že rovněž rozličných zakončení bude velký počet. Ač je český jazyk velice rozmanitý a obsahuje velké množství výjimek a nepravidelných tvarů,

různých zakončení se v něm pouze několik desítek až stovek (v závisloti na délce zakončení). Opět je ale třeba mít na paměti, že k jednomu zakončení může náležet větší počet značek, což počet možných dvojic několika násobně zvyšuje.

Takto získaná zakončení rozdělujeme do skupin podle délky (1-4 znaky). Platí, že při testování neznámého slova postupujeme při hledání v databázi zakončení od největší délky k nejmenší. Samozřejmě záleží také na tom, zda to umožňuje délka slova a vlastní zakončení daného tvaru. Jistou náhradou za hledání delších zakončení může být hledání bigramů, popsané výše, které do určité míry hledání zakončení supluje.

5.4.7 Zbylá slova

Jednopísmenných slov se v češtině vyskytuje celkem 12, přičemž se jedná většinou o předložky, takže použití v textu je časté a s dohledáním značky nebývá problém. Podobně je tomu například i u dvoupísmenných tvarů některých zájmen. Oproti tomu existují dvoupísmenná podstatná jména (např. *oj, oř, úl*), která nejsou používaná nijak zvlášť často. Jak je navíc patrné, u těchto tvarů nám nikterak nepomůže ani vyhledávání značky podle zakončení, neboť mají nulovou koncovku (jsou tudíž velmi špatně identifikovatelné).

U takových tvarů je nutné přidat značku „ručně“. Nejedná se o to, že by bylo každé takové slovo posuzováno zvlášť. Celý postup spočívá v tom, že slovu přidělíme několik (např. 10) značek, které se v celém korpusu vyskytuje nejčastěji. Respektive téměř nejčastěji, neboť je prozíravější zvolit značku pro podstatné jméno, protože na základě textů obsažených v korpusu by mohla být nejpoužívanější značkou například značka pro spojku či předložku.

Problém s určením značky samozřejmě nebývá jen u krátkých podstatných jmen s nulovou koncovkou, ale týká se i nejrůznějších zkratek, neobvyklých slovních tvarů, jako jsou některé velmi řídce používané přechodníky. A stranou nezůstávají ani slova cizí, pro která bychom většinou marně hledali jakékoli české zakončení.

6 Výpočet pravděpodobnosti

Jak uvádějí předcházející kapitoly, morfologickou značku (dále Tag[1]) nelze přidělovat dle libosti, ale musí být přidělena na základě statisticky určených údajů. Na základě četnosti výskytu daných kombinací *Slovo + Tag* a na základě četnosti výskytu jednotlivých značek pak můžeme odvodit pravděpodobnost přidělené značky.

6.1 Určení pravděpodobnosti

Nejjednodušší situace nastává tehdy, pokud je daná kombinace *Slovo + Tag* přímo obsažena ve zdrojovém korpusu. Budeme-li pak chtít spočítat, s jakou pravděpodobností se může vyskytnout dané slovo s danou značkou, bude nám k tomu stačit podíl četností výskytu (značeno $C(Slovo, Tag)$ a $C(Tag)$):

$$P(Slovo|Tag) = \frac{C(Slovo, Tag)}{C(Tag)}$$

6.2 Vyhlazování pravděpodobnosti

Pokud ale korpus danou kombinaci *Slovo + Tag* neobsahuje, narážíme při použití výše uvedeného výpočtu na problém. $C(Slovo, Tag) = 0$, takže:

$$P(Slovo|Tag) = \frac{C(Slovo, Tag)}{C(Tag)} = \frac{0}{C(Tag)} = 0$$

Taková rovnice vyjadřuje, že výskyt dané kombinace *Slovo + Tag* není možný. Takže u jakékoli kombinace, kterou neznáme předem, říká, že neexistuje, není platná. Takový výsledek je v přímém rozporu s cílem této práce, neboť jsou to právě neznámá slova, kterým se snažíme přidělit značku.

Nezbývá tedy nic jiného, než využít nějakou metodu, která by upravila hodnotu četnosti i pro neznámé (neviděné) kombinace *Slovo + Tag*. Takové postupy, které slouží k úpravě četnosti výskytu prvků v nějaké množině, se

[1] angl. výraz pro značku, označení; konkrétně pak „morphological tag“

nazývají *vyhlazování*. Existuje několik desítek různých vyhlazovacích metod, které se liší jak použitými postupy, tak složitostí výpočtů.

V našem případě byla použita tzv. *Good-Turingova vyhlazovací metoda*. Ta vychází z údaje zvaného *četnost četností*, na základě kterého stanovuje určitý odhad pro četnost výskytu prvků dané kategorie. Tento odhad následně použijeme ve výpočtu pro pravděpodobnost namísto původní absolutní četnosti.

Algoritmus pro výpočet odhadu funguje následovně (jak jej popisuje [MacCartney(2005)]):

- Cílem algoritmu je upravit hodnotu četnosti slov, která se objevila v trénovacích datech právě $r+1$ krát, tak, abychom získali upravenou hodnotu četnosti slov, která se objevil právě r-krát. (V podstatě jde o to, že u slov, která se objevil $r+1$-krát hodnotu četnosti snížíme, a o získaný počet přerozdělíme mezi slova, která se objevila r-krát, $r-1$ krát, ...)
- Především nás pak zajímá taková úprava, která by pomohla určit odhad četnosti pro slova, která se vyskytla v trénovacích datech 0-krát (tedy vůbec) na základě hodnoty četnosti slov, která se objevila právě jednou.
- Pro každou četnost r stanovíme odhad r^*:

  $$r^* = (r+1)\frac{n_{r+1}}{n_r}$$

  kde n_r je počet různých slov, která se v trénovacích datech (korpusu) vyskytla právě r-krát.
- Pak dostáváme:

  $$p_{GT}(x : C(x) = r) = \frac{r^*}{N}$$

  kde $C(x)$ je hodnota četnosti jevu (slova) x
- Logicky platí, že celkový počet slov v trénovacích datech je:

  $$N = \sum_{r=0}^{\infty} r^* n_r = \sum_{r=1}^{\infty} r n_r$$

Při výpočtu pravděpodobnosti dvojice *Slovo + Tag* tedy nepoužíváme absolutní hodnotu četnosti ve vstupních datech, ale odhad vypočtený pomocí Good-Touringova vyhlazovacího algoritmu:

$$P(Slovo|Tag) = \frac{C^{GT}(Slovo, Tag)}{C(Tag)}$$

6.3 Úprava pravděpodobnosti

Při přidělování značky dle zakončení je také třeba zohlednit fakt, že jedno zakončení se může vyskytovat ve vstupních datech s více možnými značkami. Pokud se v korpusu pojí 90% zakončení *-ný* se značkou pro přídavné jméno (např. *krásný, slunný*) a 10% se značkou pro jméno podstatné (např. *strážný hajný*), musí se to zákonitě projevit i na výpočtu pravděpodobnosti při přidělování značky.

Pokud bychom počítali pouze s četností výskytu zakončení *-ný* bez ohledu na značku, vedlo by to ke špatným výsledkům. Proto je třeba četnost páru *Slovo + Tag*, tj. $C(Slovo, Tag)$ přenásobit odpovídajícím koeficientem (v našem případě 0,9 u přídavného jména; 0,1 u jména podstatného).

Výhoda tohoto postupu spočívá v tom, že pokud se daná kombinace *Slovo + Tag* v korpusu vyskytuje, bude mít zákonitě vysokou pravděpodobnost $P(Slovo|Tag)$. Pokud se v korpusu nevyskytuje, bude pravděpodobnost přidělené značky nejvyšší u nejčastější kombinace *zakončení + Tag*. U kombinací, které se vyskytují řidčeji, pak bude pravděpodobnost odpovídajícím způsobem nižší. Tento fakt pak zajistí, že pravděpodobnost $P(Slovo|Tag)$ bude u jednotlivých přidělených značek patřičným způsobem odstupňována.

6.4 Protřídění přidělených značek

Díky způsobu, kterým přidělování morflogických značek funguje, by mohlo dojít k paradoxní situaci, že známému slovu (obsaženému) budou přiděleny pouze dvě značky, zatímco slovu neznámému několik desítek. Zdaleka se nejedná o ojedinělé případy. Například výskyt koncovky *-í*, která je vlastní jak měkkým přídavným jménům (vzor *jarní*) v rozličných pádech či rodech, podstatným jménům se vzorem *stavení*, tak i slovesům (4. slovesná třída, vzory

prosí, trpí, sází).

I zde nám slouží výpočet pravděpodobnosti výskytu daných kombinací *Slovo + Tag*.

a) Na základě hodnoty pravděpodobnosti můžeme vyřadit ty přidělené značky, které nesplňují určitou mez hodnoty pravděpodobnosti.

b) Případně můžeme místo meze absolutní použít mez relativní (např. hodnotu průměru pravděpodobností, či její násobek).

c) Rovněž lze vybrat předem stanovený počet značek, které budou použity (např. maximálně 15 značek nalezených na základě lemmat, max. 10 na základě nejčastějších zakončení).

d) Eventuelně je možné použít kombinaci výše uvedených způsobů.

Je nutné zmínit, že ani proces vyřazování značek také není zcela bezproblémový. Pokud se ve zdrojovém korpusu bude vyskytovat hledané zakončení velmi často, bude mít logicky zvýšenou pravděpodobnost i při přidělování značky a naopak. Výběrem značky čistě na základě pravděpodobnosti pak dochází k tomu, že slovu je přiřazena značka nesprávná (někdy i zcela nesmyslná) právě proto, že se s daným zakončením v korpusu vyskytuje mnohem častěji než značky správné. Pokud však dojde k přidělení správných značek, není už takový problém, že jsou společně s nimi přiděleny i značky nevhodné. Neboť výběr té nejvhodnější značky se koná až v dalších částech programu na základě kontextu. A pokud je při odhadu vrácena ona správná značka, existuje vysoká pravděpodobnost, že bude vybrána právě i na základě kontextu.

K opačnému problému - tj. vyřazení správné značky z důvodu jejího velmi řídkého výskytu ve zdrojovém korpusu na úkor značek častějších, byť špatných, může samozřejmě také dojít. Ale v takovém případě se jedná o velmi vzácnou situaci. I když nikoli zcela nemožnou.

7 Implementace a funkcionalita

Nejzásadnějším bodem implementace je nesporně fakt, že cílem této práce nebylo vytvořit samostatně stojící program, ale naopak zakomponovat odhad morfologických značek do již fungujícího programu *HMMTagger* Ing. Tomáše Brychcína. Jak je patrné z názvu programu, právě tento program pracuje se skrytými Markovovými modely, a slouží k analýze textu. Odhad morfologických značek je pak pouze jedna ze součástí komplexního celku.

Na druhou stranu tato součást funguje do značné míry autonomně. Jejím hlavním úkolem je pro zadané slovo (*Observation* - pozorování) vrátit pole všech značek (*Nodes* - stavů HMM), které přicházejí pro daný tvar v úvahu. Rovněž je třeba určit, jaká je pravděpodobnost výskytu přiřazené značky u zkoumaného slova, tj. pravděpodobnostní rozdělení B z kapitoly 4.1 věnované skrytým Markovovým modelům.

