TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky a mezioborových inženýrských studií

DIZERTAČNÍ PRÁCE

2005

Jindra Drábková

TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových inženýrských studií

TVORBA JAZYKOVÉHO MODELU ZALOŽENÉHO NA TŘÍDÁCH

DIZERTAČNÍ PRÁCE

JINDRA DRÁBKOVÁ

Liberec 2005

Tvorba jazykového modelu založeného na třídách

Dizertační práce

Ing. Jindra Drábková

Studijní program:	P2612 Elektrotechnika a informatika
Studijní obor:	2612V045 Technická kybernetika
Pracoviště:	Katedra elektroniky a zpracování signálů Fakulta mechatroniky a mezioborových inženýrských studií Technická univerzita v Liberci Hálkova 6, 461 17 Liberec
Školitel:	Prof. Ing. Jan Nouza, CSc.

Rozsah práce a příloh

Počet stran: 123
Počet obrázků: 25
Počet tabulek: 21
Počet vzorců: 61
Počet příloh: 1

+ AUTOREFERÁT

Abstrakt

Rozpoznávání spojité řeči je komplexní problém sestávající z několika úloh. Jednou z těchto úloh je tvorba jazykového modelu. Český jazyk patří mezi jazyky ohebné, což s sebou nese řadu nevýhod. Jednou z nich je velké množství slov, které tvorbu jazykového modelu komplikuje. Dizertační práce předkládá řešení, ve kterém se v jazykovém modelu použijí gramatické značky místo slov. Ke stanovení značek byly využity tři přístupy – statistický, gramatický a pravděpodobnostní.

Téma dizertační práce zasahuje do několika oblastí od počítačové lingvistiky, morfologické analýzy po tvorbu korpusu a slovníku. Všechny vyjmenované oblasti by mohly být zahrnuty pod společný název počítačové zpracování jazyka. Tato disciplína je základem pro řadu dalších odvětví, jako je např. strojový překlad, větný rozbor nebo rozpoznávání řeči.

K tvorbě bigramového jazykového modelu značek byl vytvořen vlastní označkovaný korpus. Ten byl označkován částečně ručně a z větší části automaticky. Pro automatické značkování byl navržen stochastický značkovač, který využívá označkovaný slovník obsahující přibližně 300 tisíc různých slovních tvarů. Značky byly do slovníku přidány jednak ručně a jednak na základě syntaktické metody.

Z označkovaných dat bylo vytvořeno několik bigramových jazykových modelů značek vytvořených z vět s interpunkcí i bez interpunkce. Všechny modely značek byly testovány v závislosti na velikosti slovníku a výsledky testování byly zhodnoceny. Nejlepší jazykový model značek byl použit pro experimenty se systémem pro rozpoznávání spojité řeči.

Abstract

Speech recognition is a complex challenge which consists of several tasks. Language modeling is one of these tasks. The Czech language belongs to a group of languages which can be termed as flexible languages. One of the greatest disadvantages of such flexible languages is the large number of words. This PhD thesis submits a solution to this disadvantage. It is to use the grammatical tags instead of the words. To determine these tags three different approaches were used – statistical, grammatical and stochastic.

PhD thesis theme includes several branches – computational linguistics, morphologic analyses, corpora building and vocabulary building. All of these branches could be called natural language processing. This creates a disciplinary foundation, for example, for machine translation, parsing or speech recognition.

A bigram class-based language model was built using tagged corpus. The tagging of the corpus was completed in part manually, and in part automatically. A stochastic tagger was devised to automatic tagging using tagged vocabulary which includes some 300,000 items. Tags were added to this vocabulary both manually and with using syntactic method.

Using tagged corpus it was possible to design a number of bigram class-based language models: unsmoothed, smoothed by linear interpolation, made from sentences either with or without punctuation. The effects of both punctuation and the size of vocabulary were summarized. The best bigram class-based language model was then used in experiments with the continuous speech recognizer.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek a značek

a	ohodnocení grafu
a_{ij}	pravděpodobnost přechodu
A	množina pravděpodobností přechodů, ohodnocení uzlu
Acc	Accuracy – přesnost rozpoznávání
Acc_{best}	nejlepší úspěšnost rozpoznávání
ASR	Automatic Speech Recognition
b_j	pravděpodobnostní rozložení stavu j
B	množina pravděpodobnostních rozložení každého stavu
BASE	British Academic Spoken English Corpus
BigCSLM+	bigramový jazykový model značek z 3,1 GB dat vytvořený z vět s intepunkcí
BigCSLM–	bigramový jazykový model značek z 3,1 GB dat vytvořený z vět bez intepunkce
Big8300SLM+	bigramový jazykový model značek z 8300 vět vytvořený z vět s interpunkcí
BMK	Brněnský mluvený korpus
BNC	British National Corpus
c	třída, značka
CLAWS	Constituent Likelyhood Automatic Word-Tagging System
$Corr$	Correctness – správnost rozpoznávání
$C(w)$	počet výskytů slova w
$C(w_{n-1}, w_n)$	počet výskytů dvojic slov w_{n-1}, w_n
$C_{BH}(w)$	počet slov v odložených datech
C5	základní množina značek použitá pro značkování BNC
C7	rozšířená množina značek použitá pro značkování BNC
ČNK	Český národní korpus
D	počet vynechaných slov
DESAM	označkovaný korpus publicistických textů
DTW	Dynamic Time Warping

e	hrana
E	konečná množina hran
EM	Expectation-Maximization
ENGCG	English Constraint Grammar
ENGTWOL	anglický morfologický analyzátor
FI MU	Fakulta informatiky Masarykovy univerzity v Brně
G	mapovací funkce, perplexita, orientovaný ohodnocený graf
H	počet slov v odložených datech, entropie
HMM	Hidden Markov Models
i	index, který určuje pořadí slova ve větě; rovnost $P(W) = P(W)_{max}$
I	vzájemná informace, počet vložených slov
j	stav, značka příslušná danému slovu, rovnost $Acc = Acc_{best}$
k	mixtura, značka příslušná předchozímu slovu, váhový koeficient
L	ztráta informace
LM	Language Model
M	počet symbolů výstupní abecedy, počet mixtur
MFCC	Mel-Frequency Cepstrum Coefficients
ML	Maximum Likelyhood
MMI	Maximum Mutual Information
N	počet slov, počet stavů, počet vět
O	akustická informace
OOV	out of vocabulary
$p(O)$	pravděpodobnost akustické informace O
$p(O \mid W)$	pravděpodobnost, že akustická informace O odpovídá vyslovené posloupnosti W
$p(W)$	pravděpodobnost vyslovené posloupnosti W
$p(W \mid O)$	pravděpodobnost, že vyslovená posloupnost W odpovídá akustické informaci O
$p_{+1}(w)$	vyhlazená pravděpodobnost metodou Add-One Smoothing
$p_{WB}(w)$	vyhlazená pravděpodobnost metodou Witten-Bell Discounting
$\hat{p}(w)$	vyhlazená pravděpodobnost metodou Backoff Smoothing, vyhlazená pravděpodobnost metodou Linear Interpolation Smoothing
P	přirozený logaritmus pravděpodobnosti
PDT	Prague Dependency Treebank (Pražský závislostní korpus)
PMK	Pražský mluvený korpus
POS	Part-of-Speech

R	korelační faktor
S	počet nahrazených slov
SGML	Standard Generalized Markup Language
SLM	stochastický jazykový model
SYN2000	synchronní korpus z roku 2000
TnT	Trigrams'n'Tags
$T(w_{n-1})$	počet rozdílných dvojic sousedních slov, jejichž první slovo je w_{n-1}
ÚČNK	Ústav Českého národního korpusu
ÚFAL MFF UK	Ústav formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy v Praze
V2637_T	soubor 2637 vět upravených tak, že místo všech slov jsou značky
V2637_1	soubor 2637 vět upravených tak, že místo podstatných jmen jsou značky
V2637_12	soubor 2637 vět upravených tak, že místo podstatných a přídavných jmen jsou značky
V2637_6	soubor 2637 vět upravených tak, že k příslovcím je přidána informace o slovním druhu
V7000	základní soubor 7000 vět
V	rozměr slovníku, konečná množina uzlů
W	posloupnost slov
$\hat{W}$	nejpravděpodobnější posloupnost slov
$w(n)$	Hammingovo okénko
$Z(w_{n-1})$	počet dvojic sousedních slov, které se neobjevily v trénovacích datech a jejichž první slovo je w_{n-1}
α_1	váhový koeficient
α_2	váhový koeficient
c	tolerance, zobrazení
λ	soubor parametrů Markovova modelu
λ_i	lineární koeficient
$\boldsymbol{\mu}$	vektor středních hodnot
$\boldsymbol{\pi}$	vektor pravděpodobností v počátečním stavu
$\boldsymbol{\Sigma}$	kovarianční matice
$\Sigma P(W)$	součet přirozených logaritmů pravděpodobností

Kapitola 1

Úvod

Cílem dizertační práce je vytvoření jazykového modelu založeného na třídách pro rozpoznávač spojité řeči a jeho praktické využití.

V současné době jsou nejrozšířenější jazykové modely založené na statistických informacích získaných z korpusu. Na základě takových jazykových modelů je možno s určitou pravděpodobností předpovídat následující slovo, čehož se využívá nejen v rozpoznávání řeči, ale i v rozpoznávání rukou psaného textu, v detekci pravopisných chyb apod.

Práce je rozdělena na dvě části, teoretickou a praktickou. Teoretická část popisuje úlohu rozpoznávání řeči, tvorbu akustického a jazykového modelu a vysvětluje pojem značkování. Praktická část se zabývá stanovením gramatických značek pro český jazyk, tvorbou označkovaného korpusu a slovníku. Jsou zde uvedeny návrhy stochastických značkovačů, které byly použity pro automatické označkování velkého množství dat, a prezentovány výsledky automatického značkování s bigramovým jazykovým modelem založeným na třídách.

V druhé kapitole je popsán statistický přístup k rozpoznávání řeči. Tento přístup rozděluje úlohu rozpoznání řeči na několik dílčích úloh, z nichž nejdůležitější jsou tvorba akustického modelu a tvorba jazykového modelu. V kapitole jsou vysvětleny principy tvorby obou modelů včetně popisu vyhlazování jazykového modelu. Jsou zde také uvedeny základní vztahy týkající se jazykového modelu založeného na třídách.

Pro tvorbu jazykového modelu, který je založen na třídách, a pro jeho trénování a testování je třeba velkého množství textových dat, která jsou vhodně uspořádána a označkována. K tomuto účelu slouží označkovaný korpus a slovník. Ve třetí kapitole jsou pojmy korpus a slovník vysvětleny. Součástí této kapitoly jsou i příklady nejznámějších

českých a anglických korpusů a je zde popsán postup, jak byl vytvořen vlastní korpus a označkovaný slovník.

Značkování včetně stručného popisu nejznámějších značkovačů je shrnuto ve čtvrté kapitole. Část této kapitoly je věnována i popisu nejznámějších českých množin značek.

V páté kapitole je popsána metoda, jakou byly stanoveny gramatické značky pro jazykový model založený na třídách. Seznam gramatických značek, které byly pro označkování použity, je uveden v příloze.

Různé způsoby automatického značkování vět jsou uvedeny v šesté kapitole. Součástí je grafický popis postupu značkování.

V sedmé kapitole jsou představeny navržené systémy pro automatické značkování (taggery, značkovače) a jejich použití při značkování korpusu. Součástí kapitoly jsou i výsledky porovnání jednotlivých značkovačů na testovacích datech.

Osmá kapitola uvádí konkrétní příklady různě automaticky označkovaných vět a porovnává úspěšnost jejich značkování.

Výsledky automatického značkování, které bylo provedeno s vyhlazeným i nevyhlazeným bigramovým jazykovým modelem, jsou uvedeny v deváté kapitole. Automatické značkování bylo prováděno s jazykovými modely vytvořenými z vět s interpunkcí a bez interpunkce a s různě velkým slovníkem. Vliv interpunkce a velikosti slovníku na úspěšnost automatického značkování je zhodnocen také v deváté kapitole. Součástí této kapitoly jsou experimenty se systémem pro rozpoznávání spojité řeči.

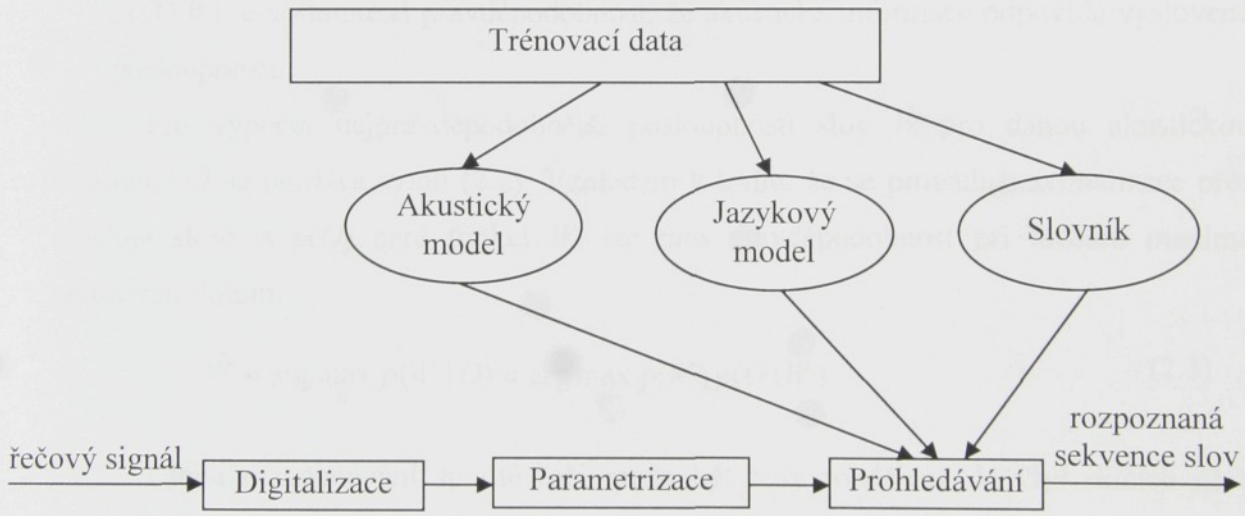
Cíle práce lze shrnout do těchto bodů:

1. Stanovení gramatických značek pro český jazyk.
2. Vytvoření vlastního označkovaného slovníku.
3. Návrh a realizace různých automatických značkovačů.
4. Vyhodnocení vytvořených značkovačů na testovacích datech.
5. Automatické značkování velkého množství dat pomocí nejlepšího značkovače.
6. Tvorba různých bigramových jazykových modelů založených na třídách.
7. Testování jazykových modelů v závislosti na interpunkci a velikosti slovníku.
8. Využití bigramového jazykového modelu založeného na třídách v systému pro rozpoznávání spojité řeči.

Kapitola 2

Statistický přístup k rozpoznávání souvislé řeči

Rozpoznání řeči je proces, při kterém akustický signál snímaný např. mikrofonem generuje posloupnost slov. Na obr. 2.1 jsou schematicky zobrazeny hlavní části systému pro rozpoznávání spojité řeči.



Obr. 2.1: Systém pro rozpoznávání spojité řeči

Nejprve je třeba provést digitalizaci řečového signálu, digitalizovaný signál zpracovat metodami krátkodobé analýzy a poté vyhledat nejpravděpodobnější kandidáty slov na základě akustického a jazykového modelu. Oba modely jsou vytvořeny na základě trénovacích dat.

Při řešení úlohy rozpoznávání spojité řeči se v současné době nejčastěji využívá statistický přístup. Předpokládejme, že $W = \{w_1, w_2, w_3, ..., w_N\}$ je posloupnost N slov

a $O = \{o_1, o_2, o_3, \ldots, o_M\}$ je akustická informace odvozená z řečového signálu. Cílem je nalézt nejpravděpodobnější posloupnost slov $\hat{W}$ pro danou akustickou informaci O:

$$\hat{W} = \arg\max_{W} p(W \mid O) \tag{2.1}$$

kde $p(W \mid O)$ je podmíněná pravděpodobnost, že vyslovená posloupnosti W odpovídá akustické informaci O,

funkce arg max v tomto vztahu znamená nalezení posloupnosti W takové, pro kterou je $p(W \mid O)$ maximální.

V případě, že použijeme Bayesovo pravidlo, platí:

$$\hat{W} = \arg\max_{W} \frac{p(W)p(O \mid W)}{p(O)} \tag{2.2}$$

kde $p(W)$ je pravděpodobnost vyslovené posloupnosti W,

$p(O)$ je pravděpodobnost akustické informace O,

$p(O \mid W)$ je podmíněná pravděpodobnost, že akustická informace odpovídá vyslovené posloupnosti.

Pro výpočet nejpravděpodobnější posloupnosti slov $\hat{W}$ pro danou akustickou informaci O se používá vztah (2.2). Vzhledem k tomu, že se provádí maximalizace přes všechna slova a $p(O)$ není funkcí W, lze tuto pravděpodobnost při hledání maxima ignorovat. Potom:

$$\hat{W} = \arg\max_{W} p(W \mid O) = \arg\max_{W} p(W)p(O \mid W) \tag{2.3}$$

Úloha rozpoznávání spojité řeči může být tedy rozdělena do čtyř dílčích úloh [PSUTKA 1995]:

1. akustické zpracování řečového signálu,
2. vytvoření akustického modelu $p(O \mid W)$,
3. vytvoření jazykového modelu $p(W)$,
4. nalezení nejpravděpodobnější posloupnosti slov.

2.1 Akustické zpracování řečového signálu

2.1.1 Digitalizace

Digitalizace signálu zahrnuje vzorkování a kvantizaci.

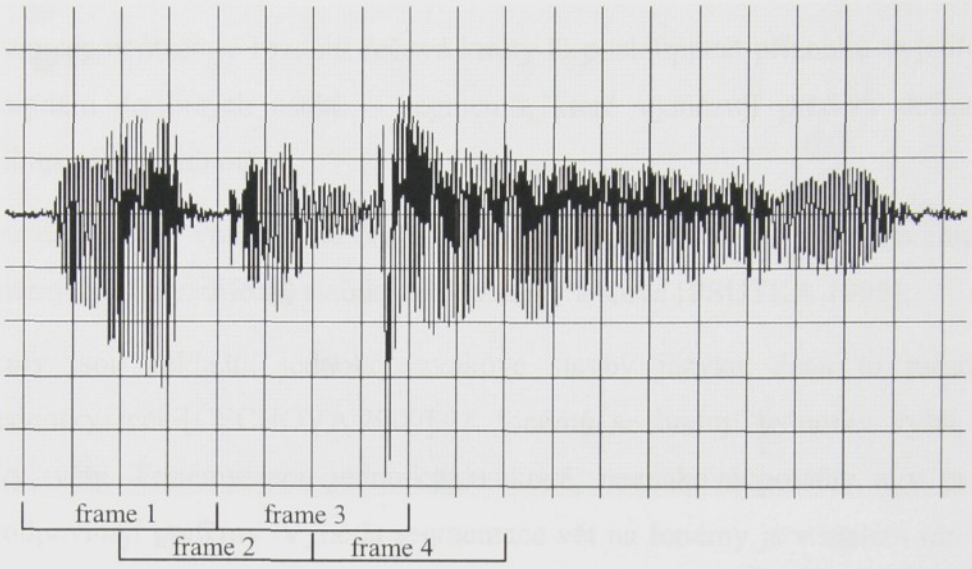
Vzorkováním se převádí signál spojitý v čase na posloupnost vzorků diskrétních v čase. Při vzorkování musí být splněn Shannonův vzorkovací teorém, kdy vzorkovací frekvence musí být nejméně dvakrát větší než maximální frekvence analogového signálu. Jestliže tato hodnota není jasně definována, signál se před vzorkováním filtruje dolní propustí. Pro telefonní pásmo (0–3,4 kHz) se používá frekvence vzorkování 8 kHz, pro kvalitní zpracování signálu mluvené řeči (0–8 kHz) se volí frekvence vzorkování vyšší, většinou 16 kHz.

Kvantizací aproximujeme analogové hodnoty vzorku signálu jednou z konečného počtu číselných hodnot. Udává se počet úrovní kvantování (obvykle se volí ve tvaru 2^B, kde B je počet bitů v binárním kódu) a kvantizační krok. Tyto parametry se vybírají tak, aby byl pokryt celý rozsah signálu. Pro kvalitní zpracování signálu se obvykle používá šestnáctibitový převod.

2.1.2 Zpracování v časové a frekvenční oblasti

Při analýze řečového signálu se využívá vlastností tzv. krátkodobých modelů produkce řeči, které jsou založeny na předpokladu, že vlastnosti řečového signálu se v průběhu času mění pomalu. Předpokládá se, že signál je stacionární na malém časovém úseku. Z tohoto důvodu se používají tzv. metody krátkodobé analýzy, při nichž se úseky řečového signálu vydělují a zpracovávají tak, jako by to byly oddělené krátké zvuky. To znamená, že se při této analýze rozděluje akustický signál na framy (mikrosegmenty). Délka framu se obvykle volí 10–30 ms. Při vzorkovací frekvenci 8 kHz tak každý frame obsahuje 80–240 vzorků. Aby byla zachována určitá spojitost po sobě jdoucích framů, je vhodné, aby se částečně překrývaly. Obvykle se volí poloviční překryv, takže při délce framu 20 ms začíná nový frame každých 10 ms (viz obr. 2.2).