Samotná činnost části pro odhad značek se skládá ze dvou fází:

1. V první fázi je třeba zpracovat data ze vstupních korpusů do podoby, která umožní následné zpracování dat a vyhledávání v nich. Tato fáze inicializace se skládá z několika dílčích kroků:
   (a) Je třeba vytvořit slovník známých slovních tvarů s odpovídajícími značkami. Tedy převést vstupní XML soubory do formy umožňující prohledávání.
   (b) Dále nalézt všechny možné slovní tvary, které mají společné lemma. Tento a první krok je možné provádět zároveň při parsování vstupních XML souborů.
   (c) Z tvarů získaných v předchozím kroku extrahovat jednotlivá zakončení a jim odpovídající značky.
   (d) Získat nejčastější bigramy, kterými končí slova, stejně tak nejčastější zakončení.
   (e) Provést výpočty nutné pro Good-Turingovo vyhlazování.
2. Druhá fáze pak zahrnuje samotné vyhledávání značek a výpočet pravděpodobnosti $P(Slovo|Tag)$. Přičemž proces vyhledávání značky je podrobněji popsán v posledním oddílu této kapitoly, zatímco výpočtem pravděpodobnosti se zabývá kapitola předchozí, kde jsou k nalezení patřičné matematické vztahy.

Výpočet pravděpodobnosti je proveden již při hledání vhodné značky. Následně jsou nalezené značky společně s přidělenou pravděpodobností uloženy do datové struktury. V momentě, kdy potřebujeme zjistit pravděpodobnost $P(Slovo|Tag)$, nahlédneme do této struktury. Požadovanou kombinaci buď nalezneme (a máme i pravděpodobnost), či využijeme výpočtu pomocí Good-Turingova odhadu.

7.1 Výchozí data

Před samotným popisem programu bych rád věnoval několik odstavců popisu vstupních dat programu, neboť právě jim je implementace do značné míry podřízena. Vstupními daty pro odhad značek jsou obsáhlé korpusy - jeden pro češtinu, druhý pro slovenštinu. Oba slovníky mají strukturu XML[1] souborů.

7.1.1 PDT

Pražský závislostní korpus (dále PDT[2]) je tvořen souborem ručně označkovaných textů (např. novinových článků). PDT obsahuje přibližně 1,5 miliónu jednotek doplněných o morfologickou značku a lemma. Více informací je možno nalézt na [Hajič(2006)]. Tento korpus pak sloužil jako hlavní testovací objekt při odhadu morfologických značek. Jeho rozsah uvádí následující tabulka (počet slov značí počet jednotek ve slovníku, nikoli různých slov).

PDT	Počet vět	Počet slov
Trénovací data	90 829	1 535 831
Testovací data	25 016	421 416

V PDT jsou rozebrány celé texty tak, že každý odstavec, věta i slovo tvoří samostatné elementy odpovídající úrovně. Každé slovo pak lze jednoznačně identifikovat podle názvu tohoto elementu (který obsahuje číslo souboru, odstavce, strany a slova). Konkrétní element slova může vypadat například takto:

[1] eXtensible Markup Language - rozšiřitelný značkovací jazyk
[2] z angl. Prague Dependency Treebank

```
<m id="m-cmpr9415-049-p3s1Aw10">            - identifikátor elementu
   <src.rf>manual</src.rf>                  - označení zdroje textu
   <w.rf>w#w-cmpr9415-049-p3s1Aw10</w.rf>   - označení slova
   <form>největší</form>                    - konkrétní tvar ve větě
   <lemma>velký</lemma>                     - lemma (základní tvar)
   <tag>AAMS1----3A----</tag>               - morfologická značka
</m>                                        - konec elementu slova
```

7.1.2 SNK

Druhým použitým zdrojem je Slovenský národný korpus (SNK), konkrétně 3. verze ručně morfologicky anotovaného korpusu. Autoři [SAV(2010)] uvádějí, že je korpus tvořen z 44,3% uměleckých, 36,7% publicistických a 19,0% odborných textů. Celkově pak obsahuje 1 207 813 tokenů (jednotek).

SNK	Počet vět	Počet slov
Trénovací data	59 308	967 672
Testovací data	14 820	240 141

7.2 Použité datové typy

Abychom mohli pracovat se vstupními daty, je třeba je převést do přístupnější formy. Ta musí umožňovat jednak seskupení jednotlivých prvků obsažených ve vstupních souborech podle určitých vztahů (např. získání všech slovních tvarů se společným lemmatem) a dále musí umožňovat snadný přístup k datům a prohledávání (když získáme lemma a slovní tvary, je potřeba extrahovat zakončení).

7.2.1 Hlavní struktura

Během vývoje této práce jsem vyzkoušel a vystřídal několik datových struktur, z nichž některé byly pro potřeby programu vhodnější než jiné. Je třeba mít na paměti, že vstupní korpus obsahuje 1,5 miliónu elementů, které je třeba zpracovat. Zrovna tak je ale třeba pracovat se zakončeními, jejichž počty se pohybují v řádech tisíců. Ve výsledku se mi nejvíce osvědčila jednoduchá struktura kombinace *hash tabulky* se *seznamem*. Ač se toto řešení

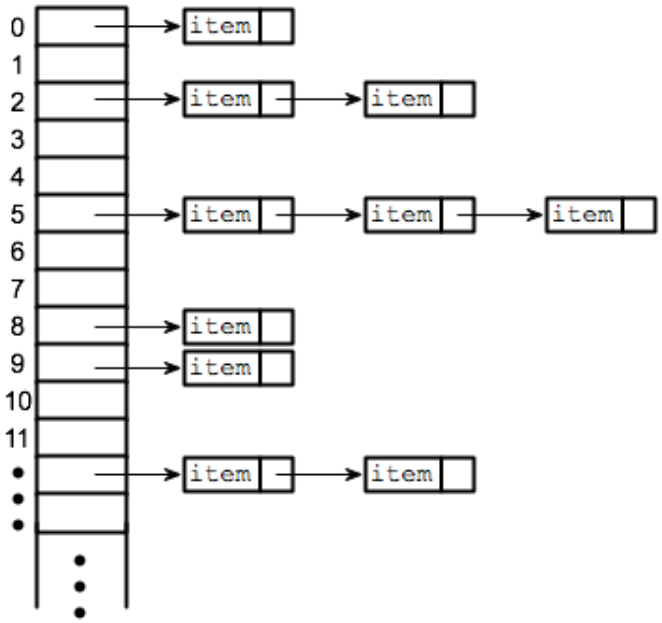
může zdát velice primitivní v době, kdy nativní prostředky jazyka *Java* nabízejí širokou paletu pokročilých datových typů pro práci s objekty, má svoje opodstatnění.

7.2.2 HashMapa x HashTabulka

Při práci s daty většinou potřebujeme nějak postihnout vztahy mezi objekty typu *Klíč (Key)* ⇒ *Položka (Value)*. Přičemž Klíč i položku přitom tvoří výhradně řetězec znaků, tedy typ *String*.

Pokud bychom potřebovali pouze přiřadit, které položky mohou být ve vztahu s daným klíčem, byla by jasným řešením *HashMapa*. Jedná se o jednoznačně nejlepší strukturu pro podobné případy. HashMap má dva volitelné parametry, jedním z nich je počáteční velikost, druhým je hodnota *load factor* (míra naplnění, při které dojde k přehashování mapy a původní kapacita se zhruba zdvojnásobí). Pokud tyto parametry nezměníme, kapacita HashMapy se navýší při 75% naplnění.

Oproti tomu u obyčejné hashTabulky je velikost pevně dána již na začátku a během naplňování se již nemění. Na jednu stranu to znamená menší přizpůsobivost, na stranu druhou méně režie a neměnné pořadí klíčů. Použitá hashovací funkce dělí součet dekadických hodnot prvních znaků (resp. znaku) klíče zvolenou konstantou (např. 256). Názorný příklad HashTabulky představuje následující obrázek:



Obrázek 7.1: Struktura HashTabulky

Pro výpočty pravděpodobnosti potřebujeme znát nejen, jaké položky patří k danému klíči, ale je třeba znát i absolutní četnost, s jakou se ta která kombinace vyskytuje. Pokud bychom použili HashMapu pro mapování <*String, String*>, bylo by třeba vytvořit další strukturu, ve které by byly uchovány četnosti, a to pro každou vytvořenou mapu. Jak je vidět, nejedná se zrovna o ideální řešení.

Pokud bychom vytvořili objekt typu *String + Četnost* a vytvořili HashMapu těchto objektů, bylo by třeba přepsat všechny metody pro vkládání, porovnávání a hledání prvků, což by snížilo přínos použití zavedeného datového typu na minimum. Avšak pokud vytvoříme hashovaný seznam klíčů tak, že každý prvek bude obsahovat seznam hodnot, které k němu patří, je to řešení stejně účinné. Navíc můžeme využít toho, že četnost klíče je rovna součtu četností hodnot. Tím pádem snadno zjistíme relativní četnost jakékoli hodnoty (neboť platí, že rel. *četnost = četnost hodnoty / četnost klíče*).

Typ základní datové struktury je tedy jasný, pojmenování pro ni bylo použito na základě její funkce - *Slovnik*. Jednotlivé položky pak obsahují dva atributy - jeden typu *String* (pro tvar slova, lemma, značku, koncovku - využití je univerzální), druhý typu *Integer* pro četnost.

7.3 Příklad hledání značky

Podle čeho jsou testovaným slovům přidělovány morfologické značky, je popsáno v kapitole 5. Stejně tak je zřejmě jasné, jak funguje vytvoření slovníků známých slov, zakončení na základě lemmat, bigramových zakončení a nejčastějších zakončení. Ale rád bych popsal, jaký je vlastní průběh hledání značky, tudíž hlavní činnost programu.

Příklad 1.

Je volána metoda *getPossibleNodesForObservation* třídy *NiklObservationEstimator* ve které je parametrem *Observation* slovo `zlá`.

1. Zkoušíme hledat `zlá` mezi známými slovy. Nalézáme slova `zlo`, `zlými`, `zlého`, nikoli však potřebný tvar.
   Získáno 0 značek

2. Slovo `zlá` je příliš krátké, než abychom mohli použít vyhledávání bigramového zakončení.

3. Zkoušíme hledat koncovku slova `zlá` mezi zakončeními získanými z lemmat. Kvůli délce slova hledáme pouze zakončení délky 1, tedy `á`. Díky tomu, že je to běžná koncovka, najdeme dvě různé značky. Jednu pro přídavné jméno, rod ženský, 1. pád singuláru (může pocházet např. ze slova hezk*á* (dívka)). Druhou pro sloveso, 3. osoba, singulár, 5. slovesná třída (např. hled*á*).
   Získány 2 značky

4. Hledáme zakončení slova `zlá` mezi nejčastějšími zakončeními. Kvůli tomu, že jde o slovo velmi krátké, hledáme zakončení pouze délky 1, tedy `á`. Kromě značek nalezených v kroku 2 nalézáme ještě další značky pro přídavné jméno. Tentokráte pro rod střední, 1. a 4. pád plurálu (např. zelen*á* (jablka), modr*á* (auta)). Dále 3 značky pro slovesa, z nichž jedna se od značky z bodu 2 liší pouze slovesným videm (např. nech*á* × necháv*á*). A zbylé značky mají příznak pro slova archaická.
   Získáno 7 značek, z toho 5 nových

5. Provedeme protřídění značek. Zjistíme, že pravděpodobnost archaických sloves je oproti ostatním značkám mizivá, takže je vyřadíme.

6. Metoda vrací pole obsahující celkem pět značek. Správné mohou být teoreticky tři z nich (značky pro přídavné jméno např. *zlá* žena - 1.pád singuláru, nebo *zlá* stvoření - 1. či 4. pád plurálu).