Před analýzou signálu je vhodné použít preemfázi ve tvaru číslicového filtru např. ve tvaru $y(n) = x(n) - a.x(n-1)$, kde $y(n)$ je n-tý vzorek signálu po preemfázi a $x(n)$ je vzorek původního signálu. Konstanta a se volí v rozsahu 0,95–0,98. Preemfáze zdůrazňuje vyšší frekvence, a tím kompenzuje jejich útlum při průchodu akustické vlny rty.



Obr. 2.2: Překryv framů

Pro potlačení váhy vzorků na krajích framu se používá Hammingovo okénko, které je definováno vztahem:

$$w(n) = \begin{cases} 0{,}54 - 0{,}46\cos[2\pi n/(N-1)] & \text{pro } 0 \le n \le N-1 \\ 0 & \text{pro ostatní } n \end{cases} \qquad (2.4)$$

2.1.3 Parametrizace

Výsledkem parametrizace jsou příznakové vektory. Metody krátkodobé analýzy většinou předpokládají, že se jako vstup zpracovávají data získaná digitalizací. Vektor příznaků popisující daný frame je získán aplikací vhodné metody akustické analýzy původního akustického řečového signálu. Protože framy jdou v čase za sebou, vznikají časové posloupnosti vektorů příznaků, které charakterizují vyslovený celek. Dimenze vektorů je dána počtem příznaků a délka posloupnosti je úměrná délce promluvy a je určena počtem framů.

Příznaky mohou být statické nebo dynamické a existuje mnoho metod analýzy akustického signálu. K nejpoužívanějším patří analýza v časové a frekvenční oblasti, kepstrální analýza, melovská filtrace, homomorfní analýza a lineární predikce. Detailní popis jednotlivých metod je uveden např. v [PSUTKA 1995]. Volba vhodných příznaků je velmi důležitým bodem procesu rozpoznávání řeči.

2.1.4 Segmentace

Úkolem segmentace je rozdělit řečové kmity či posloupnost příznaků vyjádřených z řečového signálu do jistých úseků – segmentů, které vymezují předem definované fonetické či lingvistické jednotky.

Velice důležitá je volba jednotky segmentace. Jako jednotky pro segmentaci se používají fonémy, alofóny, difóny, slabiky, poloslabiky a slova [PSUTKA 1995].

Fonémy jsou základní jednotky zvukové stavby jazyka. Jsou to minimální rozlišující jednotky řeči [ČECHOVÁ 2000]. Z fonémů se budují jednotky vyšší např. slova, slabiky, věty. Fonémy jsou jednotky zvukové, neznakové povahy a v psaném projevu jim odpovídají grafémy. Výhoda segmentace vět na fonémy je v malém množství fonémů (ve většině jazyků 30–50). Fonémy však neobsahují informaci o koartikulaci, což je závislost výslovnosti fonému na předcházejícím a následujícím zvuku a na tempu a intonaci řeči.

Alofóny jsou poziční varianty fonémů, které jsou navíc určeny svým pravým a levým kontextem. Výhodou je, že v příslušném alofónu je obsažen koartikulační efekt. Nevýhodou využití alofónů je jejich vysoký počet (až několik set).

Difón je fonetický útvar, který je užíván pro označení posloupnosti od středu samohlásky do středu souhlásky a naopak. Difóny jsou používány z důvodu, že velká část akustické informace leží v přechodech mezi souhláskami a samohláskami. Výhodou je, že obsahují důležitou koartikulační informaci. Nevýhodou je jejich velký počet (několik tisíc).

Slabiky jsou fonetické útvary, které obsahují samohláskové jádro plus volitelné počáteční a koncové souhlásky nebo skupiny souhlásek. Výhodou je, že slabika uvnitř sebe zachycuje jak koartikulaci, tak i další fonologické jevy. Nevýhodou je jejich velký počet (v mluvené češtině více než 10 000). Neexistuje také ustálený názor, kde mají být umístěny hranice slabiky.

Poloslabiky vznikají rozdělením slabik. Většina slabik obsahuje v jádru samohlásku, okolo které je shluk souhlásek. Rozdělením samohlásky ve slabice vzniknou dvě poloslabiky. Výhodou je, že v porovnání s celými slabikami je velikost inventáře menší (asi 2 000). Nevýhodou je složitost s rozdělováním samohláskového jádra.

Slova jsou spojení konečného počtu fonémů či slabik. Výhodou je, že v případě, že jsou slova vyslovována izolovaně, nemusíme používat složité algoritmy pro segmentaci

a identifikaci nižších jednotek (alofónů, fonémů apod.). Problém vymezení hranic slov u plynulé promluvy je velmi obtížný. Dochází totiž ke koartikulačním jevům mezi krajními fonémy sousedních slov, což vymezení hranic slov komplikuje.

2.1.5 Akusticko-fonetické dekódování promluvy

Úkolem akusticko-fonetického dekódování je nalézt hranice segmentů a provést jejich klasifikaci a fonetickou identifikaci. Metody akusticko-fonetického dekódovaní lze rozdělit na přístupy založené na heuristických pravidlech, přístupy založené na vzdálenosti a přístupy pravděpodobnostní [PSUTKA 1995].

Heuristický přístup je založen na znalostech, které jsou nejčastěji vyjádřeny ve tvaru produkčních pravidel. Výhodou tohoto přístupu je jeho flexibilnost (je možné pravidla přidávat a rušit). Konstrukce pravidel je však závislá na použitém jazyku, aplikované akustické analýze, zvolené fonetické jednotce a znalostech experta. Vzhledem k těmto nevýhodám (závislostem) nebyl dosud předložen žádný univerzálně použitelný algoritmus, který by problémy segmentace a identifikace segmentů vyřešil.

Přístup založený **na vzdálenosti** využívá porovnání promluvy rozdělené na segmenty se souborem referenčních segmentů, u kterých je známo správné přiřazení. Míra podobnosti dvou segmentů řeči se vyjadřuje nejčastěji mírou vzdálenosti či mírou zkreslení. Při procesu porovnávání se využívá dynamického programování.

Pravděpodobnostní přístup je většinou založen na využití tzv. skrytých Markovových modelů. Tyto modely jsou vytvářeny pro akustickou realizaci každé analyzované fonetické jednotky (slova, fonému). K natrénování modelů je potřeba velké množství promluv. Pomocí natrénovaných modelů lze vypočítat pravděpodobnost toho, že pozorované příznaky byly produkovány některým z modelů, kterými je modelována fonetická jednotka. Užitím předpokladu o maximální pravděpodobnosti pak může být vybrán takový model, který maximalizuje výslednou pravděpodobnost. Metoda modelování promluvy skrytými Markovovými modely je automatická, nevyžaduje formulování produkčních pravidel a existují dobré předpoklady pro jejich zobecnění v různých jazycích. Hlavní nevýhodou je obtížné včleňování dalších řečových znalostí (fonetických, fonologických, lexikálních, syntaktických apod.).

2.1.6 Fonetická transkripce češtiny

Fonetická transkripce se používá k přesnému a jednoznačnému přepisu textu nebo zvuků na odpovídající posloupnost fonémů. Důvodem zavedení fonetické transkripce je její využití v rozpoznávání řeči, kdy jsou slova modelována zřetězením skrytých Markovových modelů fonémů. Další možnost využití fonetické transkripce je v hlasové syntéze z psaného textu. Pro zápis fonémů je třeba použít nějakou fonetickou abecedu. Pro přepis českého textu se nejčastěji používá abeceda PAC (Phonetic Alphabet for Czech), která obsahuje 41 jednoznakových symbolů reprezentujících 40 hlásek a ticho a několik pomocných značek. Často bývá tato abeceda upravována pro konkrétní potřeby. Pro přepis textu je potřeba stanovit určitá pravidla, podle nichž lze fonetický přepis automaticky vytvářet. Většina fonémových změn může být vysvětlena levým a pravým kontextem daného fonému, takže obecné pravidlo lze definovat ve tvaru:

JESTLIŽE řetězci znaků A bezprostředně předchází řetězec znaků C a je bezprostředně následován řetězcem znaků D,
PAK se řetězec znaků A přepíše na řetězec znaků B.
Tato pravidla jsou stanovena na základě znalostí expertů (fonetiků a fonologů) a umožňují přepis obecného textu na řetězec fonémů.

2.2 Tvorba akustického modelu

Akustický model popisuje vzájemný vztah mezi řečníkem a akustickým procesorem, který transformuje řečové kmity produkované řečníkem na posloupnost vektorů příznaků. Úkolem akustického modelu je poskytnout co nejpřesnější odhad podmíněné pravděpodobnosti $p(O \mid W)$ pro libovolnou posloupnost vektorů příznaků akustické informace O a každou uvažovanou posloupnost slov W.

Existují různé metody pro akusticko-fonetické modelování. Pro tvorbu akustického modelu se využívají např. neutronové sítě [HUANG 2001] nebo expertní systémy [PSUTKA 1995]. V současné době používá většina ASR (Automatic Speech Recognition) systémů statistické modely založené na skrytých Markovových modelech (Hidden Markov Models – HMM).

Starší, zato jednodušší akusticko-fonetický model používá řečové šablony (referenční vzorky), jejichž příznakové vektory jsou uloženy v paměti počítače. Při klasifikaci se používá technika DTW (Dynamic Time Warping – dynamické borcení času),

která je založena na měření vzdáleností mezi vstupním neznámým slovem a referenčními vzorky ze slovníku. Nevýhodou této metody je závislost na konkrétním mluvčím a omezení na daný slovník reprezentovaný šablonami.

2.2.1 Skrytý Markovův model

Skrytý Markovův model (HMM) je konečná množina stavů, kterým je přiřazeno pravděpodobnostní rozložení. Přechody mezi stavy jsou ohodnoceny pravděpodobnostmi přechodu. V jednotlivých stavech může být generován výstup podle přiřazeného pravděpodobnostního rozložení. Stavy generují výstupy, ale samy zůstávají neviditelné – odtud termín skrytý Markovův model. Skrytý Markovův model s diskrétními výstupy lze definovat takto [WARAKAGODA 1996]:

- N je počet stavů,
- M je počet symbolů výstupní abecedy,
- A je množina pravděpodobností přechodů $A = \{a_{ij}\}$,

$$a_{ij} = p\{q_{t+1} = j \mid q_t = i\}, \qquad 1 \le i, j \le N \tag{2.5}$$

kde q_t je současný stav.

Platí:

$$\sum_{j=1}^{N} a_{ij} = 1, \qquad 1 \le i \le N \tag{2.6}$$

- B je množina pravděpodobnostních rozložení každého stavu $B = \{b_j(k)\}$,

$$b_j(k) = p\{o_t = v_k \mid q_t = j\}, \qquad 1 \le j \le N,\ 1 \le k \le M \tag{2.7}$$

kde v_k označuje k-tý pozorovaný symbol výstupní abecedy a o_t je stávající vektor parametrů.

Platí:

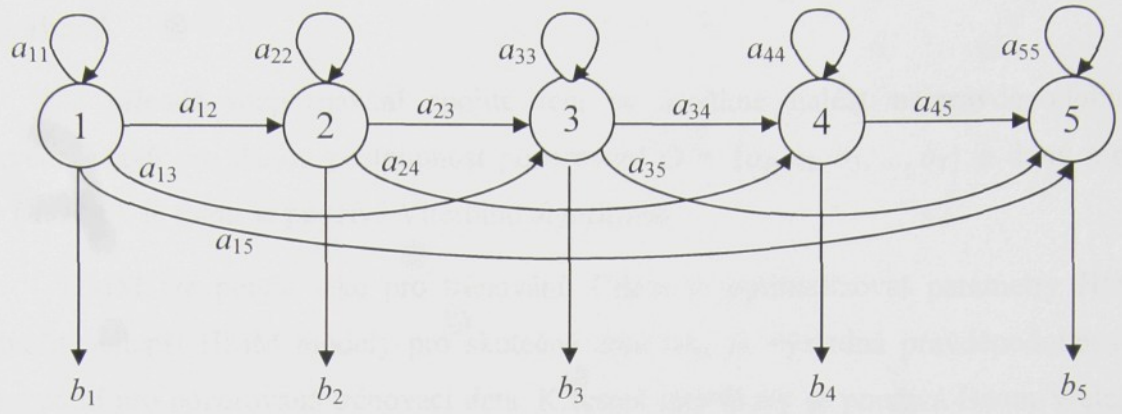
$$\sum_{k=1}^{M} b_j(k) = 1, \qquad 1 \le j \le N \tag{2.8}$$

- vektor pravděpodobností v počátečním stavu $\boldsymbol{\pi} = \{\pi_i\}$,

kde

$$\pi_i = p\{q_1 = i\}, \qquad 1 \le i \le N \tag{2.9}$$

Soubor parametrů Markovova modelu je trojice $\lambda = (A, B, \boldsymbol{\pi})$. Na obr. 2.3 je grafické znázornění Markovova modelu pro $N = M = 5$.



Obr. 2.3: Grafické znázornění skrytého Markovova modelu

V posledních 10–15 letech se při rozpoznávání řeči používají spojité HMM. Rozdíl mezi diskrétními a spojitými HMM je ve formě pravděpodobnostní výstupní funkce. U spojitých HMM se nejčastěji volí pravděpodobnostní výstupní funkce s normálním rozdělením určeným vektorem středních hodnot $\boldsymbol{\mu}_j$ a kovarianční maticí $\boldsymbol{\Sigma}_j$ pro stav j. V případě, že použijeme vícemodální gaussovské rozložení, kde se jednotlivé složky označují jako mixtury, je normální rozdělení charakterizováno vektorem středních hodnot $\boldsymbol{\mu}_{jk}$, kovarianční maticí $\boldsymbol{\Sigma}_{jk}$ pro stav j a váhovým koeficientem c_{jk} pro mixturu k.

$$b_j(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} N(\mathbf{x}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \tag{2.10}$$

kde M je počet mixtur.

U skrytých Markovových modelů předpokládáme splnění následujících podmínek.

- Každý stav je závislý pouze na stavu předchozím.
- Pravděpodobnosti přechodu jsou nezávislé na čase, ve kterém se přechod uskutečňuje.
- Stávající pozorování (výstup) je nezávislé na předchozích pozorováních (výstupech).

2.2.2 Úlohy využívající HMM

Při rozpoznávání řeči jsou skryté Markovovy modely využívány pro řešení řady úloh.

Jednou z takových úloh je rozpoznávání izolovaných slov, kde je zadán model $\lambda = (A, B, \pi)$ a akustická informace $O = \{o_1, o_2, o_3, ..., o_T\}$. Cílem této úlohy je nalezení

podmíněné pravděpodobnosti $p\{O \mid \lambda\}$, že akustická informace O je generována daným modelem λ. K výpočtu této pravděpodobnosti se používá dopředný algoritmus (Forward Algorithm).

V případě rozpoznávání spojité řeči se snažíme nalézt nejpravděpodobnější sekvenci stavů pro danou posloupnost pozorování $O = \{o_1, o_2, o_3, ..., o_T\}$ a daný model $\lambda = (A, B, \pi)$. K tomu se používá Viterbiho algoritmus.

HMM lze použít také pro trénování. Cílem je optimalizovat parametry HMM, vytvořit nejlepší HMM modely pro skutečná data tak, že výsledná pravděpodobnost je maximální pro pozorovaná trénovací data. K řešení této úlohy se používá Baum-Welchův algoritmus (Forward-Backward Algorithm).

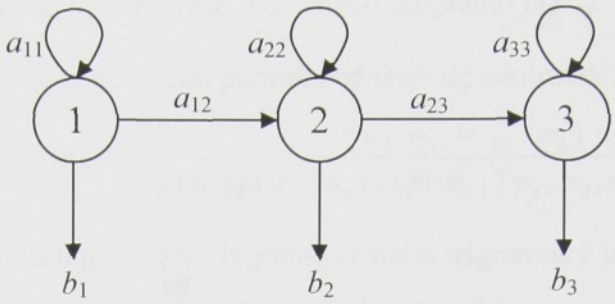
Podrobný popis všech tří metod je uveden např. v [RABINER 1989].

2.2.3 Modelování spojité řeči

Pro modelování mluvené řeči se využívají zejména levo-pravé Markovovy modely, které jsou vhodné pro modelování procesů, jejichž vývoj je spojen s postupujícím časem. Základní vlastností těchto modelů je, že proces začíná v počátečním stavu a se vzrůstajícím časem dochází k přechodům do stavů s vyššími indexy nebo setrvání ve stejném stavu. Průchod modelem je tedy zleva doprava. Proces končí v koncovém stavu.

Modely založené na skrytých Markovových modelech fonémů se v současné době používají pro modelování posloupnosti slov. Výhodou zavedení modelů fonémů je větší pružnost systému, to znamená, že je možné měnit a rozšiřovat slovník bez trénování nových modelů. Další výhodou fonémů je jejich malý počet. Nevýhodou je, že na rozdíl od grafémů (psaná podoba fonémů), které mají vždy stejnou grafickou podobu, zvuková realizace fonémů závisí na okolních hláskách. Z tohoto důvodu se pro modelování fonémů používá vícestavový (nejčastěji třístavový) skrytý Markovův model (obr. 2.4).

Markovovy modely slov vznikají jednoduše zřetězováním modelů odpovídajících fonémů. Modely slov se vytváří tak, že se nejprve aplikací fonetické transkripce získá základní fonetický tvar daného slova a pak se do pomyslného fonetického grafu odpovídajícího tomuto základnímu fonetickému tvaru „dosadí" místo jednotlivých fonémů jim odpovídající Markovovy modely [NOUZA 2001].



Obr. 2.4: Třístavový Markovův model fonému

2.3 Tvorba jazykového modelu

Úkolem jazykového modelu je stanovit jistá omezení a nalézt určitá pravidla, pomocí nichž můžeme ze slov vytvořit větu. Omezení a pravidla vycházejí z vlastností konkrétního jazyka a mohou být modelována jak stochastickými tak i nestochastickými metodami. Stochastické jazykové modely (SLM) používají pro jazykové modelování pravděpodobnostní přístup. Tyto jazykové modely přiřazují každé posloupnosti slov $W = \{w_1, w_2, w_3, ..., w_n\}$ pravděpodobnost $p(W)$, kterou je třeba odhadnout z dat. Data, která se používají pro tvorbu modelu, se nazývají trénovací data. Nejrozšířenější SLM je n-gramový jazykový model. Podle „řetězového pravidla" pravděpodobnosti platí:

$$p(W) = p(w_1, w_2, w_3, ..., w_n) = p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2)...p(w_n \mid w_1, w_2, ..., w_{n-1}) =$$
$$= \prod_{i=1}^{n} p(w_i \mid w_1, w_2, ..., w_{i-1}) \qquad (2.11)$$

Obecně lze pro odhad pravděpodobnosti výskytu slova použít n-gramový model slov, tj. $p(w_n \mid w_1, w_2, w_3, ..., w_{n-1})$. V praxi je nemožné tuto pravděpodobnost vypočítat, protože pro slovník o rozměru V a pro n-té slovo ve větě existuje V^{n-1} různých možností historií, což znamená, že ještě před samotným výpočtem posloupnosti je třeba zjistit celkem V^n různých pravděpodobností. Prvky n-gramového modelu jsou podmíněné pravděpodobnosti, které se rovnají pravděpodobnosti toho, že bude následovat jisté slovo w_n v případě, že nastala vstupní kombinace $w_1, w_2, w_3, ..., w_{n-1}$. V praxi by to znamenalo generovat obrovské množství dat, a proto se tento problém řeší aproximací. U n-gramového jazykového modelu předpokládáme, že je pravděpodobnost slova daná všemi předchozími slovy $p(w_n|w_1, w_2, w_3, ..., w_{n-1})$. Jestliže slovo závisí na předchozích

dvou slovech, mluvíme o trigramu $p(w_n|w_{n-2}, w_{n-1})$, podobně o bigramu, kdy dané slovo závisí pouze na předchozím slově $p(w_n|w_{n-1})$ nebo unigramu $p(w_n)$.

Prvky n-gramové matice jsou podmíněné pravděpodobnosti:

$$p(w_n \mid w_1, w_2, w_3, ..., w_{n-1}) = \frac{p(w_1, w_2, w_3, ..., w_n)}{p(w_1)p(w_2 \mid w_1)...(p(w_{n-1} \mid w_1, w_2, w_3, ..., w_{n-2})} \quad (2.12)$$

V praxi se nejčastěji používá bigramový nebo trigramový jazykový model. Jestliže pravděpodobnosti vyjádříme pomocí četností, které získáme z trénovacích dat, pak pro bigram platí:

$$p(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)}{C(w_{n-1})} \quad (2.13)$$

kde $C(w_{n-1})$ je počet výskytů slova w_{n-1},

$C(w_{n-1}, w_n)$ je počet výskytů dvojic slov w_{n-1}, w_n.

Hodnoty pravděpodobností jsou z definice menší než jedna a pro velké množství dat jsou velice malé. Z tohoto důvodu mnoho programů pro výpočet podmíněné pravděpodobnosti bigramů používá logaritmy pravděpodobností.

Základní (nevyhlazený) bigramový jazykový model je matice, jejíž prvky jsou podmíněné pravděpodobnosti určené pro všechny možné dvojice sousedních slov, které se objeví v trénovacích datech. Posloupnosti slov, které se v trénovacích datech neobjeví, mají hodnotu pravděpodobnosti rovnu nule. Od takto vytvořeného jazykového modelu se odvozují všechny vyhlazovací metody.

2.4 Vyhlazování jazykového modelu

Vyhlazování jazykového modelu se používá z důvodu velkého počtu nulových hodnot, které se vyskytují v bigramové matici. Každý trénovací korpus je konečný a nemůže obsahovat všechny dvojice slov. Nulová hodnota se v matici může objevit v případě, že se daná posloupnost slov nebo dané slovo v trénovacím korpusu neobjevily. V testovacím korpusu se ale objevit může. Proto se používají vyhlazovací algoritmy, které nulovým hodnotám v matici přiřadí malé nenulové pravděpodobnosti.

2.4.1 Add-One Smoothing

Tento typ vyhlazování je nejjednodušší vyhlazovací technikou. V případě unigramů se k počtu všech slov včetně těch, které se v textu nevyskytly, přičte 1. Nechť N je počet všech slovních tvarů, V je počet slov ve slovníku, $C(w)$ je počet výskytů slova w a $C(w_{n-1}, w_n)$ je počet výskytů dvojic slov w_{n-1}, w_n.

Potom pro nevyhlazený unigram platí:

$$p(w) = \frac{C(w)}{N} \tag{2.14}$$

Pravděpodobnost pro vyhlazený unigram se vypočítá podle:

$$p_{+1}(w) = \frac{C(w)+1}{N+V} \tag{2.15}$$

Obdobně platí pro nevyhlazený bigram:

$$p(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)}{C(w_{n-1})} \tag{2.16}$$

Pravděpodobnost pro vyhlazený bigram se vypočítá podle:

$$p_{+1}(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)+1}{C(w_{n-1})+V} \tag{2.17}$$

Nevýhodou je fakt, že pro viděné dvojice slov je pravděpodobnost „podhodnocená" a pro neviděné dvojice slov je pravděpodobnost „nadhodnocená".

2.4.2 Witten-Bell Discounting

Tento typ vyhlazování je založen na myšlence, že pomocí počtu slov, které se v korpusu vyskytly jednou, lze odhadnout počet slov, které se v korpusu dosud neobjevily. Následující vztahy budeme uvažovat pro bigramový model. Jestliže je podmíněná pravděpodobnost $p(w_n \mid w_{n-1})$ rovna 0, je vyhlazená pravděpodobnost vyšší v případě, že se slovo w_{n-1} objeví jako první slovo v mnoha dvojicích, a nižší, když se slovo w_{n-1} objeví jako první slovo v málo dvojicích. Problém nastane v případě, že se v trénovacích datech slova w_{n-1}, w_n neobjeví. Pak zůstává pravděpodobnost vyhlazeného bigramu stále nulová. V následujících vztazích jsou uvedeny pravděpodobnosti pro vyhlazený bigram.

Jestliže $C(w_{n-1}, w_n) = 0$:

$$p_{WB}(w_n \mid w_{n-1}) = \frac{T(w_{n-1})}{Z(w_{n-1})(C(w_{n-1})+T(w_{n-1}))} \tag{2.18}$$

Jestliže $C(w_{n-1}, w_n) > 0$:

$$p_{WB}(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)}{C(w_{n-1}) + T(w_{n-1})} \tag{2.19}$$

kde $C(w_{n-1})$ je počet výskytů slova w_{n-1},

$C(w_{n-1}, w_n)$ je počet výskytů dvojic slov w_{n-1}, w_n,

$T(w_{n-1})$ je počet rozdílných dvojic sousedních slov, jejichž první slovo je w_{n-1},

$Z(w_{n-1})$ je počet dvojic sousedních slov, které se neobjevily v trénovacích datech a jejichž první slovo je w_{n-1}.

2.4.3 Backoff Smoothing

Backoff Smoothing odhaduje pravděpodobnosti $\hat{p}$ n-gramů, které se v textu neobjevily, použitím nižší úrovně n-gramů. Tento typ vyhlazování používá nižší úrovně podmíněných pravděpodobností n-gramů pro výpočet vyšších jen v případě, že vyšší úroveň n-gramu má nulovou hodnotu podmíněné pravděpodobnosti. Tento algoritmus se často používá, protože kombinuje informace z různých n-gramů a řeší flexibilně vzájemný kompromis mezi přesností a komplexností. Vztah pro trigramy je uveden ve vztahu (2.20).

$$\hat{p}(w_n \mid w_{n-2}, w_{n-1}) = \begin{cases} p(w_n \mid w_{n-2}, w_{n-1}), & \text{pro } C(w_{n-2}, w_{n-1}, w_n) > 0 \\ \alpha_1 p(w_n \mid w_{n-1}), & \text{pro } C(w_{n-2}, w_{n-1}, w_n) = 0 \text{ a } C(w_{n-1}, w_n) > 0 \\ \alpha_2 p(w_n), & \text{jinde} \end{cases} \tag{2.20}$$

Koeficienty α_1, α_2 jsou váhové koeficienty, které zaručují, že součet podmíněných pravděpodobností daného slova přes všechny odpovídající n-gramy je roven hodnotě 1.

2.4.4 Linear Interpolation Smoothing

Vyhlazování nazvané lineární interpolace (Linear Interpolation Smoothing) používá pro odhad podmíněných pravděpodobností $\hat{p}$ vyšších úrovní vždy všech n-gramů nižších úrovní. Každý člen je vážen lineárním koeficientem λ_i. Počet slov ve slovníku je V.

V případě trigramů se používá vzorec:

$$\hat{p}(w_n \mid w_{n-1}, w_{n-2}) = \lambda_3 p(w_n \mid w_{n-1}, w_{n-2}) + \lambda_2 p(w_n \mid w_{n-1}) + \lambda_1 p(w_n) + \lambda_0 / V \tag{2.21}$$

Analogický vztah platí pro bigramy:

$$\hat{p}(w_n \mid w_{n-1}) = \lambda_2 p(w_n \mid w_{n-1}) + \lambda_1 p(w_n) + \lambda_0 / V \tag{2.22}$$

Hodnoty podmíněných pravděpodobností trigramů, bigramů a unigramů se stanoví z trénovacích dat. Pro odhad koeficientů λ_i se používají „odložená data" (data oddělená od hlavní trénovací množiny). Při použití trénovacích dat pro odhad koeficientů je koeficient u trigramů λ_3 (resp. bigramů λ_2) roven jedné a ostatní koeficienty jsou nulové. Koeficienty jsou stanoveny tak, aby maximalizovaly pravděpodobnost odložené části dat $p(W_H)$:

$$p(W_H) = \prod_{i=1...H} p(w_i \mid w_{i-1}) \tag{2.23}$$

kde H je počet slov v odložených datech

Pro výpočet koeficientů λ_i se používá EM algoritmus (Expectation – Maximization Algorithm), jehož postup pro bigramy je popsán níže [MRVA 2000].

1. Z trénovacích dat se vypočítají unigramy $p(w_n)$, bigramy $p(w_n| w_{n-1})$ a $1/V$, kde V je počet slov ve slovníku.
2. Na odložených datech (held-out datech) se spočítají nepodmíněné relativní frekvence dvojic slov $C_{BH}(w_{n-1}, w_n)$.
3. Pro první iteraci hodnot koeficientů λ_i je nejvhodnější použít rovnoměrné rozdělení, v případě bigramů je $\lambda_0 = \lambda_1 = \lambda_2 = 1/3$.
4. Provede se výpočet podmíněné pravděpodobnosti:

$$\hat{p}(w_n \mid w_{n-1}) = \lambda_2 p(w_n \mid w_{n-1}) + \lambda_1 p(w_n) + \lambda_0 / V \tag{2.24}$$

5. Provedou se výpočty:

$$C(\lambda_0) = \sum_{i=1...H} C_{BH}(w_{i-1}, w_i) \frac{\lambda_0}{V \cdot \hat{p}(w_{i-1} \mid w_i)} \tag{2.25}$$

$$C(\lambda_1) = \sum_{i=1...H} C_{BH}(w_{i-1}, w_i) \frac{\lambda_1 p(w_i)}{\hat{p}(w_{i-1} \mid w_i)} \tag{2.26}$$

$$C(\lambda_2) = \sum_{i=1...H} C_{BH}(w_{i-1}, w_i) \frac{\lambda_2 p(w_{i-1} \mid w_i)}{\hat{p}(w_{i-1} \mid w_i)} \tag{2.27}$$

kde H je počet slov v odložených datech

6. Spočítají se nové hodnoty λ_i podle rovnice:

$$\lambda_i = \frac{C(\lambda_i)}{\sum_{j=0}^{2} C(\lambda_j)} \quad \text{pro } i = 0,1,2 \tag{2.28}$$

Spočítá se rozdíl mezi starými a novými hodnotami λ_i. Jestliže je aspoň jeden rozdíl větší než stanovená tolerance ε, algoritmus se od kroku 4 opakuje.

Popsaným EM algoritmem se stanoví koeficienty λ_i, které se použijí pro výpočet vyhlazených bigramů.

2.5 Jazykový model založený na třídách

Jazykový model založený na třídách (Class-Based Language Model) určuje závislosti mezi třídami (značkami) slov a mezi značkami a slovy místo závislostí mezi konkrétními slovy. Pro tvorbu takového jazykového modelu je třeba jednotlivým slovům přiřadit značky (slova zařadit do tříd). Značky jsou většinou stanoveny na základě syntaktických a sémantických vlastností slov.

Předpokládejme, že existuje mapovací funkce G, která přiřazuje každému slovu w_n v korpusu značku c_n.

$$G : c_n = G(w_n) \tag{2.29}$$

Trénovací množina slov (w_1, w_2, …, w_T) se tak rozšíří na množinu dvojic – slovo a příslušná značka: ($\langle w_1,G(w_1)\rangle$, $\langle w_2,G(w_2)\rangle$, …, $\langle w_T,G(w_T)\rangle$).

Jestliže je definována mapovací funkce, je možné nahradit slova značkami. Pak v případě bigramů pro pravděpodobnost $p(W)$, kde $W = \{w_1, w_2, w_3, ..., w_n\}$, platí [JURAFSKY 2000]:

$$p(W) = \prod_{i=1}^{n} p(w_i \mid c_i) \cdot p(c_i \mid c_{i-1}) \tag{2.30}$$

kde $p(w_i \mid c_i)$ je podmíněná pravděpodobnost, s jakou bude k dané značce přiřazeno dané slovo,

$p(c_i \mid c_{i-1})$ je bigram značek.

Jazykový model lze přepsat následujícím způsobem:

$$p(w_n \mid w_1^{n-1}) = p(w_n \mid c_n)p(c_n \mid c_1^{n-1}) \tag{2.31}$$

kde w_1^{n-1} je historie slov,

c_1^{n-1} je historie značek.

Pro bigramový jazykový model založený na třídách potom platí:

$$p(w_n \mid w_{n-1}) = p(w_n \mid c_n)p(c_n \mid c_{n-1}) \tag{2.32}$$

Tento jazykový model sestává ze dvou složek:

- bigram značek

$$p(c_n \mid c_{n-1}) = \frac{C(c_n, c_{n-1})}{C(c_{n-1})} \tag{2.33}$$

kde $C(c_n, c_{n-1})$ je počet výskytů dvojice značek c_n, c_{n-1},

$C(c_{n-1})$ je počet výskytů značky c_{n-1}.

- podmíněná pravděpodobnost, s jakou bude k dané značce přiřazeno dané slovo

$$p(w_n \mid c_n) = \frac{C(w_n, c_n)}{C(c_n)} \tag{2.34}$$

kde $C(c_n)$ je počet výskytů značky c_n,

$C(w_n, c_n)$ je počet současného výskytu slova w_n se značkou c_n, pokud je jednomu slovu přiřazena jen jedna značka, pak $C(w_n, c_n) = C(w_n)$.

Existuje několik způsobů, jak stanovit funkci G a přiřadit jednotlivá slova do tříd. Jeden z přístupů je přístup morfologický, kdy se slova rozdělují do tříd podle slovních druhů. Další přístup využívá seskupení podobně se chovajících slov ve větě do tříd. Například samostatné třídy by zahrnovaly jména dnů, měsíců, měst apod. Pro rozdělení slov do tříd je možno využít také statistických metod. Jedna z nejpoužívanějších metod je popsána v [BROWN 1992].

V tomto případě definujeme reálnou funkci L, která maximalizuje:

$$L(G) = \frac{1}{N}\sum_{i=1}^{N} \log_2 (p(w_i \mid c_i) p(c_i \mid c_{i-1})) \tag{2.35}$$

Funkci upravíme následujícím způsobem [IRCING 2003a]:

$$\begin{aligned}
L(G) &= \frac{1}{N}\sum_{i=1}^{N} \log_2 (p(w_i \mid c_i) p(c_i \mid c_{i-1})) = \frac{1}{N}\sum_{i=1}^{N} \log_2 (p(w_i \mid c_i) p(c_i)\, p(c_i \mid c_{i-1}) / p(c_i)) = \\
&= \frac{1}{N} \log_2 (p(w_i, c_i)\, p(c_i \mid c_{i-1}) / p(c_i)) = \frac{1}{N}\sum_{i=1}^{N} \log_2 p(w_i, c_i) + \frac{1}{N}\sum_{i=1}^{N} \log_2 \frac{p(c_i \mid c_{i-1})}{p(c_i)} = \\
&= \frac{1}{N}\sum_{i=1}^{N} \log_2 p(w_i) + \frac{1}{N}\sum_{i=1}^{N} \log_2 \frac{p(c_i \mid c_{i-1}) p(c_{i-1})}{p(c_i) p(c_{i-1})} = -H(w) + \frac{1}{N}\sum_{i=1}^{N} \log_2 \frac{p(c_i, c_{i-1})}{p(c_{i-1}) p(c_i)} = \\
&= -H(w) + \sum_{c_1, c_2 \in C} p(c_1, c_2) \log_2 \frac{p(c_1, c_2)}{p(c_1) p(c_2)} = -H(w) + I(c_1, c_2)
\end{aligned} \tag{2.36}$$

Výsledkem této úpravy je součet záporné hodnoty entropie $H(w)$ a vzájemné informace $I(c_1, c_2)$ dvou sousedních sloučených tříd. Protože entropie nezávisí na funkci G, je maximalizace funkce $L(G)$ stejná jako maximalizace vzájemné informace $I(c_1, c_2)$.

Vzájemná informace určuje, jak se pravděpodobnost $p(c_1, c_2)$ odchyluje od součinu pravděpodobností $p(c_1)$ a $p(c_2)$. Čím vyšší je vzájemná informace, tím více je jedna veličina závislá na druhé. Vzájemnou informaci dvou tříd c_1 a c_2 lze vyjádřit vztahem:

$$I(c_1,c_2) = \sum_{c_1,c_2 \in C} p(c_1,c_2)\log_2 \frac{p(c_1,c_2)}{p(c_1)p(c_2)} \tag{2.37}$$

2.5.1 Hladový algoritmus

Hladový algoritmus (Greedy Algorithm) se používá pro slučování slov do tříd, přičemž při sloučení dochází k minimální ztrátě vzájemné informace. Postup slučování lze shrnout do čtyř bodů:

1. Každé slovo je v jedné třídě (vlastní).
2. Sloučí se dvě třídy tak, aby se maximalizovala vzájemná informace *I*.
3. Nastaví se nová množina dat s jednou seskupenou třídou bez dvou tříd, které byly seskupeny.
4. Opakují se body 2 a 3, dokud se nedosáhne zadané velikosti.

Pro slovník o velikosti V je výpočetní náročnost tohoto algoritmu úměrná V^5 [HAJIČ 2000a]. Pro slovník o velikosti $V = 100$ proces slučování trvá při použití současné výpočetní techniky několik hodin. Slovníky obvykle bývají daleko větší, takže se tento algoritmus pro slučování nepoužívá. Brown v [BROWN 1992] prezentoval hladový algoritmus, který je založen na minimalizaci ztráty vzájemné informace během slučování. Algoritmus používá „elegantní" postupy při výpočtech a aktualizaci matice ztrát během slučování. U přepočítávání bigramové matice není třeba provést přepočet všech hodnot, ale jenom těch, které jsou závislé na sloučení. Takže se odstraní sloupce a řádky tříd, které se sloučí do jedné, a přidá se řádek a sloupec sloučené třídy.

Předpokládejme, že po $V - k$ sloučení máme k tříd. Nechť

$$p_{kl}(l) = \sum_{r=1}^{k} p_k(l,r) \tag{2.38}$$

$$p_{kr}(r) = \sum_{l=1}^{k} p_k(l,r) \tag{2.39}$$

Definujme

$$q_k(l,r) = p_k(l,r)\log \frac{p_k(l,r)}{p_{kl}(l)\cdot p_{kr}(r)} \tag{2.40}$$

$$s_k(a) = \sum_{l=1}^{k} q_k(l,a) + \sum_{r=1}^{k} q_k(a,r) - q_k(a,a) \tag{2.41}$$

Označme $a + b$ novou sloučenou třídu. Potom odečtená část ze vzájemné informace po sloučení dvou tříd a, b je:

$$sub_k(a,b) = s_k(a) + s_k(b) - q_k(a,b) - q_k(b,a) \tag{2.42}$$

Přičtená část po sloučení tříd a, b je:

$$add_k(a,b) = \sum_{l=1,l\neq a,b}^{k} q_k(l,a+b) + \sum_{r=1,r\neq a,b}^{k} q_k(a+b,r) + q_k(a+b,a+b) \tag{2.43}$$

Pro vzájemnou informaci po sloučení a, b platí:

$$I_k(a,b) = I_k - sub_k(a,b) + add_k(a,b) \tag{2.44}$$

Cílem je nalézt takovou dvojici (a, b), pro kterou je ztráta informace $L_k(a, b)$ nejmenší.

$$L_k(a,b) = sub_k(a,b) - add_k(a,b) \tag{2.45}$$

Po sloučení (a, b jsou sloučené třídy v a) je třeba provést aktualizaci. Index použitý pro další iteraci je $k - 1$ a při aktualizaci musí být splněny podmínky: $i \neq a$, $i \neq b$, $j \neq a$, $j \neq b$.

$$s_{k-1}(i) = s_k(i) - q_k(i,a) - q_k(a,i) - q_k(i,b) - q_k(b,i) + q_{k-1}(a,i) + q_{k-1}(i,a) \tag{2.46}$$

$$\begin{aligned} L_{k-1}(i,j) = {} & L_k(i,j) - s_k(i) + s_{k-1}(i) - s_k(j) + s_{k-1}(j) + \\ & + q_k(i+j,a) + q_k(a,i+j) + q_k(i+j,b) + q_k(b,i+j) - \\ & - q_{k-1}(i+j,a) - q_{k-1}(a,i+j) \end{aligned} \tag{2.47}$$

Pro další iteraci jsou potřeba hodnoty $s_{k-1}(a)$ a $L_{k-1}(a, i)$. Hodnoty $s_{k-1}(b)$ a $L_{k-1}(b, i)$ již dále nejsou potřebné a není nutné je uchovávat pro další výpočty. Celý proces se ukončí v případě, že $k = 1$, to znamená, že všechna slova jsou sloučena do jedné třídy, nebo když počet tříd dosáhne předurčené velikosti.

2.6 Metriky používané k ohodnocení systémů

Pro ohodnocení systémů rozpoznávání řeči jsou nezbytné určité metriky, na jejichž základě je možné jednotlivé systémy porovnávat. Pro porovnání těchto systémů se využívají tři typy chyb [HUANG 2001]:

- substituce (S) – počet nesprávných slov, která byla nahrazena správným slovem
- vymazání (D) – počet vynechaných slov
- vložení (I) – počet vložených slov

Jednou z používaných metrik je přesnost rozpoznání (accuracy), která využívá výše uvedených chyb a vypočítá se podle:

$$Acc = \frac{N - D - S - I}{N} \tag{2.48}$$

kde N je počet slov ve správné větě.

Někdy se místo přesnosti rozpoznání používá správnost rozpoznávání (correctness). Tato metrika ve skutečnosti přestavuje procento správně rozpoznaných slov a vypočítá se podle:

$$Corr = \frac{N - D - S}{N} \tag{2.49}$$

Nejběžnější metriky používané k ohodnocení n-gramových jazykových systémů jsou entropie a perplexita. Kromě těchto metrik lze jazykové modely porovnat na zkušební množině dat (testovacích datech) [MRVA 2000].

Entropie je míra informace. Může být použita jako míra pro určení množství informace v určité gramatice (abecedě), pro vyjádření toho, jak dobře je přiřazena určitá gramatika danému jazyku. Jsou-li zadány dvě gramatiky (abecedy) pro jeden korpus, podle entropie lze zjistit, která gramatika je pro daný korpus vhodnější. Čím vyšší entropii má jazykový model, tím obtížněji se na jeho základě předpovídá budoucí text.

Entropie je maximální v případě rovnoměrného rozdělení pravděpodobnosti a zmenšuje se, pokud se vyskytne událost s vyšší pravděpodobností výskytu. Při rovnoměrném rozdělení dostaneme nejvyšší stupeň nejistoty, neboť každá událost může nastat se stejnou pravděpodobností.

Entropie H náhodné veličiny X, kterou mohou být slova, písmena, slovní druhy, se vypočítá ze vztahu:

$$H(X) = -\sum_{x \in \Omega} p(x) \log_2 p(x) \tag{2.50}$$

kde $p(x)$ je pravděpodobnostní funkce náhodné veličiny X.