Příklad 2.

Pro zjednodušení: Hledáme značky pro slovo `dům`:

1. Hledáme slovo ve slovníku známých slov. Máme štěstí, jedná se o slovo často používané. Jsou nalezeny hned dvě značky - pro podstatné jméno, rod mužský, 1. a 4. pád čísla jednotného.
   Získány 2 značky
2. Vracíme obě nalezené značky, vyhledávání dalších není nutné.

Příklad 3.

Hledáme značky pro slovo `kontrafagotistkami`

1. Hledáme slovo ve slovníku známých slov. Není to slovo zcela běžné, takže neuspějeme. (Bohužel v korpusu není obsažen žádný text popisující obsazení sekce dřevěných dechových nástrojů v symfonickém orchestru.)
   Získáno 0 značek
2. Hledáme jednotlivá bigramová zakončení. Tedy `istk+ami`, `stk+ami`, `tka+mi`, `tk+ami`. Opět získáme značku pro podst. jméno, rod ženský, 7. pád singuláru. A to hned z několika bigramů (např. femin*ist+kami*, ko*stk+ami*, bo*tka+mi*, marione*tk+ami*). *Získána 1 značka*
3. Hledáme zakončení tohoto slova mezi zakončeními získanými z lemmat. Maximální délka hledaného zakončení je omezena na pět znaků. Hledáme tedy postupně: `tkami`, `kami`. První nalezená značka je pro zakončení `ami`, a to pro podst. jméno, rod ženský, 7. pád plurálu (např. ze slov masy, mas*ami*). Dále nalezneme značku pro `mi`, tentokráte pro přídavné jméno, rod mužský, 7. pád plurálu (silný, silný*mi*). A další dvě značky pro zakončení `i` týkající se podstatných jmen (plavcem, plavc*i*; práce, prác*i*).
   Získány 4 značky, ale jen 3 nové

4. Hledáme zakončení `kami`, `ami`, `mi`, `i`. Zatímco zakončení `kami` a velmi běžné `ami` vedou opět k nalezení již známé značky (pro podst. jméno, rod ženský, 7. pád singuláru), zakončení `mi` vede k získání značek pro přídavné jméno v 7. pádě plurálu pro všechny možné rody (např. krásný*mi* princeznami, statečný*mi* rytíři, nedobytný*mi* hrady, mocný*mi* městy). A zakončení `i` se pak ukazuje jako „nejzrádnější“, neboť se vyskytuje u velkého množství podstatných jmen často hned v několika pádech (např. stroj*i* - 3. a 6. pád singuláru, 7. pád plurálu; zd*i* - 2., 3., 5., 6. pád singuláru, 1., 4., 5. pád plurálu; ...).
   Pro případ jako je tento byla pro hledání možných značek nalezena podmínka, že pokud po hledání mezi nejčastějšími zakončeními délky 2 máme nalezeno 5 či více potenciálních značek, nehledáme již mezi nejčastějšími zakončeními délky jedna. Což vede jednak k urychlení procesu hledání značek ale hlavně ke snížení výskytu nesmyslných kombinací *Slovo + Značka*.
   Získáno 5 značek, z toho 3 nové

5. Při porovnání pravděpodobností jednotlivých značek zjistíme, že pravděpodobnosti pro značky přídavného jména pro rod mužský životný i neživotný jsou velmi nízké, stejně tak pro podst. jméno rodu mužského (získaná z lemmat). Tyto značky vyřadíme.

6. Budou navráceny 4 značky, z nichž je pouze jedna správná (podst. jméno, rod ženský, 7. pád singuláru). Nicméně konkrétně u koncovky *-ami* si můžeme být jisti, že pravděpodobnost správné značky bude oproti ostatním výrazně vyšší, takže ve výsledku bude vybrána.

8 Testování a úspěšnost odhadu

Program pro odhad morfologických značek byl testován na množině souborů, které jsou součástí korpusů PDT a SNK. Konkrétně u PDT se jedná o výběr slov, u kterých je nižší pravděpodobnost výskytu v ostatních souborech korpusu (použitých pro tvorbu slovníku). U SNK pak jde přímo o část korpusu, která je určena pro testovací účely.

Z těchto slov je pak vytvořen jeden soubor, který obsahuje pouze výrazy typu: *Slovo/Značka* (např. `kancelář/NNFS1-----A----`). Přičemž na jednom řádku souboru se vyskytuje nejvýše jedna věta (někdy jde pouze o skupinu slov). Takovýto styl zápisu je podmínečný, protože je takto zpracováván jinou částí programu HMMTagger. A právě na jeho základě jsou vytvářeny pravděpodobnostní modely.

Samotný úspěch či neúspěch odhadu značek je posuzován podle toho, zda je skutečná, správná, značka mezi přidělenými značkami nebo ne. V ideálním případě by se měla shodovat značka přiřazená ručně v korpusu se značkou s nejvyšší pravděpodobností přiřazenou při odhadu.

Jak je v této práci několikrát zdůrazněno, jakož i předvedeno na řadě názorných příkladů, obsahuje čeština řadu výjimek, neobvyklých tvarů a jiných jazykových zvláštností, které mají obecně špatný vliv na účinnost (či samotnou možnost) odhadu značek. Proto je směrodatným údajem pro účinnost odhadu značek nikoli samotná procentuální hodnota, ale je to porovnání účinnosti značení bez použití odhadu a účinnosti při použití této části programu. Právě takové porovnání názorně ukazují následující odstavce.

8.1 PDT

Následující odstavce uvádějí, jakými způsoby je možno ovlivnit jak účinnost programu, tak rychlost jeho běhu. Takové změny jsou pak zachyceny v tabulkách pro možnost porovnání. Podrobnější informace o korpusu PDT lze nalézt v kapitole 7.1.1.