Vzhledem k tomu, že logaritmus ve vzorci (2.50) má základ 2, je výsledná hodnota vyjádřena v bitech. Entropii lze interpretovat také jako nejmenší počet bitů potřebných k zakódování textu, písmen apod.

Perplexita G vyjadřuje počet možností a je definována takto:

$$G(x) = 2^{H(X)} \tag{2.51}$$

Perplexita pro jazykový model ze slov je definována jako:

$$G = 2^{LP} \tag{2.52}$$

kde

$$LP = -\frac{1}{N}\sum_{i=1}^{N} \log_2 p(w_i \mid w_{i-1}) \tag{2.53}$$

Perplexita pro jazykový model ze slov je velikost množiny slov, ze které jazykový model náhodně vybírá následující slovo.

Kapitola 3

Korpus a slovník

Jazykový korpus je velmi rozsáhlý soubor textů přirozeného jazyka [ČNK 2000]. Současné korpusy psaného jazyka se obvykle vytvářejí sběrem celých textů, a to tak, aby co nejvěrněji reprezentovaly daný jazyk. Tyto texty, zahrnující např. novinové články, krásnou literaturu, odborné publikace, jsou tříděny a věcně, strukturně a lingvisticky označkovány (opatřeny dodatečnými identifikačními, strukturními a jazykovědnými údaji). Korpus bývá uložen v textové podobě a je organizován se zřetelem na využití pro určitý cíl [VYMAZAL 2001]. Data nejrůznějšího druhu se v korpusu nacházejí ve své přirozené kontextové podobě, což umožňuje jejich všestranné a objektivní studium. Velký rozsah vybudovaného korpusu minimalizuje to, že převládnou zvláštní a okrajová užití slovních tvarů nad základními a typickými.

Korpus obsahuje různé slovní tvary (wordform) na rozdíl od lexikonu, v němž jsou základní tvary slov (lemma) popř. pravidla pro skloňování, časování apod. V anglickém názvosloví se používá slovo **types** pro počet odlišných slov v korpusu a **tokens** pro celkový počet slov v korpusu.

Primárním zdrojem korpusových dat jsou elektronické dokumenty. Dnes jsou to prakticky veškeré tištěné dokumenty – noviny, časopisy, knihy. Texty lze získávat též prostřednictvím sítě internet. Řada moderních korpusů je aspoň zčásti dostupná v počítačové síti internet.

Korpusy se mohou lišit zaměřením textů. Obecné korpusy se snaží zachytit charakteristické rysy jazyka a obsahují mnoho různě tématicky zaměřených dokumentů. U obecných korpusů je cílem zachytit jazyk v celé jeho šíři na základě lingvisticky podložených kritérií. Slouží zejména k vytváření slovníků. Speciální korpus je soubor textů

zaměřený na nějakou užší oblast podle stanoveného kritéria (např. korpus určitého nářečí, korpus autorský nebo korpus konkrétního literárního díla).

Kromě textových korpusů je možné se setkat i s mluvenými korpusy. Mluvený korpus je sbírka vyslovených dat. Kromě slov mohou obsahovat pauzy, přeřeknutí, zaplněné pauzy (např. hm), které se v psaném korpusu neobjeví.

Uložení korpusu v počítači může mít různou podobu. Nejvolnější formou je pouhý archív dokumentů v různém formátu a v různém kódování podle zdroje. Organizovanější formou jsou textové banky, ve kterých jsou texty uloženy v jednotném formátu a doplněny o další informace (rozdělení na články, uvedení zdroje apod.). Obecně bývají korpusy nejčastěji uloženy v SGML formátu. SGML (Standard Generalized Markup Language) je standardní jazyk určený k formálnímu popisu struktury dokumentů. SGML formát umožňuje uložení textů v elektronické podobě a je nezávislý jak na softwarové, tak na hardwarové platformě. Dokumenty SGML obsahují text a multimediální prvky a dále mohou obsahovat nadpisy všech úrovní, odstavce a několik formátovacích prvků.

3.1 Příklady korpusů

K nejznámějším korpusům v České republice patří **Český národní korpus** (ČNK), který byl založen v roce 1994 a pracuje na něm Ústav Českého národního korpusu (ÚČNK) na Filozofické fakultě Univerzity Karlovy v Praze. Největší složka ČNK je synchronní psaný korpus **SYN2000**, který obsahuje přibližně 100 miliónů textových slov. V korpusu je zastoupena publicistika (60 %), odborná literatura (25 %) a beletrie (15 %). U každého slova (tj. výskytu slova v textu) je prostřednictvím mezinárodně kompatibilních značek uvedena informace o jeho zdroji (původním textu), o jeho morfologických vlastnostech (slovním druhu a gramatických kategoriích) a rovněž jeho lemma (základní tvar) [ČERMÁK 2004]. K dalším známým korpusům patří **Pražský závislostní korpus** (PDT) vytvořený na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze v Ústavu formální a aplikované lingvistiky (ÚFAL MFF UK). V současnosti obsahuje dva milióny ručně morfologicky anotovaných dat a další softwarové nástroje pro prohledávání korpusu, anotaci dat a jazykovou analýzu. PDT je sestaven z článků z novin a časopisů [PDT 2005]. Na Fakultě informatiky Masarykovy univerzity (FI MU) v Brně byl vytvořen označkovaný korpus publicistických textů **DESAM**. Každému slovnímu tvaru v textu je přiřazen základní tvar a gramatická značka obsahující slovní druh a příslušné gramatické kategorie. Všechny nejednoznačnosti byly ručně zjednoznačněny a korpus byl dále ručně pročištěn

(byly odstraněny překlepy a chyby na různých úrovních značkování). Název korpusu je odvozen od desambiguace (zjednoznačnění). Korpus obsahuje 1,2 miliónu slov [VYMAZAL 2001].

Jeden z nejznámějších korpusů pro anglický jazyk je **British National Corpus** (BNC). Jde o sbírku 100 miliónů slov psané i mluvené britské angličtiny. Psaná část tvoří 90 % a obsahuje např. články z novin, periodik a časopisů, vědecké publikace, beletrii. Mluvená část (10 %) obsahuje neformální konverzace nahrané dobrovolníky různého věku, pohlaví, regionu, sociální skupiny, dále nahrávky formálních obchodních schůzek, zábavných pořadů apod. [BNC]. Ke značkovaným korpusům patří **Brown Corpus**, který obsahuje jeden milión slov americké angličtiny z roku 1961. Tento korpus obsahuje 374 vzorků informativní prózy a 126 vzorků imaginativní prózy. Aby byl korpus všeobecný, byl sestavován z textů z 15 různých oblastí (např. tisk, beletrie, humor, fikce) [BROWNC]. Jeden z řady mluvených korpusů pro anglický jazyk je **British Academic Spoken English Corpus** (BASE), který obsahuje přednášky a semináře většinou nahrané na video. Korpus se skládá z nahrávek ze 162 přednášek a 41 semináře. Většina z nich je přepsaná a zkontrolovaná [BASE].

Mezi nejznámější mluvené korpusy v České republice patří **Pražský mluvený korpus** (PMK) a **Brněnský mluvený korpus** (BMK). Oba tyto mluvené korpusy jsou součástí Českého národního korpusu.

Pražský mluvený korpus je prvním korpusem mluvené češtiny a zachycuje autentickou mluvenou češtinu, hlavně obecnou a tématicky nespecializovanou, resp. neomezovanou, z oblasti Prahy a jejího okolí. Počet všech slovních tvarů je 819 267, počet různých slovních tvarů je 49 089.

Brněnský mluvený korpus je elektronickým přepisem 250 anonymních magnetofonových nahrávek z let 1994–1999 zachycujících 294 mluvčích. Počet všech slovních tvarů je 596 009, počet různých slovních tvarů je 39 615.

3.2 Morfologická analýza

Morfologická analýza je zobrazení, které každému slovu (slovnímu tvaru) přiřadí dvojici lemma (základní heslový tvar) – značka. Lemma je lexikální jednotka, slovníkové heslo. Nejčastěji bývá reprezentováno základním tvarem slova, ale může to být i číselný odkaz [OSOLSOBĚ 1995].

Pro stanovení jazykového modelu a jeho testování je potřeba velké množství dat (korpus) rozdělit na dvě části. Trénovací množinu a testovací množinu. Parametry statistického jazykového modelu jsou vypočítány z trénovacích dat a použity pro testovací data. Rozdělení dat na trénovací a testovací se využívá také pro porovnávání různých jazykových modelů.

Je třeba, aby byl korpus velmi pečlivě sestaven. V případě, že je korpus příliš specifický, jazykový model nelze zobecnit (použít pro obecné věty). Na druhou stranu jazykový model vytvořený z obecných vět nebude použitelný pro příliš specifické a např. odborně zaměřené věty.

3.3 Tvorba slovníku

Pro zpracování řeči i textu je třeba mít k dispozici vhodný slovník. Slovník je seznam typů slov, což jsou odlišné položky slovníku. To znamená, že slova pán a pánovi jsou dvě odlišné položky slovníku se stejným jazykovým kmenem. Slovník, který je používán v této práci, byl vytvořen na Technické univerzitě v Liberci z internetové verze článků z Lidových novin, z internetových novin Neviditelný pes, ze 4 diplomových prací a 27 novel. Kromě typu slova obsahuje i jeho četnost. Pro rozpoznávání řeči je nutné přidat do slovníku transkripci, pro zpracování textu lze do slovníku přidat další informaci o typu slova. Vzhledem k tomu, že úkolem práce bylo vytvoření jazykového modelu založeného na třídách (Class-Based Language Model), je ke každému typu slova přidána do slovníku informace o příslušných gramatických značkách, které lze danému typu slova přiřadit. V tabulce 3.1 je zobrazena jedna položka slovníku. Pole „skupina" obsahuje počet gramatických značek jako dvojmístné číslo a seznam těchto značek – vše oddělené lomítkem. Pole „počet" zobrazuje počet výskytů slova.

Tab. 3.1: Položka slovníku

slovo	skupina	počet
technika	03/01zj1/01maj4/01mj2	4230

Pro přiřazení příslušných značek ke každému slovu ve slovníku byly postupně použity tři metody (přístupy). Nejprve byly přidány do slovníku značky těch slov, které byly obsaženy ve větách označkovaných ručně. Tato slova tvořila jen velmi malou část slovníku (4 %). Proto byla použita syntetická metoda založená na generování všech možných slovních tvarů s jejich morfologickými značkami. Tvary byly generovány

s použitím pravidel pro tvorbu slov a s použitím množiny kořenů, předpon, přípon a koncovek (poskytnuto ÚFAL MFF UK v Praze). Touto metodou bylo označkováno dalších asi 49 % slov ze slovníku. Třetí přístup byl částečně manuální. Pomocí filtrů s typickými předponami, příponami a koncovkami byly přiřazeny příslušné značky k dalším slovům ve slovníku. Některým slovům ve slovníku byly příslušné značky přiřazeny ručně.

3.4 Příprava dat

Pro tvorbu jazykového modelu byl vytvořen korpus z novinových článků, článků z internetu, z částí knížek apod. Vytvořený korpus neobsahuje odborné a vědecké články. Všechna data jsou převedena na prostý text a splňují níže uvedené požadavky.

- Každá věta je na jednom řádku.
- Jednotlivá slova včetně interpunkce jsou oddělena mezerou.
- V korpusu nejsou použita čísla a zkratky zakončené tečkou – obojí je přepsáno do slov.
- Spojovník je od slov oddělen mezerami.
- Velká a malá písmena jsou v korpusu ponechána beze změny stejně jako další zkratky (např. ODS).
- Každá věta je vždy ukončena tečkou, vykřičníkem, otazníkem nebo uvozovkami.
- Věty neobsahují nespisovná slova (např. hladovej) a gramaticky nesprávná spojení (např. ty děvčata).

Vytvořený korpus obsahuje celkem 8 800 vět (130 718 slov včetně interpunkce). Z toho 3 300 vět sloužilo jako trénovací data a 500 vět jako testovací data. Pro další experimenty bylo upraveno 5 000 vět. V tabulce 3.2 jsou uvedeny počty slov ve větách s interpunkcí a bez interpunkce, procento interpunkce a procento OOV (out of vocabulary – slov, která nejsou ve slovníku).

Tab. 3.2: Statistika korpusu

počet vět	trénovací data	testovací data	data pro automatické značkování	celkem
	3 300	500T	5 000	8 800
počet slov s interpunkcí	44 331	8 867	77 520	130 718
počet slov bez interpunkce	38 084	7 836	67 440	113 360
interpunkce [%]	14,09	11,63	13,00	13,28
OOV [%]	0	0,34	1,65	1,02

Kapitola 4

Značkování

Značkování (Part-of-Speech Tagging) je přiřazení gramatické značky (tagu) každému slovu a obvykle i interpunkčnímu znaménku v korpusu. Značkovače se používají např. v rozpoznávání řeči a syntaktické analýze. Vstupem do značkovače je řetězec slov (věta) a množina značek a výstupem je posloupnost značek, kdy pro každé slovo je vybrána nejpravděpodobnější značka. Značky lze přiřazovat ručně nebo automaticky. Přiřazení značky není vždy jednoznačné. Desambiguace (zjednoznačnění) je velmi obtížný problém. Milióny slov nelze značkovat ručně a prakticky není možné se obejít bez chyb. Podle [ČERMÁK 2004] dosahují nejlepší programy pro desambiguaci aplikované na korpusy angličtiny 97–98% úspěšnosti. Úspěšnost morfologické desambiguace korpusu SYN2000 (Český národní korpus) dosahuje zhruba 94 % [ČERMÁK 2004]. Uvedený rozdíl vyplývá zejména z odlišných typologických vlastností češtiny a angličtiny. Angličtina je jazyk s poměrně velmi pevným slovosledem, takže se jak pravděpodobnostními metodami založenými na četnosti posloupností slov a jejich značek tak i nepravděpodobnostními metodami založenými na pravidlech značkuje mnohem úspěšněji.

4.1 Značkování textu

Existují různé přístupy ke značkování textu. Nejznámější značkovače (taggers) jsou značkovače založené na pravidlech (rule-based taggers), stochastické značkovače (stochastic taggers) a značkovače, které kombinují tyto dva způsoby (hybrid taggers).

4.1.1 Značkovače založené na pravidlech

Značkovače založené na pravidlech obsahují databázi, ve které jsou ručně napsaná pravidla, která specifikují, jaká značka má být danému slovu přiřazena. Tento typ

značkování je založen na dvouúrovňové architektuře. Nejdříve se přiřadí seznam potenciálních značek každému slovu. Poté dojde k výběru správné značky s použitím seznamu ručně psaných pravidel resp. omezení o jednoznačnosti. Anglický jazyk má danou strukturu s pevným slovosledem, takže není problém pravidla stanovit.

Anglický morfologický analyzátor **ENGTWOL** používá pro značkování 4 000 pravidel a slovník s 56 000 záznamy [VOUTILAINEN 1993]. Tento morfologický analyzátor je součástí syntaktického analyzátoru **ENGCG** (English Constraint Grammar), který byl vyvinut na univerzitě v Helsinkách a který provádí morfologicko-syntaktickou analýzu běžného anglického textu. Systém se skládá z několika modulů. První modul – preprocesor – stanovuje mimo jiné hranice vět a slov. V další fázi je provedena morfologická analýza pomocí analyzátoru **ENGTWOL**, při níž je použita dvouúrovňová analýza a heuristická analýza neznámých slov. V poslední fázi se rozhoduje o morfologických a syntaktických nejednoznačnostech.

Pro český jazyk jsou vytvořeny značkovače založené na pravidlech a morfologické analyzátory např. v Ústavu formální a aplikované lingvistiky (ÚFAL) MFF UK v Praze [OLIVA 2000].

4.1.2 Stochastické značkovače

Stochastické značkovače řeší problém nejednoznačnosti použitím trénovacího korpusu. Ten se používá k výpočtu pravděpodobnosti daného slova, kterému je přiřazena daná značka a kontext. Jeden z nejrozšířenějších stochastických značkovačů využívá skryté Markovovy modely (HMM). Tento značkovač (HMM tagger) vybírá posloupnost značek, která maximalizuje pravděpodobnost $p(slovo|značka) \cdot p(značka|předchozích\ n\ značek)$. Pro bigramový značkovač potom platí:

$$c_i = \arg\max_j p(w_i \mid c_j) \cdot p(c_j \mid c_{i-1}) \qquad (4.1)$$

HMM tagger provádí výběr nejlepší sekvence značek pro větu, nikoli pro jednotlivá slova. Výběr se provádí pomocí Viterbiho algoritmu, který je založen na Bellmanově principu optimality. HMM tagger je nejčastěji trénován na ručně značkovaných datech.

Stochastic Parts Program [CHURCH 1988] je stochastický značkovač, který maximalizuje pravděpodobnost $p(značka|slovo) \cdot p(značka|předchozích\ n\ značek)$. Pro bigramový značkovač potom platí:

$$c_i = \arg\max_j p(c_i \mid w_j) \cdot p(c_j \mid c_{i-1}) \qquad (4.2)$$

Funkce argmax v rovnicích (4.1) a (4.2) znamená nalezení takového *j*, pro které je součin uvedených podmíněných pravděpodobností maximální.

K stochastickým značkovačům patří např. značkovač **TnT** (Trigram'n'Tags). Tento značkovač není optimalizován pro konkrétní jazyk. Je možné ho použít pro různé množiny značek a pro různé jazyky. K trénování využívá označkovaný korpus.

Czech HMM-based Tagger je značkovač vyvinutý v ÚFAL MFF UK v Praze. U tohoto značkovače se provádí předzpracování morfologickým analyzátorem.

4.1.3 Hybribní značkovače

Hybridní značkovače kombinují pravděpodobnostní a nepravděpodobnostní přístup. Značkování sestává z několika kroků, ve kterých se využívá jak gramatických pravidel většinou ručně sestavených, tak pravděpodobnostních metod s využitím trénovacího korpusu.

Mezi hybridní značkovače pro angličtinu patří např. značkovač **CLAWS4**, který byl vyvinut na univerzitě v Lancasteru. Skládá se z několika částí. Nejdříve je provedena segmentace korpusu na slova a věty. Poté je každému slovu přiřazena jedna nebo více značek. K přiřazení se používá slovník, seznam koncovek a seznam pravidel pro značkování neznámých slov. Pravděpodobnostní přístup se v tomto značkovači používá pro výběr nejlepší značky. Na základě pravidel se provádí zjednoznačnění nebo oprava značek. V poslední fázi je provedena grafická úprava výstupního souboru.

4.2 Příklady množin značek

4.2.1 Český jazyk

Pro český jazyk je vytvořeno několik různých množin značek (tagů), které jsou využívány k označkování různých korpusů.

Nejznámější jsou **Kompaktní tag systém pro češtinu** a **Poziční tag systém pro češtinu** vytvořené v ÚFAL MFF UK v Praze [HAJIČ 2000b]. Kompaktní tag systém se používá v českém morfologickém slovníku. Poziční tag systém se používá ke značkování českých textů a byl použit např. pro označkování Českého národního kropusu. Každý tag je reprezentován řetězcem 15 znaků. Každá pozice v řetězci určuje jistou gramatickou vlastnost. Tento systém kompletně popisuje morfologické kategorie v českém jazyce. V tabulce 4.1 je uveden detailní popis jednotlivých pozic. Např. slovu „doběhla" odpovídá

značka VpQW---XR-AA--1. První pozice je slovní druh, V – sloveso, další pozice udává detailní informaci o slovním druhu (p – příčestí minulé, činný rod), 3. pozice je rod (Q – rod ženský jednotný nebo střední množný), 4. pozice udává číslo (W – jednotné pro ženský rod, množné pro střední rod), 8. pozice je osoba (X – jakákoli), na 9. pozici je čas (R – minulý), 11. pozice představuje negaci (A – afirmace), na 12. pozici je informace o slovesném rodu (A – činný rod), poslední pozice udává variantu (1 – méně frekventované) a na pozicích, kde jsou uvedeny pomlčky, daná gramatická vlastnost pro dané slovo neexistuje.

Tab. 4.1: Popis jednotlivých pozic u pozičního tag systému pro český jazyk

1	POS	Slovní druh
2	SUBPOS	Slovní poddruh
3	GENDER	Rod
4	NUMBER	Číslo
5	CASE	Pád
6	POSSGENDER	Rod vlastníka
7	POSSNUMBER	Číslo vlastníka
8	PERSON	Osoba
9	TENSE	Čas
10	GRADE	Stupeň
11	NEGATION	Negace
12	VOICE	Slovesný rod
13	RESERVE1	Rezerva
14	RESERVE2	Rezerva
15	VAR	Varianta, styl

Další tag systém pro češtinu není poziční a jednotlivé značky jsou definovány jako posloupnost dvojic typu atribut – hodnota. Atribut, který se značí malým písmenem, reprezentuje některou z možných gramatických kategorií. Hodnota, která se značí velkým písmenem nebo číslicí, vyjadřuje aktuální hodnotu, jíž daná kategorie u daného tvaru nabývá [ČERMÁK 2004]. Tento tag sytém se používá pro morfologický analyzátor češtiny **Ajka**, který byl vyvinut v Brně na FI MU [SEDLÁČEK 1999]. Přehled některých symbolů, které se používají pro tvorbu značek, je uveden v tabulce 4.2. Například slovu „doběhla“ odpovídá značka k5eAaPmAp3gFnS.