Poznámka

Pro úplnost uvádím, že doba běhu uvedená v tabulkách značí pouze dobu samotného odhadu značek. Dále je třeba poznamenat, že některé hodnoty účinnosti vznikaly postupně, a nemusí v nich být zahrnuta všechna nastavení, která byla shledána jako nejvýhodnější.

Poznámka č. 2

Uvedené hodnoty běhu programu jsou do značné míry relativní, ale byly pořízeny za takřka stejných podmínek. Prostředky pro měření času jsou obsaženy v programu HMMTagger, který rovněž provádí výpočet úspěšnosti značkování. Výpočet probíhal na sestavě: Intel®Core^{TM}i3; 2,3 GHz; 4 GB RAM; Windows 7 64bit.

8.1.1 Velikost Hashtabulky

Prvním parametrem, na který jsem se při testování zaměřil, byl rozsah hashtabulky použité pro vytváření všech datových struktur obsahujících jak různé druhy koncovek, tak reprezentaci vstupních dat z korpusu. Pro jednoduchost byl použit u všech struktur stejný rozsah. Jeho hodnota pak ovlivňuje především rychlost vyhledávání v seznamech a rychlost vkládání.

Použitý rozsah	**Doba běhu** [s]
256	177, 19
384	160, 18
512	145, 26
768	150, 864
1024	157, 98

Jak je patrné, nejlepších hodnot dosahuje program u rozsahu 512, který zaručuje ideální zaplnění hashtabulky. Při rozsahu nižším jsou jednotlivé seznamy přeplněné, což způsobuje delší průchod. Při rozsahu vysokém pak dochází k tomu, že je část tabulky nevyužita.

8.1.2 Počet přidělovaných značek

Dalším velice podstatným parametrem je (maximální) počet značek, které jsou slovu přidělovány na základě konkrétní koncovky. U některých slov může být pro jedno dané zakončení přiděleno až několik desítek značek, což samozřejmě vede k tomu, že ve výsledku je většina přidělených značek nesprávná, tudíž i nadbytečná.

Omezení zaručuje to, že k dané koncovce bude přiděleno maximálně tolik značek, kolik je zadáno (5, 10, ...). Značky jsou vybírány podle přidělené pravděpodobnosti, a uvažujeme, že vyšší pravděpodobnost bude znamenat i vhodnější značku. Tento parametr je možné nastavit zvlášť pro zakončení získaná z lemmat i pro zakončení nejčastější.

z lemmat	**z nejčastějších**	**Úspěšnost** [%]	**Doba běhu** [s]
7	7	91,353	122,34
8	8	91,393	134,51
10	10	91,065	116,44
15	15	91,106	124,64

Tím, že omezíme počet přidělených značek na zakončení, významně urychlíme chod programu. Menší počet přidělených značek znamená rychlejší prohledávání i méně nesprávných výsledků. Nicméně pokud počet přidělených značek omezíme moc, může dojít k tomu, že právě ta vhodná už mezi vybranými nebude.

Dalším logickým krokem bylo omezení počtu přidělených značek nesymetricky, přičemž se ukázalo, že je výhodnější přidělovat méně značek pro zakončení získaná z lemmat, ale více pro zakončení získaná z nejčastějších koncovek. Doba běhu se tím logicky zvýší, ale stejně tak dochází ke zvýšení úspěšnosti značkování.

z lemmat	**z nejčastějších**	**Úspěšnost** [%]	**Doba běhu** [s]
8	10	91,112	105,37
8	15	91,167	118,04
8	20	91,222	134,96
8	25	91,254	143,84
15	25	91,175	149,56

Jako výchozí pak byla použita hodnota max. 8 značek z lemmatických zakončení, max. 25 značek z nejčastějších zakončení. Při dalším zvyšování

dochází k přidělování opravdu nových značek jen zřídka, takže nemá větší význam limit nějak navyšovat.

8.1.3 Omezení hledání

S počtem přidělovaných značek je velmi úzce spojeno i omezení pro samotné hledání značek. Zjednodušeně během hledání spočítáme přidělené značky, a pokud je jich dost, nebudeme hledat další.

Pomocí bigramů jsou u delších slov nalezeny většinou maximálně dvě značky. U kratších slov se nedají použít. Proto vyhedávání zakončení mezi těmi získanými z lemmat provádíme vždy. Po této fázi pak následuje přepočet značek. Pokud jich máme velmi málo (např. méně než pět), provádíme vyhledávání všech možných zakončení. Pokud jich máme o něco více (např. 5-10) vyhledáváme pouze zakončení délky 3 a 4 (pokud je to možné), neboť delší zakončení by mělo znamenat i lepší shodu se značkou. A pokud je značek ještě více, neprovádíme vyhledávání mezi nejčastějšími zakončeními vůbec.

Dolní mez	**Horní mez**	**Úspěšnost** [%]	**Doba běhu** [s]
5	15	91,244	122,67
7	14	91,242	144,05
8	16	91,254	143,84
10	20	91,248	145,26

Ač na samotnou účinnost značkování nemá tento parametr tak zásadní vliv, je zřejmé, že rychlost značkování dokáže o několik desítek vteřin zvýšit či snížit. Je to dáno tím, že u slov, pro která máme značek dost, neprovádíme žádná další hledání (která by nejspíše nepřinesla významnou změnu v úspěšnosti značkování).

8.1.4 Eliminace značek

Stejně jako můžeme omezit počet přidělených značek, můžeme právě z těch přidělených vybrat jen značky s nejvyšší pravděpodobností. Jednou možností výběru je výpočet průměrné pravděpodobnosti a vyřazení těch značek, které mají pravděpodobnost nižší. Drouhou možností je pak výběr určitého počtu značek s nejvyšší pravděpodobností.