Tab. 4.2: Přehled symbolů používaných pro tvorbu značek pro systém Ajka

Slovní druh – k	**Pád – c**	**Druh zájmena – x**
Podstatná jména: 1	1. pád: 1	Osobní: P
Přídavná jména: 2	2. pád: 2	Přivlastňovací: O
Zájmena: 3	3. pád: 3	Ukazovací: D
Číslovky: 4	4. pád: 4	Vymezovací: T
Slovesa: 5	5. pád: 5	
Příslovce: 6	6. pád: 6	**Druh zájmena – y**
Předložky: 7	7. pád: 7	Refexivní: F
Spojky: 8		Tázací: Q
Částice: 9	**Osoba – p**	Vztažné: R
Citoslovce: 0	1. osoba: 1	Záporné: N
Zkratky: A	2. osoba: 2	Neurčité: I
by, aby, kdyby: Y	3. osoba: 3	
	1. nebo 2. nebo 3. osoba: X	**Druh číslovky – x**
Rod – g		Základní: C
Mužský životný: M	**Typ (Mód) – m**	Řadové: O
Mužský neživotný: I	Infinitiv: F	Druhové: R
Ženský: F	Indikativ prézentu: I	Gramatika: G
Střední: N	Imperativ: R	Gramatika: H
Rodina (příjmení): R	Příčestí činné (minulé): A	
	Příčestí trpné: N	**Druh číslovky – y**
Číslo – n	Přechodník přítomný (současnost): S	Záporná: N
Jednotné: S	Přechodník minulý (dřívější děj): D	Neurčitá: I
Množné: P	Indikativ futura: B	
Duál: D		**Druh zájmenného příslovce – x**
Hromadné označení členů rodiny (Novákovi): R	**Vid – a**	Ukazovací: D
	Perfektum: P	Vymezovací: T
Stupeň – d	Imperfektum: I	Způsobové: M
Pozitiv: 1	Obouvidé: B	Stavové: S
Komparativ: 2		
Superlativ: 3	**Druhy spojek – x**	**Druh zájmenného příslovce – y**

	Souřadicí: C	Tázací: Q
Negace – e	Podřadicí: S	Vztažné: R
Afirmace: A		Záporné: N
Negace: N	**Speciální vzor – x**	Neurčité: I
	Půl: P	
		Typ tvaru – z
		tvar s příklonným -s: S

4.2.2 Anglický jazyk

Pro anglický jazyk – a to jak pro britskou tak pro americkou angličtinu – existuje řada tag systémů.

Penn Treebank POS Tagset pro americkou angličtinu obsahuje 48 značek, z toho je 36 gramatických značek (závislých na slovním druhu) a 12 dalších značek (pro interpunkci, měnu apod.). Gramatické značky mají dva až čtyři znaky, značky pro interpunkci mají jeden znak. Podrobný popis značek je např. v [SANTORINI 1990].

Brown Corpus je označkován gramatickými značkami, které jsou vybírány z množiny 87 značek. Kromě gramatických značek jsou do této množiny zahrnuty i značky pro interpunkci. Detailní popis značek je uveden např. v [FRANCIS 1979].

Pro označkování **British National Corpus (BNC)** je použita množina 61 značek označená C5. Každá značka sestává ze tří znaků. První dva určují slovní druh a třetí znak označuje podkategorii. Pro podrobnější značkování se používá množina 146 značek označená C7. Podrobnější popis značek je v [LEECH 2000].

Kapitola 5

Stanovení značek

Nevýhodou tvorby jazykového modelu i z velkého korpusu je nedostatek dat. Není možné, aby se v korpusu vyskytla všechna slova a slovní spojení. Jedním z řešení je seskupení podobně se chovajících slov do tříd. Tím získáme reálný odhad i pro slovní spojení, která se dosud v korpusu nevyskytla, ale ztratíme přesnost při nahrazení slova značkou (třídou).

Při stanovení značek byl využit přístup gramatický, statistický i pravděpodobnostní. První fáze zahrnovala stanovení značek podle slovních druhů a jejich gramatických kategorií (viz tabulka 5.1).

Tab. 5.1: Slovní druhy a jejich morfologické vlastnosti

Slovní druh	**Gramatické kategorie**
Podstatná jména	rod, číslo, pád
Přídavná jména	rod, číslo, pád
Zájmena	druh, rod, číslo, pád
Číslovky	druh, rod, číslo, pád
Slovesa	osoba, číslo, čas, způsob
Příslovce	druh
Předložky	pád, se kterým se pojí
Spojky	
Citoslovce	
Částice	
Vlastní jména	druh, pád
Interpunkce	

Při takto stanovených značkách může být k jednomu slovu přiřazeno až několik desítek značek. Jejich celkový počet se pohybuje okolo 500.

Vzhledem k tomu, že by značkování mělo být využito k tvorbě jazykového modelu založeného na třídách pro rozpoznávání spojité řeči, jsme se snažili dodržet tři zásady:

- stanovit co nejmenší počet značek,
- nejfrekventovanějším slovům přiřadit samostatné značky,
- dodržet, aby k jednomu slovu bylo přiřazeno nejvýše deset značek.

V druhé fázi byla vybrána slova s největší frekvencí a každému takovému slovu byla přiřazena samostatná značka. Tato slova se vyskytují v textu tak často, že tvoří přibližně 30 % jakéhokoli textu. Počet značek pro jednotlivé slovní druhy je uveden v tabulce 5.2.

Tab. 5.2: Počet značek stanovených podle frekvence slov pro jednotlivé slovní druhy

Slovní druh	Počet značek
Podstatná jména	7
Přídavná jména	0
Zájmena	30
Číslovky	11
Slovesa	31
Příslovce	15
Předložky	34
Spojky	15
Citoslovce	0
Částice	2
Vlastní jména	1
Interpunkce	4
Celkový počet	150

Nejfrekventovanější slova a jejich procentuální zastoupení ve slovníku a v korpusu jsou uvedeny v tabulce 5.3.

Tab. 5.3: Procentuální zastoupení nejfrekventovanějších slov ve slovníku a v korpusu

Slovo	Slovník		Korpus	
	Počet	Procentuální zastoupení [%]	Počet	Procentuální zastoupení [%]
a	4 071 924	3,00	2 826	2,16
v	3 313 921	2,44	1 805	1,38
se	2 934 202	2,16	2 947	2,25
na	2 520 329	1,86	1 913	1,46
je	1 522 516	1,12	1 282	0,98

že	1 376 966	1,01	1 160	0,89
s	1 133 405	0,84	737	0,56
z	1 086 698	0,80	763	0,58
o	1 064 948	0,78	672	0,51
to	955 627	0,70	1 028	0,79
do	931 620	0,69	821	0,63
i	847 125	0,62	692	0,53
ve	687 975	0,51	479	0,37
k	668 315	0,49	505	0,39
za	616 589	0,45	415	0,32
pro	596 021	0,44	368	0,28
ale	576 565	0,42	537	0,41
si	550 073	0,41	682	0,52
by	543 221	0,40	405	0,31

Poslední fází bylo seskupení některých značek stanovených v první fázi tak, aby byly seskupeny značky s podobnými gramatickými nebo syntaktickými vlastnostmi. Ke slučování byl použit program vytvořený v programovacím jazyku Perl, který realizuje hladový algoritmus (Greedy Algorithm) – viz kapitola 2.5.1. Program ze zadaného textu postupně seskupuje slova do jednotlivých tříd. V případě velkého počtu dat je možné stanovit minimální četnost slov pro slučování. Program končí, když jsou všechna slova seskupena do jedné třídy nebo když je dosaženo předurčeného počtu tříd, který lze opět stanovit. Výstupem je binární strom postupu slučování a vzájemná informace vypočítaná na začátku slučování.

K experimentům, ve kterých byl použit hladový algoritmus, bylo připraveno necelých 3 000 vět. Věty byly upraveny podle požadavků v odstavci 3.4 a dále pak čtyřmi níže uvedenými způsoby.

1. Místo všech slov byly značky, které obsahovaly gramatické kategorie příslušné slovnímu druhu (viz tabulka 5.1) – V2637_T.
2. Značky byly jen místo podstatných jmen – V2637_1.
3. Značky byly místo podstatných a přídavných jmen – V2637_12.
4. K příslovcím byla přidána informace o slovním druhu – V2637_6.

Hladový algoritmus byl použit také na 7 000 vět (V7000), které byly upraveny jen podle požadavků uvedených v odstavci 3.4. V tabulce 5.4 jsou uvedeny hodnoty vzájemné

informace pro výše upravené věty. Z tabulky je zřejmé, že v případě nahrazení slov značkami dochází ke ztrátě vzájemné informace.

Tab. 5.4: Vzájemná informace

Věty	Vzájemná informace
V2637_T	2,099
V2637_1	4,678
V2637_12	4,206
V2637_6	6,584
V7000	6,639

Pro každou množinu vět byla na základě počtu různých slov, které se objeví v textu, stanovena minimální četnost slučovaných slov. Počet tříd, kterého má být při slučování dosaženo, byl určen podle toho, zda bylo cílem rozdělit slova do velkého počtu tříd nebo naopak vytvořit malý počet tříd s velkým počtem slov. V tabulce 5.5 jsou uvedeny hodnoty těchto parametrů včetně počtu různých a slučovaných slov ve větách.

Tab. 5.5: Hodnoty vstupních parametrů pro hladový algoritmus včetně počtu různých a slučovaných slov ve větách

Věty	Minimální četnost slučovaných slov	Počet tříd	Počet různých slov	Počet slučovaných slov
V2637_T	3	200	475	371
V2637_1	5	30	6 374	657
V2637_12	5	30	4 477	624
V2637_6	5	100	10 296	860
V7000	10	150	22 837	1 100

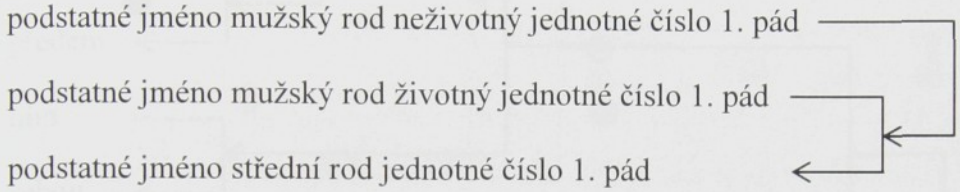
5.1 Seskupování slov ve V2637_1

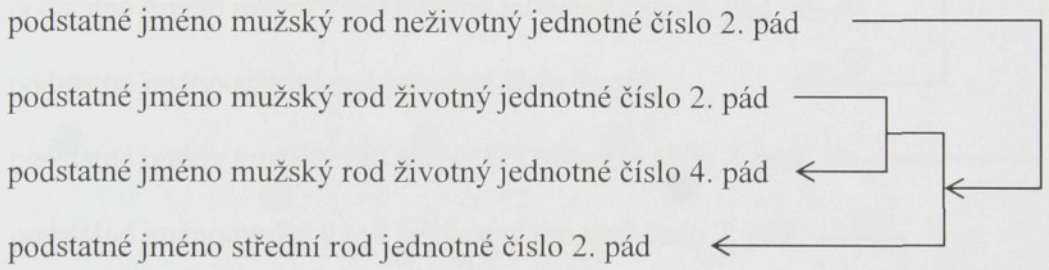
Cílem seskupování slov v textu upraveném tak, že podstatná jména byla nahrazena značkami charakterizujícími jejich gramatické kategorie (V2637_1), bylo najít značky, které je možné seskupit. Značky přiřazené podstatným jménům obsahovaly informaci o rodu, čísle a pádu podstatného jména. Celkový počet vět v textu určeném pro seskupování (V2637_1) byl 2 637, počet slov byl 35 062 a podstatná jména tvořila téměř 21 % všech slov (7 302). Slova, která nebyla seskupena s žádným jiným slovem, jsou uvedena v tabulce 5.6. Do dalších 26 tříd byla rozdělena zbylá slova. Tyto třídy obsahovaly nejméně 4 a nejvíce 98 slov.

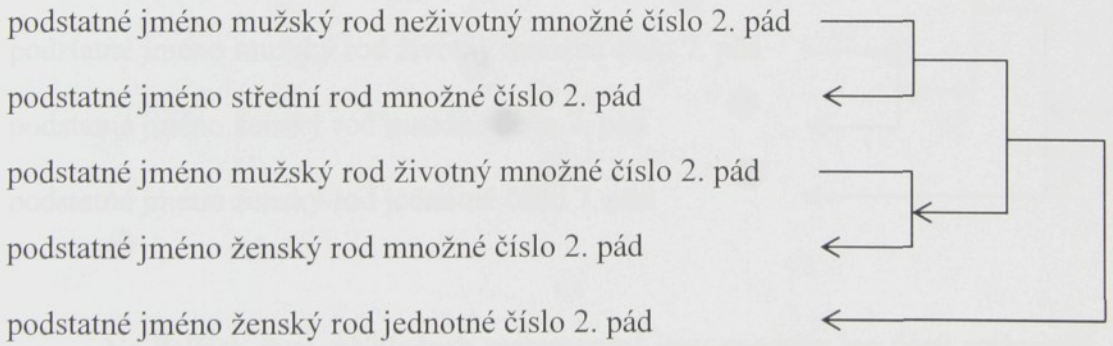
Tab. 5.6: Slova a interpunkční znaménka, která nebyla seskupena v textu V2637_1

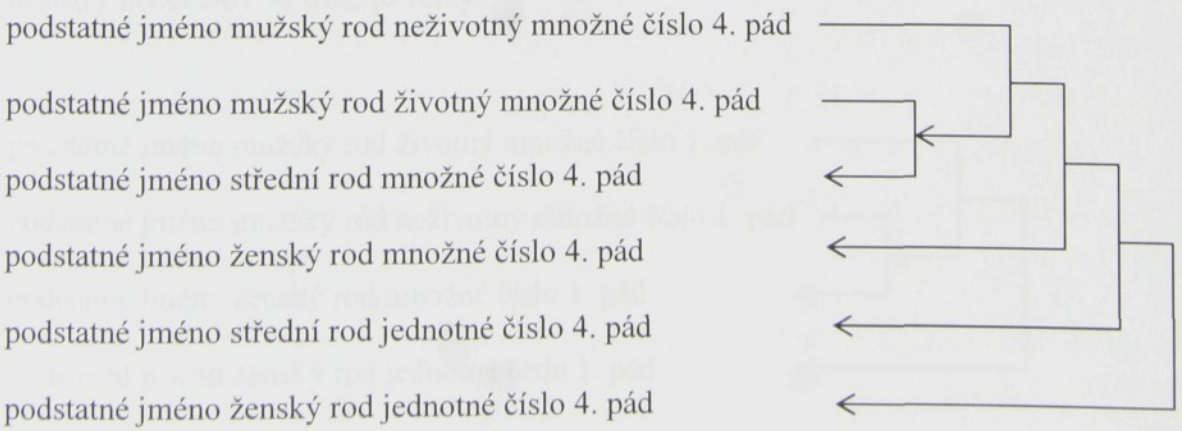
slovní druh	**třída**
interpunkční znaménka	, .
spojka	a
předložka, zvratné zájmeno	se

Příklady postupu seskupování některých podstatných jmen jsou uvedeny níže. Většina podstatných jmen byla seskupena podle pádu a čísla nezávisle na rodu.

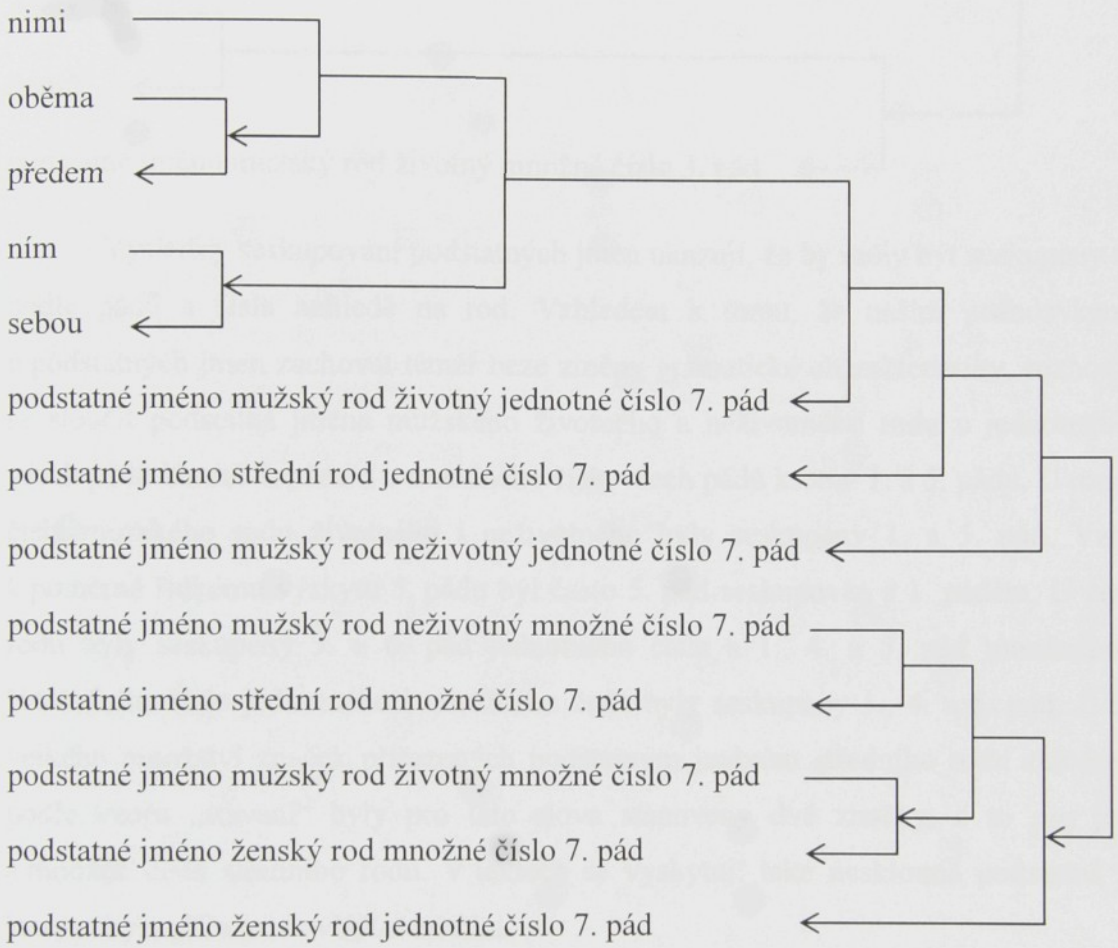




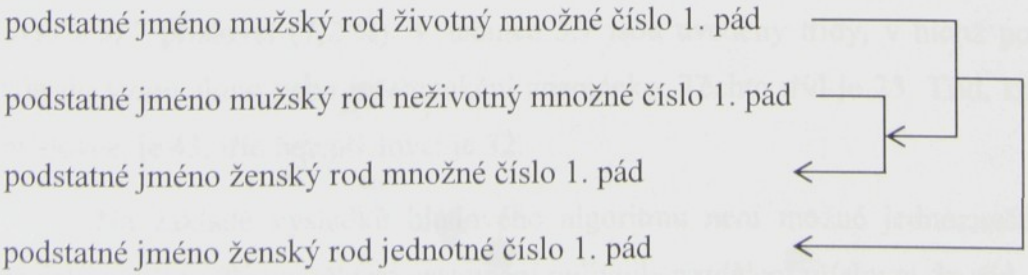


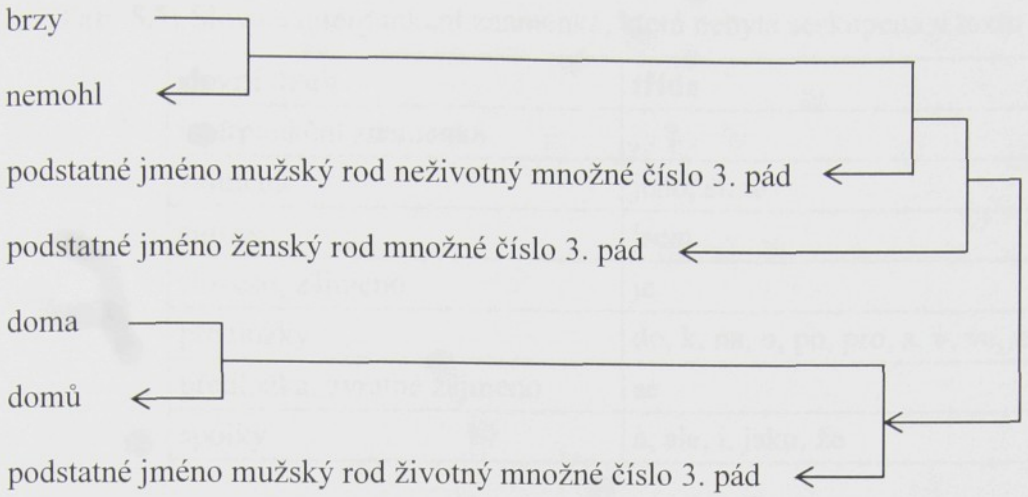


Některá podstatná jména byla seskupena i s jinými slovními druhy. Z níže uvedeného postupu seskupování je zřejmé, že v případě rozdělení do více tříd by bylo množné a jednotné číslo odděleno. K poslednímu slučování totiž dochází mezi množným číslem a jednotným číslem se slovy nimi, oběma, předem, ním, sebou.