Na základě průměru

Kromě samotné hodnoty průměru je možné použít i nějaký jeho násobek. Jak se však ukázalo, není tato metoda zcela účinná, pokud jsme předem omezili počet přidělovaných značek. V takovém případě dochází i k vyřazení značek správných, neboť četnost, s jakou se vyskytují u daného zakončení (a tudíž výslednou pravděpodobnost) ovlivnit nedokážeme.

Mez pro vyřazení	**Úspěšnost** [%]	**Doba běhu** [s]
Aritmet. průměr	88,671	71,90
0,5 * průměr	89,080	83,90
0,1 * průměr	90,153	104,08
0,05 * průměr	90,467	107,38

Výběr značek

Při použití této metody jednoduše vybereme zadané množství značek s nejvyšší pravděpodobností. Nicméně se opět ukazuje, že takový výběr by měl větší význam v případě, kdybychom neomezili počet přidělených značek. Pokud je ale množství značek omezeno, nemá tak silný účinek.

Počet vybraných	**Úspěšnost** [%]	**Doba běhu** [s]
10	90,087	100,79
15	90,737	131,81
20	91,112	129,06
30	91,241	149,23

Obecně platí, že použití jakékoli vyřazovací metody má za následek prodloužení chodu programu, neboť s sebou přináší další série vyhledávání a průchodů seznamy značek. Oproti tomu omezení počtu přidělovaných značek znamená urychlení, protože k dosažené výsledku nám stačí prohledávání méně.

8.1.5 Celková úspěšnost

Co se celkové úspěšnosti přidělování značek týká, neznamená použití odhadu značek na první pohled závratný posun. Při porovnání běhu programu bez odhadu značek a s odhadem zjistíme, že celková účinnost se zvýšila při po-

užitých nastaveních - max. přidělených značek pro lemmatická zakončení: 8; max. přidělených značek pro nejčastější zakončení: 25; limity pro hledání značek 5 a 15; bez použití eliminace - přibližně o 2,75%. Na druhou stranu je třeba podotknout, že při testování 421416 slov by to mohlo znamenat 11589 slov, pro která jsme nově nalezli značku. Pro přehlednost uvádím výsledné hodnoty v tabulce:

Přidělování značek	**Úspěšnost** [%]	**Doba běhu** [s]
Bez použití odhadu	88,527	1336,66
S použitím odhadu	91,254	144,05 (+ 80)
Rozdíl	2,727	1112

Ačkoli samotná úspěšnost značkování nezaznamenala markantní posun, je třeba podotknout, že se velmi zrychlila doba běhu značení. V hodnotě z výše uvedené tabulky není započten čas potřebný pro tvorbu slovníku (přibližně 80 vteřin), jedná se pouze o dobu běhu samotného odhadu. Nicméně na to, abychom při odhadu značek dosáhli stejné hodnoty účinnosti, potřebujeme pouhých 70 vteřin. Celkově s inicializací programu tedy takový odhad zabere 150 vteřin, tedy dvě a půl minuty. Což je pouhých 11 % doby běhu původní verze.

Pokud porovnáme program s podobnými taggery vyvíjenými na jiných univerzitách a dalších zařízeních, zjistíme, že úspěšnost značkování je o několik procent nižší (doposud nejúspěšnější taggery se pohybují okolo 96%-98%). Takové programy jsou založeny buď na zcela jiných principech, nebo jsou velice úzce specializované, což se samozřejmě projeví i na jejich účinnosti. Ale rovněž musím poznamenat, že u zmíněných programů probíhá proces přidělování značek v řádech (až desítek) hodin. Zatímco produkt této práce umožňuje značkování v intervalu několika minut.

8.2 SNK

Rozsah testovacího souboru pro korpus SNK byl oproti PDT přibližně o třetinu menší, stejně tak i rozsah trénovacích dat (bližší popis je obsažen v kapitole 7.1.2). Tím, že je zpracováváno o několik set tisíc slov méně, je dána i kratší doba běhu programu. Stejně tak i rychlejší tvorba slovníků - inicializační fáze programu trvá 35-40 vteřin.

8.2.1 Velikost Hashtabulky

Díky tomu, že slovenština obsahuje některé specifické znaky, vychází pro mnoho slov i jiný výsledek hashovací funkce (dekadická hodnota až čtyř prvních znaků), což znamená odlišné rozložení slov. I díky rozdílnému rozsahu testovacího i zdrojového souboru zde dochází k tomu, že hodnoty 768 či 1024 se ukazují jako vhodnější pro rozsah hashtabulky.

Použitý rozsah	**Doba běhu** [s]
256	116,25
384	110,18
512	106,24
768	107,22
1024	114,69

Přičemž jako nejvhodnější jsem zvolil hodnotu 768, neboť u hodnoty 1024 už není časové zrychlení nijak výrazné. Protože nejspíše opět dochází k tomu, že hashtabulka není ideálně zaplněna.

8.2.2 Počet přidělovaných značek

Množství přidělovaných značek k jednotlivým koncovkám jsem omezil stejnými hodnotami jako u korpusu PDT.

z lemmat	**z nejčastějších**	**Úspěšnost** [%]	**Doba běhu** [s]
7	7	88,447	60,85
8	8	88,449	70,33
10	10	88,519	74,24
15	15	88,574	87,80

Stejně jako u korpusu PDT se jako nejvýhodnější ukázala asymetrická kombinace parametrů: maximálně 8 značek pro koncovky získané z lemmat a maximálně 25 značek pro koncovky nejčastější. I když některá přísnější omezení nabízejí výraznou úsporu času, přednost dostala varianta s nejvyšší naměřenou úspěšností značkování.

z lemmat	**z nejčastějších**	**Úspěšnost** [%]	**Doba běhu** [s]
8	10	88,529	80,90
8	15	88,613	98,29
8	20	88,667	102,97
8	25	88,691	107,22
15	25	88,607	108,79

8.2.3 Omezení hledání

Při hledání limit, které omezují případné hledání dalších značek, se ukazuje, že výhodnou kombinací jsou jak hodnoty 8 a 16, ale rovněž i 10 a 20.