Na dalších dvou příkladech seskupování jsou uvedeny jen části seskupení, protože celkový počet slov ve třídě je velký.





Výsledky seskupování podstatných jmen ukazují, že by měly být seskupeny značky podle pádů a čísla nehledě na rod. Vzhledem k tomu, že naším požadavkem bylo u podstatných jmen zachovat téměř beze změny gramatické charakteristiky, rozhodli jsme se sloučit podstatná jména mužského životného a neživotného rodu u jednotného čísla všech pádů kromě 4. pádu a u množného čísla všech pádů kromě 1. a 5. pádu. U množného čísla mužského rodu životného i neživotného byly seskupeny 1. a 5. pád. Vzhledem k poměrně řídkému výskytu 5. pádu byl často 5. pád seskupován s 1. pádem. U ženského rodu byly seskupeny 3. a 6. pád jednotného čísla a 1., 4. a 5. pád množného čísla. U středního rodu jednotného i množného čísla byly seskupeny 1., 4. a 5. pád. Z důvodu velkého množství značek přiřazených podstatným jménům středního rodu skloňovaných podle vzoru „stavení" byly pro tato slova stanoveny dvě značky, a to pro jednotné a množné číslo středního rodu. V textech se vyskytují také nesklonná podstatná jména, kterým byla přiřazena zvláštní značka.

5.2 Seskupování slov ve V2637_6

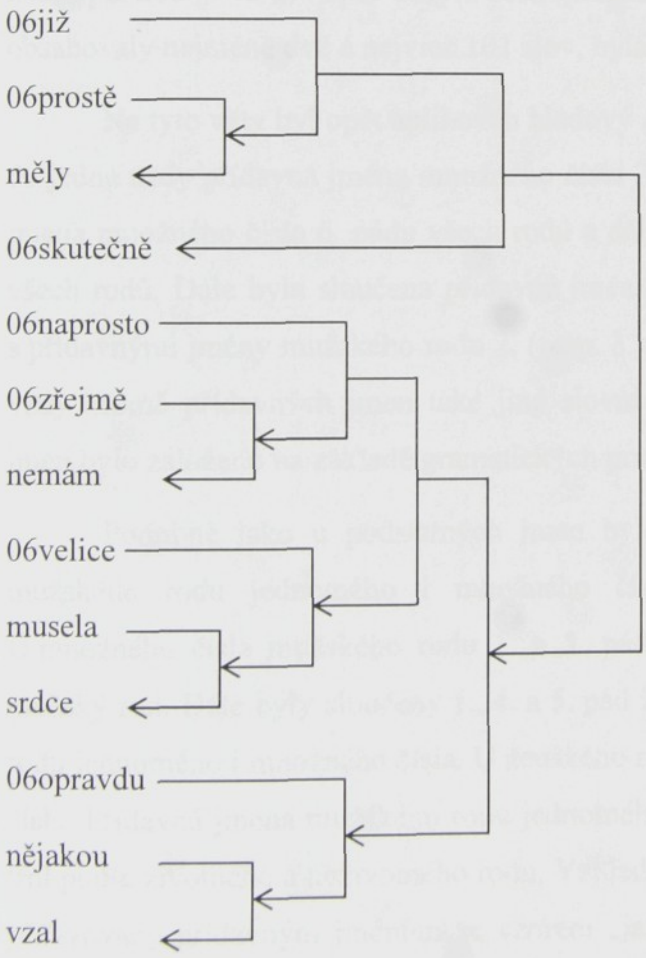
Seskupování slov v textu upraveném tak, že k příslovcím byla přidána informace o slovním druhu (V2637_6), pomohlo definovat značky pro příslovce. Celkový počet vět v textu určeném pro seskupování slov (V2637_6) byl 2 637, počet slov byl 35 062, z toho bylo 2 527 příslovcí (7,2 %). V tabulce 5.7 jsou uvedeny třídy, v nichž po seskupování zůstalo jedno slovo nebo interpunkční znaménko. Těchto tříd je 25. Tříd, kde se objevují příslovce, je 43, tříd bez příslovcí je 32.

Na základě výsledků hladového algoritmu není možné jednoznačně příslovcím přidělit značky. Přesto některá seskupení ovlivnila rozdělení příslovcí do tříd.

Tab. 5.7: Slova a interpunkční znaménka, která nebyla seskupena v textu V2637_6

slovní druh	**třída**
interpunkční znaménka	, . ?
zájmena	jeho, si, to
sloveso	jsem
sloveso, zájmeno	je
předložky	do, k, na, o, po, pro, s, v, ve, z, za
předložka, zvratné zájmeno	se
spojky	a, ale, i, jako, že

V jedné třídě se kromě dalších slov vyskytla příslovce „dlouho“, „poté“, „teď“, „nejprve“, „potom“. Příslovce „ráno“ a „večer“ byla také seskupena s dalšími slovy do jedné třídy a podobně příslovce nebo „hned“ a „ihned“. Příklad postupu seskupování slov, kde většinu tvoří příslovce způsobu, je uveden níže. Na základě těchto výsledků byly pro příslovce času a způsobu vytvořeny zvláštní značky.



Dvojice „kam“ a „proč“ a dvojice „kde“ a „kdy“ byly společně s dalšími slovy seskupeny podle výsledků hladového algoritmu do dvou tříd. Takovým příslovcím byla na základě výsledků hladového algoritmu přiřazena značka „příslovce otázky“. Další značky (příslovce místa, příslovce počtu, 2. a 3. stupeň a ostatní příslovce) již z hladového algoritmu nejsou úplně zřejmé a byly stanoveny na základě gramatických a syntaktických pravidel.

5.3 Seskupování slov ve V2637_12

Cílem seskupování ve větách upravených tak, že místo podstatných a přídavných jmen byly značky charakterizující jejich gramatické kategorie, bylo seskupení některých značek pro přídavná jména. Značky přiřazené podstatným a přídavným jménům obsahovaly informaci o rodu, čísle a pádu podstatného nebo přídavného jména. Celkový počet vět byl stejný jako v předchozích dvou případech. Podstatná jména tvořila téměř 21 % a přídavná jména asi 8 % (2 786) z celkového počtu 35 062 slov. Slova „a“, „se“ a interpunkce „.“ a „,“ opět nebyla seskupena s žádnými slovy. Do dalších 26 tříd, které obsahovaly nejméně dvě a nejvíce 101 slov, byla rozdělena zbylá slova.

Na tyto věty byl opět aplikován hladový algoritmus a na jeho základě byla sloučena do jedné třídy přídavná jména množného čísla 2. pádu všech rodů, do další třídy přídavná jména množného čísla 6. pádu všech rodů a dále přídavná jména množného čísla 7. pádu všech rodů. Dále byla sloučena přídavná jména středního rodu 2. (resp. 3., 6. a 7.) pádu s přídavnými jmény mužského rodu 2. (resp. 3., 6. a 7.) pádu. V seskupených třídách byly vždy kromě přídavných jmen také jiné slovní druhy, takže další seskupení přídavných jmen bylo založeno na základě gramatických pravidel.

Podobně jako u podstatných jmen byla sloučena přídavná jména 1. a 5. pádu mužského rodu jednotného i množného čísla a ženského rodu jednotného čísla. U množného čísla mužského rodu 1. a 5. pádu je navíc rozdělen životný a neživotný mužský rod. Dále byly sloučeny 1., 4. a 5. pád ženského rodu množného čísla a středního rodu jednotného i množného čísla. U ženského rodu byly seskupeny 3. a 6. pád jednotného čísla. Přídavná jména mužského rodu jednotného čísla 4. pádu zůstala rozdělena do dvou tříd podle životného a neživotného rodu. Vzhledem k velkému množství značek, které jsou přiřazovány přídavným jménům se vzorem „jarní“, byla vytvořena speciální značka pro tato přídavná jména.

5.4 Seskupování slov ve V2637_T

Při seskupování slov ve 2 637 větách upravených tak, že místo slov byly značky odpovídající gramatickým kategoriím jednotlivých slov, byla minimální četnost slov nastavena na 3 z důvodu malého počtu různých slov. Počet tříd, kterého bylo při slučování dosaženo, byl stanoven na 100. Výsledkem slučování bylo 43 tříd, které obsahovaly jednu značku, a 57 tříd, které obsahovaly od 2 do 17 značek odpovídajících gramatickým kategoriím slov. K první skupině tříd patřily např. příslovce času, příslovce způsobu nebo předložka pojící se s 2., 3., 4., 6. nebo 7. pádem. Podstatná a přídavná jména byla seskupena podobně jako v kapitole 5.1 a 5.3. Často byly seskupeny zájmena a číslovky nebo zájmena a přídavná jména nebo zájmena, přídavná jména a číslovky. Slovesa byla částečně rozdělena podle času. Jedna třída sloves obsahovala příčestí trpné. Příslovce, předložky a spojky byly ve třídách většinou samostatně.

Jedna z podmínek, které jsme pro slučování stanovili, byla zachování slovního druhu. Tato zásada byla porušena v jednom případě, a to pro sloučení přídavných jmen, zájmen a číslovek, které mají stejné syntaktické vlastnosti. Šlo o zájmena neurčitá a záporná (např. nějaký, žádný) a číslovky řadové (např. první, druhý). Ve stejné třídě s těmito slovními druhy byla podle hladového algoritmu také zájmena přivlastňovací. Těm byly ponechány buď samostatné značky rozdělené podle rodu, pádu a čísla, nebo byly seskupeny nezávisle na rodu, podobně jako přídavná jména.

5.5 Seskupování slov ve V7000

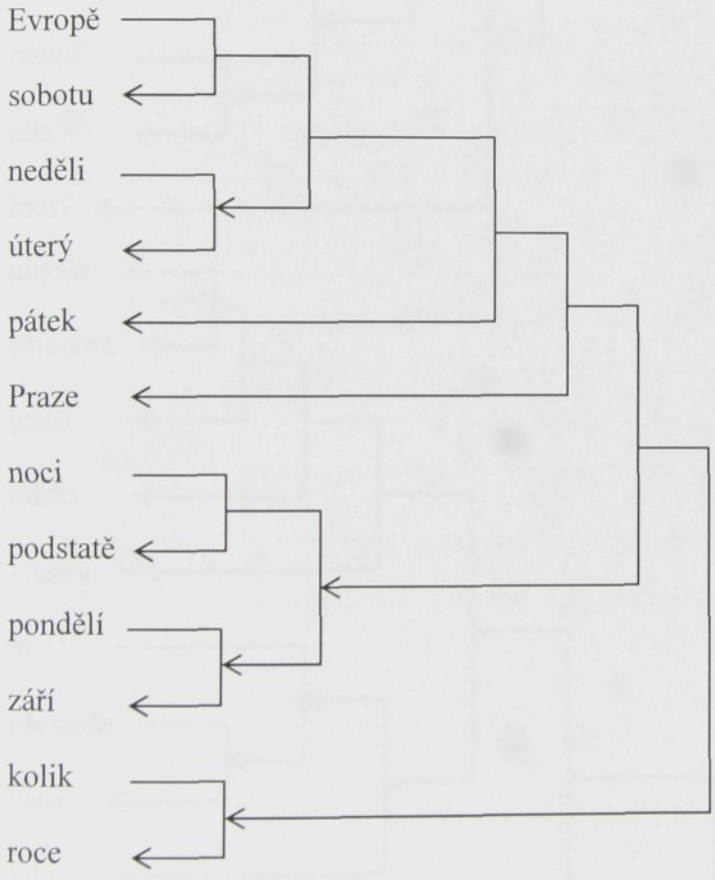
Hladový algoritmus byl použit i pro seskupení slov v 7 000 větách. Při seskupování se potvrdily výsledky ze statistického přístupu, kdy nejfrekventovanější slova zůstala v samostatných třídách. Při zpracování vět byla minimální četnost slov nastavena na 10 z důvodu velkého počtu různých slov ve V7000. Slučování bylo zastaveno při 150 třídách. Samostatné značky byly přiřazeny slovům a interpunkčním znaménkům uvedeným v tabulce 5.8.

Tab. 5.8: Slova a interpunkční znaménka, která nebyla seskupena v textu V7000

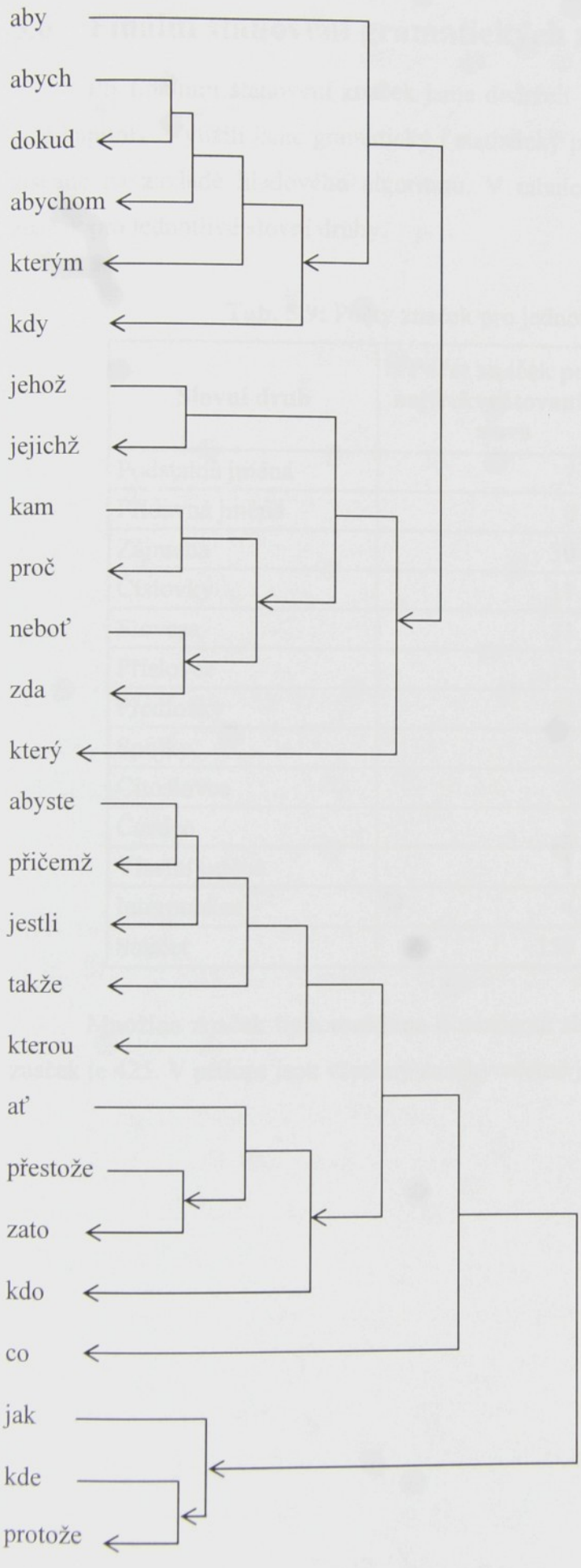
slovní druh	skupina
interpunkční znaménka	, . ?
podstatná jména	tisíc

zájmena	si, to
slovesa	by, byl, byla, jsem, jsme, jsou
sloveso, zájmeno	je
příslovce	tak, už
předložky	do, k, na, o, od, po, pro, s, v, ve, z, za
předložka, zvratné zájmeno	se
spojky	a, ale, až, i, jako, nebo, než, že
částice	jen

Poměrně přesně byly rozděleny do tříd číslovky, některá slovesa, zájmena a příslovce. Na níže uvedeném postupu slučování je zobrazeno slučování slov následujících po předložce „v".



Další dva příklady seskupování jsou třídy slov – spojek, příslovcí nebo zájmen, které většinou následují po interpunkci.


aby
abych
dokud
abychom
kterým
kdy
jehož
jejichž
kam
proč
neboť
zda
který
abyste
přičemž
jestli
takže
kterou
ať
přestože
zato
kdo
co
jak
kde
protože

5.6 Finální stanovení gramatických značek

Při finálním stanovení značek jsme dodrželi všechny tři zásady uvedené v úvodu páté kapitoly. Využili jsme gramatický i statistický přístup a vzali jsme v úvahu výsledky získané na základě hladového algoritmu. V tabulce 5.9 jsou uvedeny počty finálních značek pro jednotlivé slovní druhy.

Tab. 5.9: Počty značek pro jednotlivé slovní druhy

Slovní druh	Počet značek pro nejfrekventovanější slova	Počet dalších značek
Podstatná jména	7	39
Přídavná jména	0	24
Zájmena	30	97
Číslovky	11	55
Slovesa	31	22
Příslovce	15	8
Předložky	34	6
Spojky	15	4
Citoslovce	0	1
Částice	2	1
Vlastní jména	1	15
Interpunkce	4	1
Součet	150	273

Množina značek byla rozšířena o neznámá slova a o začátek věty. Celkový počet značek je 425. V příloze jsou všechny značky včetně jejich popisu uvedeny.

Kapitola 6

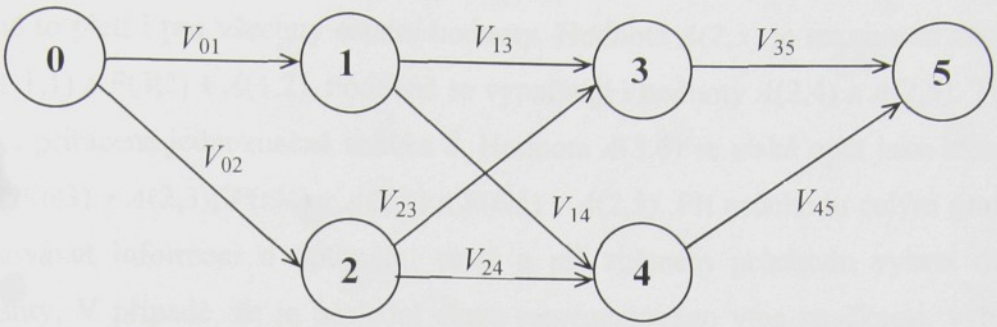
Automatické značkování

Pro automatické značkování vět jsme vytvořili v programovacím jazyku Perl stochastické značkovače, které pro nalezení nejlepší posloupnosti značek pro danou posloupnost slov využívají dynamické programování. Dynamické programování je založeno na Bellmanově principu optimality, který říká, že každá podstrategie optimální strategie je optimální.

Všechny příslušné značky, které jsou jednotlivým slovům věty přiřazeny, si lze představit jako uzly orientovaného ohodnoceného grafu. Orientovaný ohodnocený graf G je čtveřice (V, E, c, a), kde V je konečná množina uzlů, E je konečná množina hran a c je zobrazení, které přiřazuje každé hraně e uspořádanou dvojici uzlů $(x, y) \in V^2$ a a je ohodnocení grafu. Uzel x se nazývá počáteční vrchol hrany e a uzel y se nazývá koncový uzel hrany e. V práci jsou používány prosté grafy, to znamená, že ve stejném směru může vést nejvýše jedna hrana. Cílem automatického značkování je najít optimální orientovanou cestu v grafu. Orientovaná cesta v grafu je konečná posloupnost $v_1, e_1, v_2, e_2, ..., v_{n+1}$, kde v_i jsou uzly grafu G a e_i jeho hrany, přičemž platí, že v_i je počáteční vrchol hrany e_i a v_{i+1} je koncový vrchol hrany e_i pro všechna $i \in \langle 1, n \rangle$ a že se v posloupnosti žádný uzel neopakuje. Pro grafickou interpretaci věty je používán orientovaný ohodnocený graf, který má právě jeden uzel, který není koncovým uzlem žádné hrany a který se nazývá počátečním uzlem.

Existuje řada způsobů, jak graf popsat. Pro interpretaci věty je nejvhodnější použít jeho nákres. Uzly grafu se v nákresu zobrazují jako body, hrany grafu jako jejich spojnice, u orientovaného grafu se šipkou. Jednotlivé uzly jsou označeny čísly a u ohodnoceného orientovaného grafu je u každé hrany uvedeno ohodnocení (viz obr. 6.1).

Všechny dále uvedené metody značkování byly použity pro tvorbu značkovačů (taggerů), pomocí nichž bylo provedeno automatické značkování.



Obr. 6.1: Schéma ohodnoceného orientovaného grafu

6.1 Automatické značkování s ohodnocením hran grafu

Metoda automatického značkování s ohodnocením hran grafu využívá k ohodnocení hran bigramovou matici, a to nevyhlazenou nebo vyhlazenou lineární interpolací. Každé hraně, která spojuje dva uzly, je jako ohodnocení přiřazena pravděpodobnost výskytu odpovídající značky za podmínky, že ji předcházela určitá značka. Tato podmíněná pravděpodobnost se načítá z matice bigramů. Protože podmíněné pravděpodobnosti jsou velmi malá čísla, je zvykem uvádět v matici jejich logaritmy. Pro veškeré výpočty byl použit přirozený logaritmus z důvodu vhodnějšího měřítka zejména pro porovnání malých hodnot pravděpodobnosti. Pro průchod tímto grafem lze využít upravený Bellmanův vzorec, který každému uzlu přiřazuje maximum ze součtů přirozených logaritmů podmíněných pravděpodobností příslušného uzlu a hodnoty předchozích uzlů.

$$A(i,j) = \max_{\forall k}\{P(j \mid k) + A(i-1,k)\} \tag{6.1}$$

kde i je index, který určuje pořadí slova ve větě,

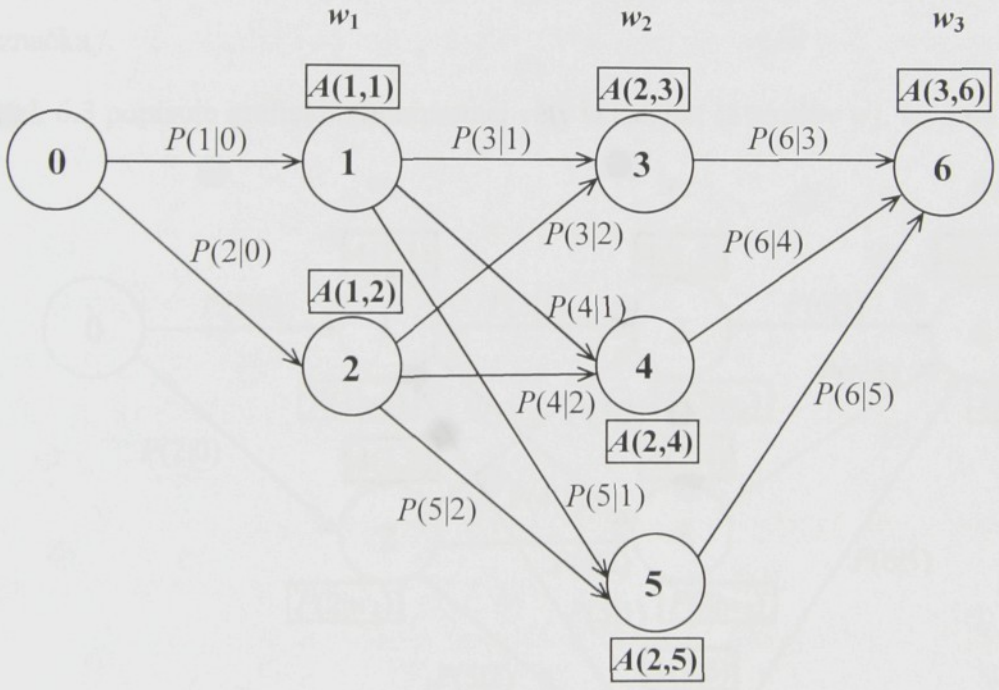
j je značka příslušná danému slovu,

k je značka příslušná předchozímu slovu,

$P(j \mid k)$ je přirozený logaritmus pravděpodobnosti značky j za podmínky, že ji předcházela značka k.

Na obrázku 6.2 je grafické znázornění daného postupu. Začátek věty je označen uzlem 0. Čísla v jednotlivých uzlech určují značky, které jsou přiřazeny příslušnému slovu. Prvnímu slovu w_1 odpovídají dvě značky (1 a 2). Přirozený logaritmus pravděpodobnosti

výskytu značky 1 na začátku věty je $P(1|0)$, značky 2 $P(2|0)$. Hodnoty $A(1,1)$ a $A(1,2)$ jsou shodné s hodnotami $P(1|0)$ a $P(2|0)$. Druhému slovu w_2 odpovídají tři značky (3, 4 a 5). $P(3|1)$ je přirozený logaritmus pravděpodobnosti s jakou se vyskytuje značka 3 za značkou 1. Podobně to platí i pro všechny ostatní hodnoty. Hodnota $A(2,3)$ je maximum ze součtů $P(3|1) + A(1,1)$ a $P(3|2) + A(1,2)$. Podobně se vypočítají i hodnoty $A(2,4)$ a $A(2,5)$. Třetímu slovu w_3 je přiřazena jednoznačně značka 6. Hodnota $A(3,6)$ se získá opět jako maximum ze součtů $P(6|3) + A(2,3)$, $P(6|4) + A(2,4)$ a $P(6|5) + A(2,5)$. Při průchodu celým grafem je třeba uchovávat informaci o optimální cestě a při zpětném průchodu vybrat nejlepší reprezentanty. V případě, že je poslední slovo reprezentováno více značkami, vybere se maximální hodnota a optimální cesta se určí opět zpětným průchodem. Může dojít k situaci, kdy jsou hodnoty A pro jedno slovo stejné, potom je vybrána ta značka, která je v pořadí první.



Obr. 6.2: Grafické znázornění použití dynamického programování pro automatické značkování vět s ohodnocením hran

6.2 Automatické značkování s ohodnocením hran a uzlů

Metoda automatického značkování s ohodnocením hran a uzlů využívá při dynamickém programování kromě podmíněných pravděpodobností z vyhlazené bigramové matice také pravděpodobností, s jakými jsou daným slovům přiřazeny dané značky. Tato pravděpodobnost je vyhlazena metodou Add-One Smoothing. U grafické interpretace věty

w_1 přiřazena značka 1 resp. 2. Hodnoty $A(1,1)$ a $A(1,2)$ jsou rovny součtu $P(1|0) + P(1|w_1)$ a $P(2|0) + P(2|w_1)$. Druhému slovu w_2 mohou být přiřazeny tři značky (3, 4 a 5). Hodnota $A(2,3)$ je maximum ze součtů $P(3|1) + P(3|w_2) + A(1,1)$ a $P(3|2) + P(3|w_2) + A(1,2)$. Podobně se vypočítají i hodnoty $A(2,4)$ a $A(2,5)$. Třetímu slovu w_3 je přiřazena jednoznačně značka 6. Hodnota $A(3,6)$ se získá opět jako maximum ze součtů $P(6|3) + P(6|w_3) + A(2,3)$, $P(6|4) + P(6|w_3) + A(2,4)$ a $P(6|5) + P(6|w_3) + A(2,5)$.

6.3 Automatické značkování s ohodnocením hran a uzlů a s četností slov

U automatického značkování s ohodnocením hran a uzlů a s četností slov se využívá kromě podmíněných pravděpodobností z vyhlazené bigramové matice a pravděpodobností, s jakými budou daným slovům přiřazeny dané značky, ještě pravděpodobností, s jakými se slova vyskytnou ve slovníku. U grafické interpretace je pak přidána ke každému uzlu ještě informace o přirozeném logaritmu této pravděpodobnosti – k uzlům 1 a 2 $P(w_1)$, k uzlům 3, 4, a 5 $P(w_2)$ a k uzlu 6 $P(w_3)$. Upravený Bellmanův vzorec pak vypadá takto:

$$A(i,j) = \max_{\forall k}\{P(j \mid k) + P(j \mid w_i) + P(w_i) + A(i-1,k)\} \tag{6.3}$$

kde i je index, který určuje pořadí slova ve větě,

j je značka příslušná danému slovu,

k je značka příslušná předchozímu slovu,

w_i je dané slovo,

$P(j \mid k)$ je přirozený logaritmus pravděpodobnosti značky j za podmínky, že ji předcházela značka k,

$P(j \mid w_i)$ je přirozený logaritmus pravděpodobnosti, s jakou slovu w_i bude přiřazena značka j,

$P(w_i)$ je přirozený logaritmus pravděpodobnosti výskytu slova ve slovníku.

Podle úspěšnosti automatického značkování textu (viz kapitola 7) byl proveden výběr nejlepšího značkovače pro experimenty s rozpoznávačem spojité řeči.

Kapitola 7

Značkování korpusu

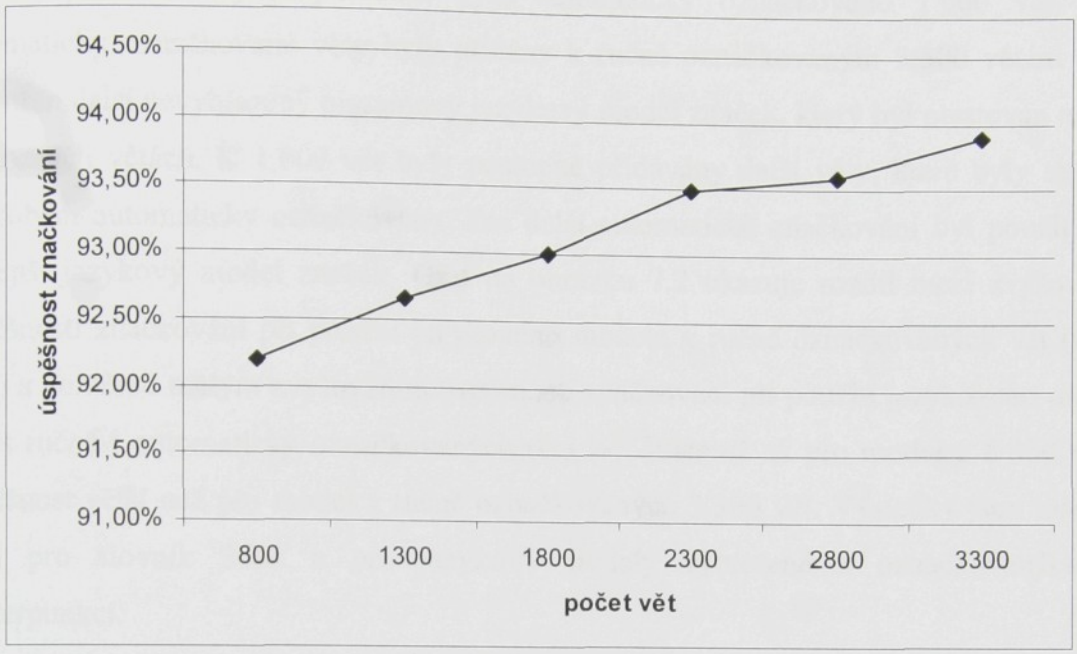
Pro experimenty bylo ručně nebo poloautomaticky označkováno 3 800 vět, z toho 500 vět bylo použito pro testování. Pro ruční značkování vět byl vytvořen program, který pro každé slovo ve větě nabídl ze slovníku příslušné gramatické značky. U slov, která ve slovníku chyběla, byla přiřazena značka ručně. Z 800 manuálně označkovaných vět byla vytvořena nevyhlazená bigramová matice, jejíž prvky byly přirozené logaritmy podmíněných pravděpodobností. Tato matice byla použita pro poloautomatické značkování, při kterém program označil u každého slova nejpravděpodobnější značku. U nesprávně určených značek byla provedena ruční oprava. U slov, která ve slovníku chyběla, byla přiřazena značka ručně. Na základě takto ručně označkovaných vět byl vytvořen bigramový jazykový model značek.

Všechny experimenty byly prováděny se třemi různě velkými slovníky a jazykové modely byly vytvářeny z vět s interpunkcí i bez interpunkce. Pro automatické značkování byly vytvořeny tři různé stochastické značkovače založené na Stochastic Parts Program (viz odstavec 4.1.2).

7.1 TaggerB

Stochastický značkovač s nevyhlazeným jazykovým modelem značek (TaggerB) používá k ohodnocení hran grafu prvky nevyhlazené bigramové matice. Nevyhlazený bigramový jazykový model značek byl vytvořen nejprve z 800 vět. K těmto větám byly postupně přidávány další, a tak byly vytvořeny další jazykové modely značek z 1 300, 1 800, 2 300, 2 800 a 3 300 ručně označkovaných vět. Tyto bigramové modely značek byly postupně použity k automatickému označkování 500 vět, které sloužily jen pro testování. Z grafu uvedeného na obrázku 7.1 je patrné postupné zlepšování k téměř 94% úspěšnosti

značkování. V tomto případě byl použit slovník 300k (313 217 slov – viz kapitola 9.1) a jazykové modely značek byly vytvářeny z vět s interpunkcí.



Obr. 7.1: Úspěšnost značkování u ručně označkovaných vět

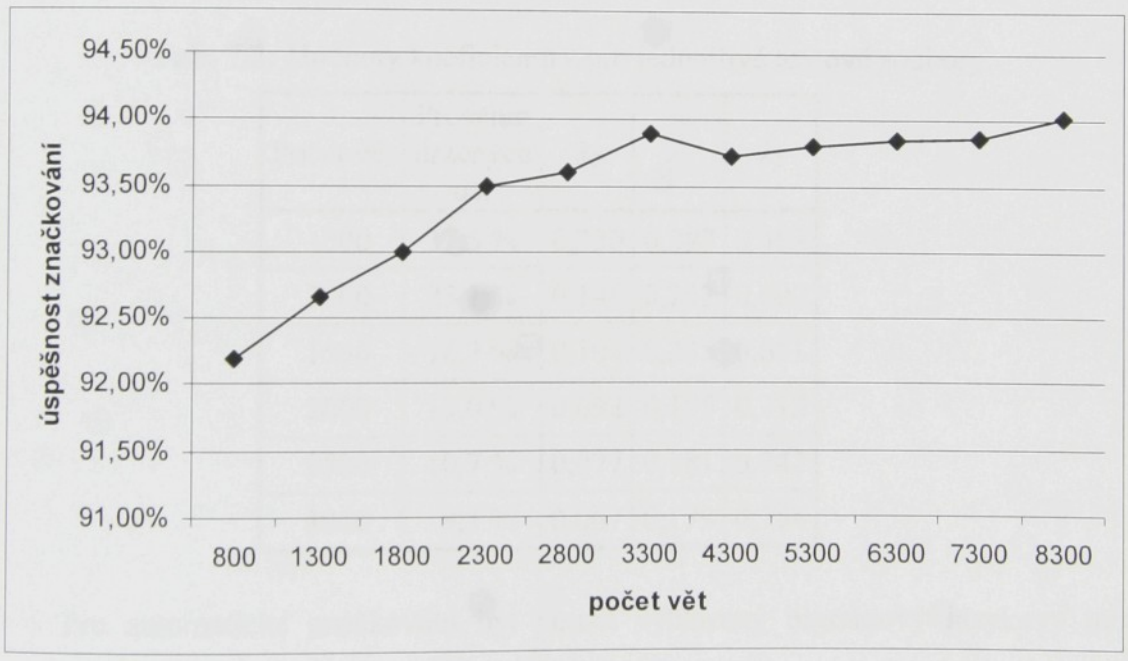
Z těchto modelů byl vybrán nejlepší nevyhlazený bigramový jazykový model značek, který byl použit pro automatické označkování dalších vět.

V případě, že se slovo ve slovníku nenachází, je mu automaticky přiřazena značka pro neznámá slova (13nez). U nevyhlazeného jazykového modelu značek byla stanovena určitá hodnota podmíněné pravděpodobnosti pro dvojice, které se v korpusu neobjevily. Tato hodnota byla určena jako dolní hranice z podmíněných pravděpodobností bigramové matice a přirozený logaritmus této pravděpodobnosti je –10.

Zvláštním způsobem je provedeno označkování prvního slova ve větě. První slovo ve větě je automaticky převedeno na malá písmena. Až poté je vyhledáno ve slovníku. Pokud je ve slovníku nalezeno, je mu přiřazena příslušná značka (značky). Pokud ve slovníku není, je první písmeno převedeno na velké. Takto upravené slovo je opět vyhledáváno ve slovníku. V případě, že je nalezeno, je mu přiřazena příslušná značka (značky). Není-li nalezeno, je mu přiřazena značka pro neznámé slovo (13nez). Tento způsob částečně řeší problém slov, která jsou na začátku věty a mohou se vyskytnout jak s velkým písmenem jako vlastní jména, tak i s malým písmenem. Protože se na prvním místě věty ve většině případů vlastní jména neobjevují, byl použit právě tento způsob. Nejsou však vyřešeny případy, kdy se na začátku věty vyskytne zkratka, ve které jsou

všechna písmena velká a případy, kdy jde o vlastní jméno na začátku věty, které se může vyskytnout i s malým písmenem (např. Česká a česká).

Výše uvedeným způsobem bylo automaticky označkováno 1 000 vět. Tyto automaticky označkované věty byly přidány k ručně označkovaným 3 300 větám a byl vytvořen další nevyhlazený bigramový jazykový model značek, který byl otestován na 500 testovacích větách. K 1 000 vět byly postupně přidávány další věty, které byly stejným způsobem automaticky označkovány. Pro další automatické značkování byl použit vždy nejlepší jazykový model značek. Graf na obrázku 7.2 ukazuje rozdíl mezi zvyšováním úspěšnosti značkování při použití jazykového modelu z ručně označkovaných vět (téměř 2 %) a poměrně malým zvyšováním úspěšnosti značkování při použití jazykového modelu z vět ručně i automaticky označkovaných (0,3 %). Přičemž až pro model z 8 300 vět je úspěšnost větší než pro model z ručně označkovaných 3 300 vět. Výsledky jsou uvedeny opět pro slovník 300k a pro jazykové modely vytvořené z označkovaných vět s interpunkcí.



Obr. 7.2: Úspěšnost stochastického značkovače TaggerB

V tabulce 7.1 jsou uvedena procenta nenulových bigramů ve větách z korpusu.

Tab. 7.1: Procenta nenulových bigramů

Počet vět	800	1300	1800	2300	2800	3300	4300	5300	6300	7300	8300
Nenulové bigramy [%]	2,61	3,45	4,26	5,06	5,74	6,23	7,28	8,13	8,94	9,39	9,86

7.2 TaggerSB

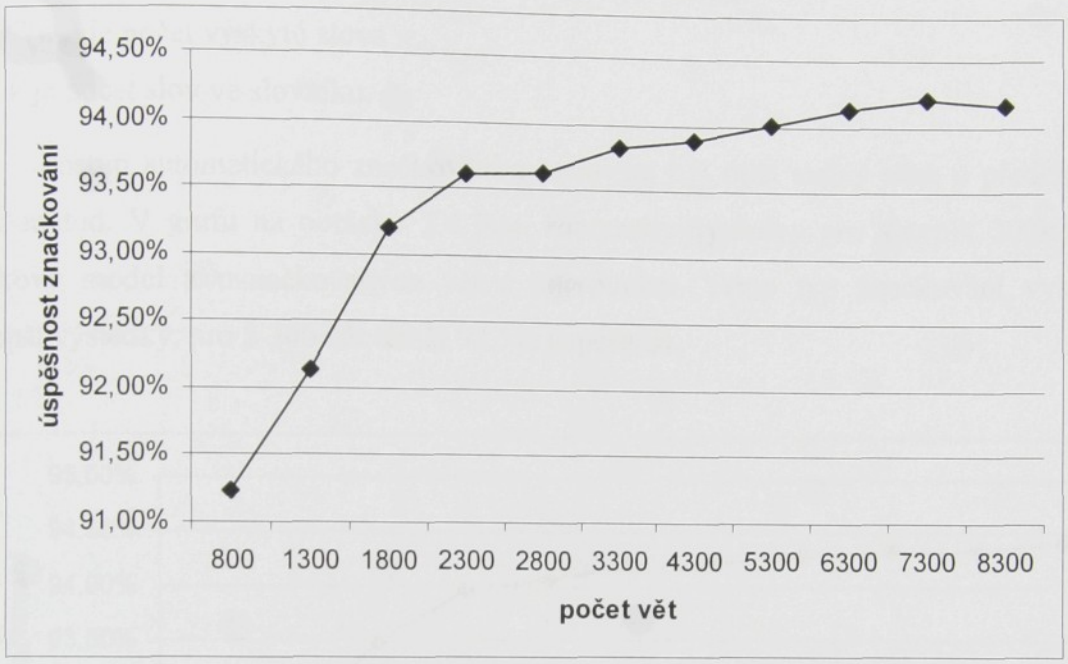
Stochastický značkovač s vyhlazeným jazykovým modelem značek (TaggerSB) používá pro ohodnocení hran v grafu prvky vyhlazené bigramové matice. Pro vyhlazování modelu byla použita metoda lineární interpolace (viz kapitola 2.4.4). U tohoto typu vyhlazování je potřeba držet nějaká data stranou (held-out data) pro stanovení koeficientů λ (viz kapitola 2.4.5). Z trénovací množiny bylo pro tento účel vyřazeno 300 ručně označkovaných vět. Nejdříve byl vytvořen vyhlazený jazykový model značek z 500 ručně označkovaných vět a z 300 odložených. Postupně byly k 500 větám přidávány další ručně označkované věty, přičemž odložená data byla stále stejná. V tabulce 7.2 jsou uvedeny hodnoty koeficientů λ pro jednotlivé textové soubory a údaj kolik procent tvoří odložená data. V případě, že odložená data tvoří kolem 10 %, koeficient u bigramů vychází přibližně 0,75, koeficient u unigramů $\lambda_1 = 0{,}18$ a $\lambda_0 = 0{,}07$. Rozložení koeficientů λ v případě, že procento odložených dat je větší, je rovnoměrnější. Koeficient u bigramů nemá tak velkou váhu (0,48) a koeficienty λ_1 a λ_0 mají hodnotu přibližně 0,25.