Dolní mez	**Horní mez**	**Úspěšnost** [%]	**Doba běhu** [s]
5	15	88,657	118,22
7	14	88,685	113,07
8	16	88,691	107,22
10	20	88,690	120,44

8.2.4 Eliminace značek

Na základě průměru

Bohužel, i u korpusu SNK se ukazuje, že vyřazování značek na základě průměrné pravděpodobnosti není účinné. I zde totiž dochází k vyřazování značek, které jsou správné, na úkor značek nevhodných, ale četněji se vyskytujících.

Mez pro vyřazení	**Úspěšnost** [%]	**Doba běhu** [s]
Aritmet. průměr	85,907	46,88
0,5 * průměr	86,451	52,89
0,1 * průměr	87,477	67,73
0,05 * průměr	87,821	75,64

Výběr značek

Rovněž výběr několika značek s nejvyšší pravděpodobností se poněkud míjí účinkem, nicméně uvádím naměřené hodnoty alespoň pro porovnání s korpusem PDT.

Počet vybraných	**Úspěšnost** [%]	**Doba běhu** [s]
10	87,268	60,70
15	88,047	75,18
20	88,510	84,78
30	88,687	96,64

8.2.5 Celková úspěšnost

Oproti korpusu PDT je u SNK celková úspěšnost odhadu celkově nižší. Při zvolených parametrech: velikost hashtabulky 768; maximálně 8 značek pro lematická zakončení; maximálně 25 značek pro nejčastější zakončení; limity pro hledání značek 8 a 16; bez použití eliminace značek. Rozdíl mezi použitými verzemi programu pak činí přibližně 2,4%, co se úspěšnostii značkování týká.

Přidělování značek	**Úspěšnost** [%]	**Doba běhu** [s]
Bez použití odhadu	86,305	1329,30
S použitím odhadu	88,691	107,22 (+ 40)
Rozdíl	2,386	1182

Pokud budeme porovnávat dobu běhu taggeru ve verzi s odhadem značek a bez něj, dojdeme k rozdílu ještě většímu než u PDT. To je dáno především nižším rozsahem vstupních dat. Nic to ale nemění na tom, že rychlost značkování bez použití odhadu se téměř nezměnila. Zatímco při použití odhadu nastává urychlení jak při vytváření slovníků, tak při samotném přiřazování značek. Při použití odhadu bylo dosaženo 87,7% urychlení.

9 Závěr

Morfologické značky nalézají stále větší uplatnění při pokročilé práci s textem, jeho analýze a zpracování. Tato práce si klade za cíl seznámit čtenáře jak s principy morfologického značkování a postupy používanými při odhadu značek, tak i s úskalími, které značkování slov flektivních jazyků, mezi které patří i čeština, skrývá.

První část práce je zaměřena na popis systému morfologického značkování obecně a dále na morfologické jevy, kterých je při odhadu značek využíváno. Jsou v ní rovněž uvedeny i příklady jazykových zvláštností, na které je třeba dát pozor při odhadu značek, aby bylo dosaženo co nejvyšší úspěšnosti. V části druhé jsou popsány postupy a prostředky používané při odhadu značek. Poslední část pak popisuje nejdůležitější aspekty samotné implementace odhadu morfologických značek a efektivitu činnosti programu.

S prací jsem spokojen, jelikož splňuje požadavky, které na ni byly kladeny, a zvyšuje účinnost dosavadního systému morfologického značkování. U značkování češtiny rozdíl činí přibližně 2,7%, u slovenštiny je rozdíl při použití odhadu značek 2,4%. Neméně významným faktem je zrychlení procesu přidělování značek o více než 80%. Během tvorby této bakalářské práce jsem načerpal mnoho nových vědomostí, které budu moci využít i v budoucnu, což rovněž považuji za její přínos.

Literatura

[Brychcín(2010)] BRYCHCÍN, T. Efektivní dekódování s N-gramovými jazykovými modely. Master's thesis, Západočeská univerzita v Plzni, 2010.

[Dokulil(1986)] DOKULIL, M. *Mluvnice češtiny, 1. díl.* Praha : Academia, 1986. ISBN 21-099-86.

[Šmilauer(1973)] ŠMILAUER, V. *Nauka o českém jazyku.* Praha : SPN, 1973. ISBN 14-218-73.

[Hajič(2004)] HAJIČ, J. *Disambiguation of Rich Inflection (Computational Morphology of Czech). Vol. 1.* Praha : Karolinum Charles University Press, 2004.

[Hajič(2006)] HAJIČ, J. *The Prague Dependency Treebank* [online]. 2006. Dostupné z: `http://ufal.mff.cuni.cz/pdt2.0/`.

[Hajič(2010)] HAJIČ, J. *Popis morfologických značek - poziční systém* [online]. 2010. Dostupné z: `http://ucnk.ff.cuni.cz`.

[Havránek(1960)] HAVRÁNEK, B., JEDLIČKA, A. *Česká mluvnice.* Praha : SPN, 1960.

[Jelínek(2008)] JELíNEK, T. Nové značkování v Českém národním korpusu. *Naše řeč.* 2008, 91, s. 13–20.

[MacCartney(2005)] MACCARTNEY, B. *NLP Lunch Tutorial: Smoothing*, 2005.

[Rabiner(1989)] RABINER, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE.* 1989, 77, s. 257–286.

[SAV(2004)] JÚLŠ SAV. *Morfologická anotácia textov Slovenského národného korpusu*, 2004.

[SAV(2010)] JÚLŠ SAV. *Slovenský národný korpus* [online]. 2010. Dostupné z: `http://korpus.juls.savba.sk/stats.html`.