Tab. 7.2: Hodnoty koeficientů λ pro jednotlivé textové soubory

Počet vět	Procento držených dat	λ_0	λ_1	λ_2
500	37,5 %	0,220	0,297	0,483
1000	23,1 %	0,146	0,252	0,602
1500	16,7 %	0,104	0,235	0,661
2000	13,0 %	0,092	0,195	0,713
2500	10,7 %	0,077	0,181	0,742
3000	9,1 %	0,067	0,179	0,754

Pro automatické značkování byl použit vyhlazený bigramový jazykový model značek. Problematika prvních slov ve větě je řešena stejným způsobem jako u nevyhlazeného jazykového modelu značek. Stejným způsobem bylo provedeno postupné automatické značkování vět a testování různých jazykových modelů. V grafu na obrázku 7.3 jsou zobrazeny výsledky pro slovník 300k a pro jazykové modely značek vytvořené z vět s interpunkcí. Opět je zřejmý prudší nárůst úspěšnosti značkování pro ručně označkované věty (3 %) než pro věty označkované automaticky a přidané k ručně označkovaným (0,4 %). V tomto případě došlo k mírnému poklesu až po přidání 5 000

automaticky označkovaných vět k 3 300 ručně označkovaným větám. Přesto je úspěšnost automatického značkování při použití vyhlazeného bigramového jazykového modelu značek z 8 300 vět vyšší než úspěšnost automatického značkování s použitím vyhlazeného jazykového modelu značek z 3 300 ručně označkovaných vět.



Obr. 7.3: Úspěšnost stochastického značkovače TaggerSB

7.3 TaggerSBP

Stochastický značkovač TaggerSBP používá k ohodnocení hran grafu prvky vyhlazené bigramové matice a k ohodnocení uzlů pravděpodobnost výskytu daného slova v dané třídě $p(c_n \mid w_n)$. Označkované věty, ze kterých je tato pravděpodobnost vypočítána, jsou uloženy v souboru ve vertikálním formátu, což znamená, že každé slovo věty je na jednom řádku. Kromě slova je na řádku odpovídající značka, která je od slova oddělena tabulátorem. Pro výpočet podmíněné pravděpodobnosti $p(c_n \mid w_n)$ jsou všechna první slova ve větě převáděna na malá písmena. V případě, že se slovo w_n v trénovacích datech neobjeví, ale ve slovníku obsaženo je, může být slovu přiřazena každá značka, která je danému slovu přiřazena ve slovníku, se stejnou pravděpodobností. Tato hodnota však na celkové vyhledání optimální cesty nemá vliv. V tomto případě je možné pravděpodobnost $p(c_n \mid w_n)$ ignorovat. Pro vyhlazování bigramového jazykového modelu značek $p(c_n \mid c_{n-1})$ je opět použita metoda lineární interpolace. Podmíněná pravděpodobnost $p(c_n \mid w_n)$ je

vyhlazena pomocí nejjednodušší vyhlazovací techniky Add-One Smoothing a počítá se podle vzorce:

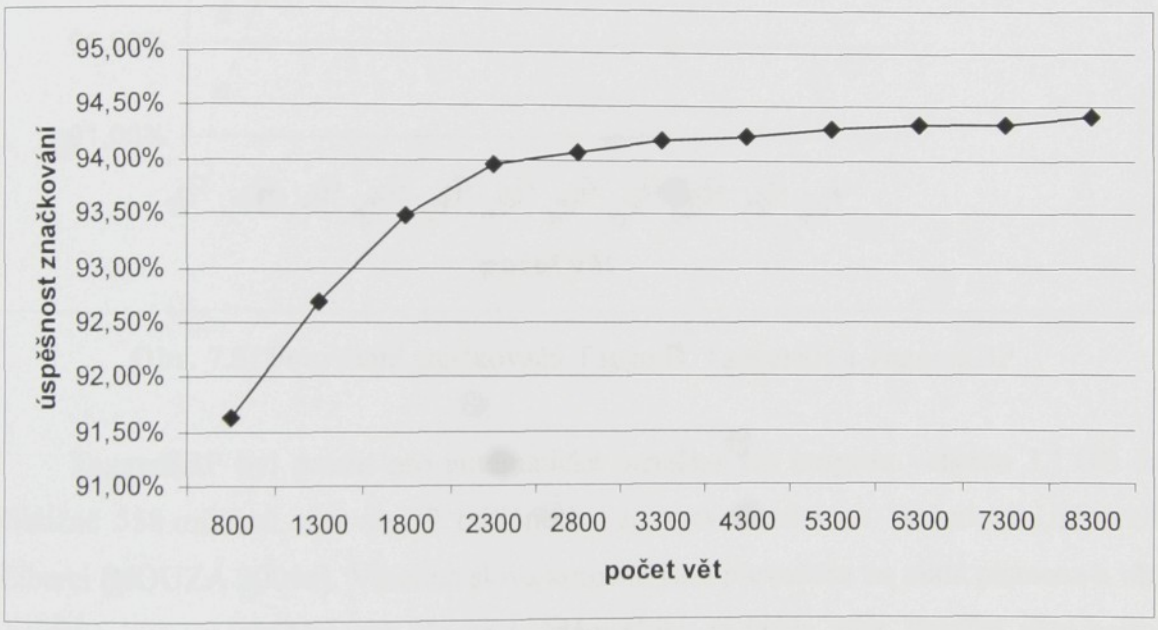
$$p(c_n|w_n) = \frac{C(w_n, c_n)+1}{C(w_n)+V} \tag{7.1}$$

kde $C(w_n, c_n)$ je počet výskytů dvojice slovo w_n – značka c_n,

$C(w_n)$ je počet výskytů slova w_n,

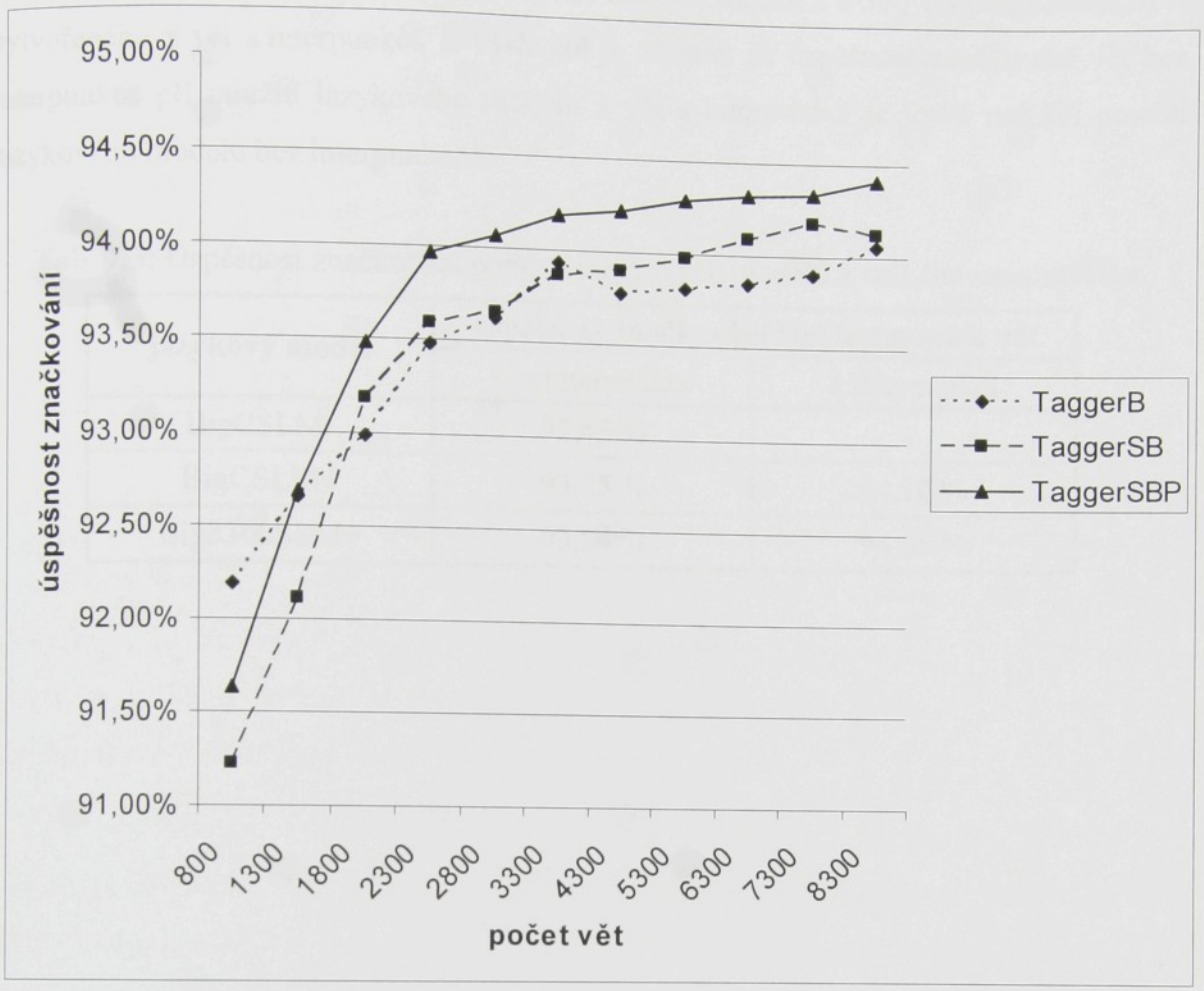
V je počet slov ve slovníku.

Postup automatického značkování a testování byl opět stejný jako u předchozích dvou metod. V grafu na obrázku 7.4 jsou zobrazeny výsledky pro slovník 300k a pro jazykový model z označkovaných vět s interpunkcí. Tento typ značkování vykazuje nejlepší výsledky, pro 8 300 vět téměř 94,5% úspěšnost.



Obr. 7.4: Úspěšnost stochastického značkovače TaggerSBP

Graf na obrázku 7.5 porovnává všechny tři způsoby automatického značkování. Z grafu je zřejmé, že TaggerB s nevyhlazeným bigramovým jazykovým modelem značek sestaveným z malého počtu vět vykazuje lepší výsledky než TaggerSB a TaggerSBP s vyhlazenými bigramovými jazykovými modely značek sestavenými ze stejného počtu vět. Úspěšnost značkovačů TaggerSB a TaggerSBP však po přidání dalších vět rychle stoupá. Z uvedených výsledků lze usuzovat, že všechny tři značkovače vykazují znaky učícího se systému.



Obr. 7.5: Porovnání značkovačů TaggerB, TaggerSB a TaggerSBP

TaggerSBP byl použit pro automatické označkování korpusu velkého 3,1 GB dat (přibližně 558 miliónů slov ve 29 miliónech vět) vytvořeného na Technické Univerzitě v Liberci [NOUZA 2004a]. Všechna slova korpusu jsou převedena na malá písmena a věty v korpusu jsou uspořádány tak, že na každém řádku je jedna věta. Použité věty nejsou manuálně kontrolovány, takže se v nich objevují nespisovná slova, čísla, různé zkratky apod. V textu je přibližně 4,5 % slov, kterým je přiřazena značka pro neznámé slovo. Z takto označkovaných vět a z 8 300 označkovaných vět byly vytvořeny další dva vyhlazené bigramové jazykové modely značek – z vět s interpunkcí (BigCSLM+) a z vět, kde byla interpunkce vynechána (BigCSLM–). Oba modely byly otestovány na testovacích datech. Model vytvořený z vět s interpunkcí byl testován na testovacích datech s interpunkcí i bez interpunkce.

V tabulce 7.3 je uvedena úspěšnost značkování testovacích dat značkovačem TaggerSBP při použití obou jazykových modelů v porovnání s úspěšností značkování při

použití vyhlazeného bigramového jazykového modelu značek z 8 300 vět (Big8300SLM+) vytvořeného z vět s interpunkcí. Z výsledků je zřejmé, že úspěšnost značkování vět bez interpunkce při použití jazykového modelu z vět s interpunkcí je vyšší než při použití jazykového modelu bez interpunkce.

Tab. 7.3: Úspěšnost značkování pomocí jazykových modelů z velkého množství dat

jazykový model	**úspěšnost značkování 500 testovacích vět**	
	bez interpunkce	s interpunkcí
BigCSLM-	92,45 %	–
BigCSLM+	93,25 %	94,10 %
Big8300SLM+	93,54 %	94,45 %

Kapitola 8

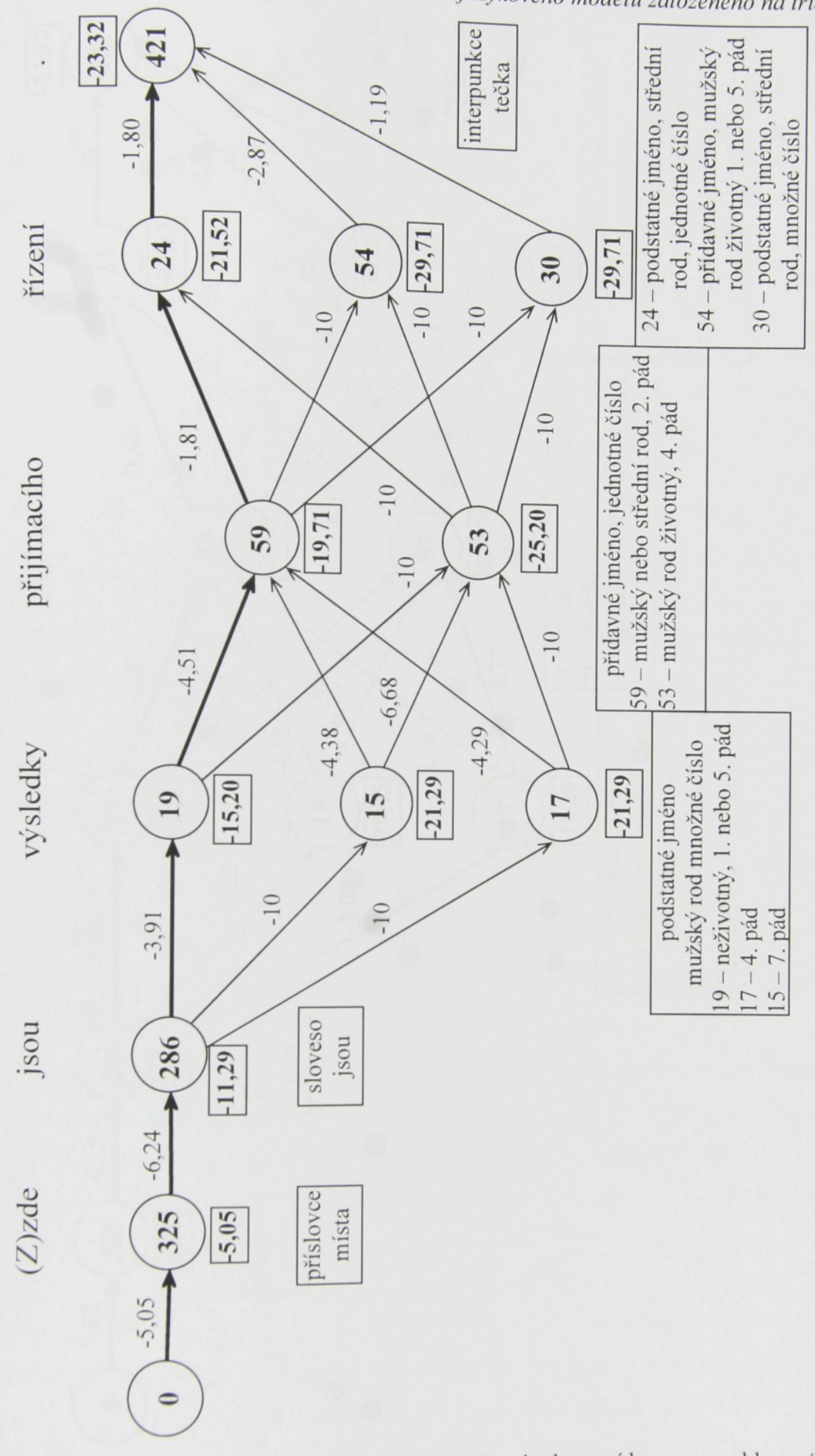
Značkování vět – příklady

V této kapitole jsou uvedeny příklady ohodnocených orientovaných grafů. V grafech je zobrazen postup automatického značkování včetně výběru nejlepší cesty pro konkrétní věty. Při automatickém značkování byl ve všech uvedených příkladech použit slovník o velikosti 313 217 různých slov (slovník 300k – viz kapitola 9.1) a věty jsou značkovány včetně interpunkce.

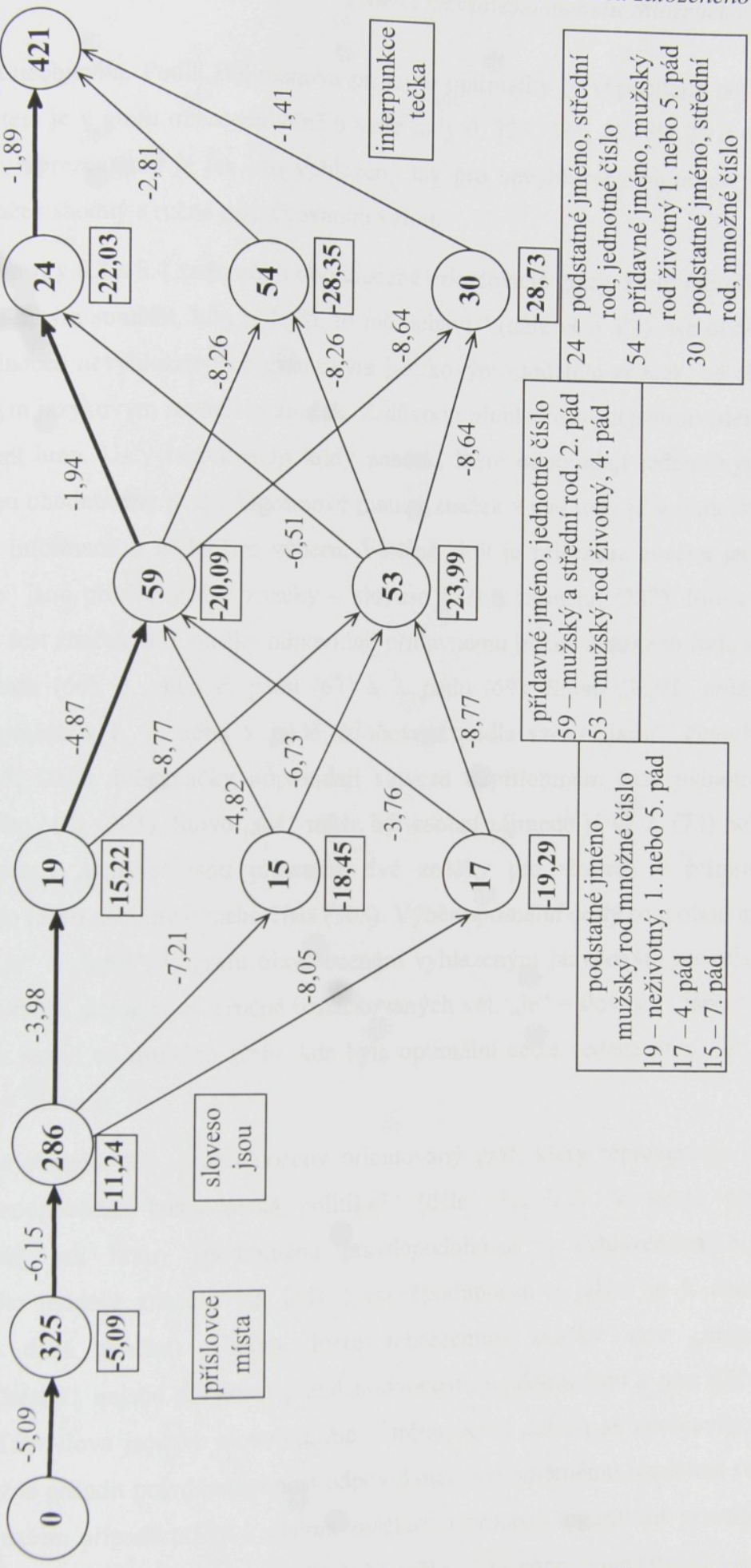
Na obrázcích 8.1 a 8.2 jsou znázorněny ohodnocené orientované grafy, které popisují automatické značkování věty „Zde jsou výsledky přijímacího řízení." (dále věta V1). Oba grafy se liší jen ohodnocením. V obrázku 8.1 se používá pro automatické značkování TaggerB a v obrázku 8.2 TaggerSB.

Jednotlivé uzly grafu reprezentují kódy značek, které jsou přiřazeny jednotlivým slovům věty. Ke slovům „zde", „jsou" a interpunkčnímu znaménku „." jsou přiřazeny značky jednoznačně (příslovce místa – 325, sloveso jsou – 286 a interpunkce tečka – 421). Slovu „výsledky" jsou přiřazeny tři značky. Všechny tři značky reprezentují podstatné jméno mužského rodu množného čísla. Může jít o 1. nebo 5. pád neživotného rodu (19), nebo 4. pád (17), nebo 7. pád (15). Slovo „přijímacího" může být přídavné jméno jednotného čísla buď 2. pádu mužského nebo středního rodu (59), nebo 4. pádu mužského rodu životného (53). Slovu „řízení" jsou přiřazeny tři značky, podstatné jméno středního rodu jednotného čísla (24), podstatné jméno středního rodu množného čísla (30) a přídavné jméno mužského rodu životného 1. nebo 5. pádu (54).

U každé hrany je hodnota přirozeného logaritmu podmíněné pravděpodobnosti načtená z bigramové matice, v obrázku 8.1 z nevyhlazené, v obrázku 8.2 z vyhlazené. U každého uzlu je v rámečku hodnota nejlepšího výběru. Hodnoty –10, které jsou přiřazeny některým hranám grafu v obrázku 8.1, ukazují, že se daná dvojice značek



Obr. 8.1: Graf automatického značkování věty V1, ohodnocení hran z nevyhlazené bigramové matice

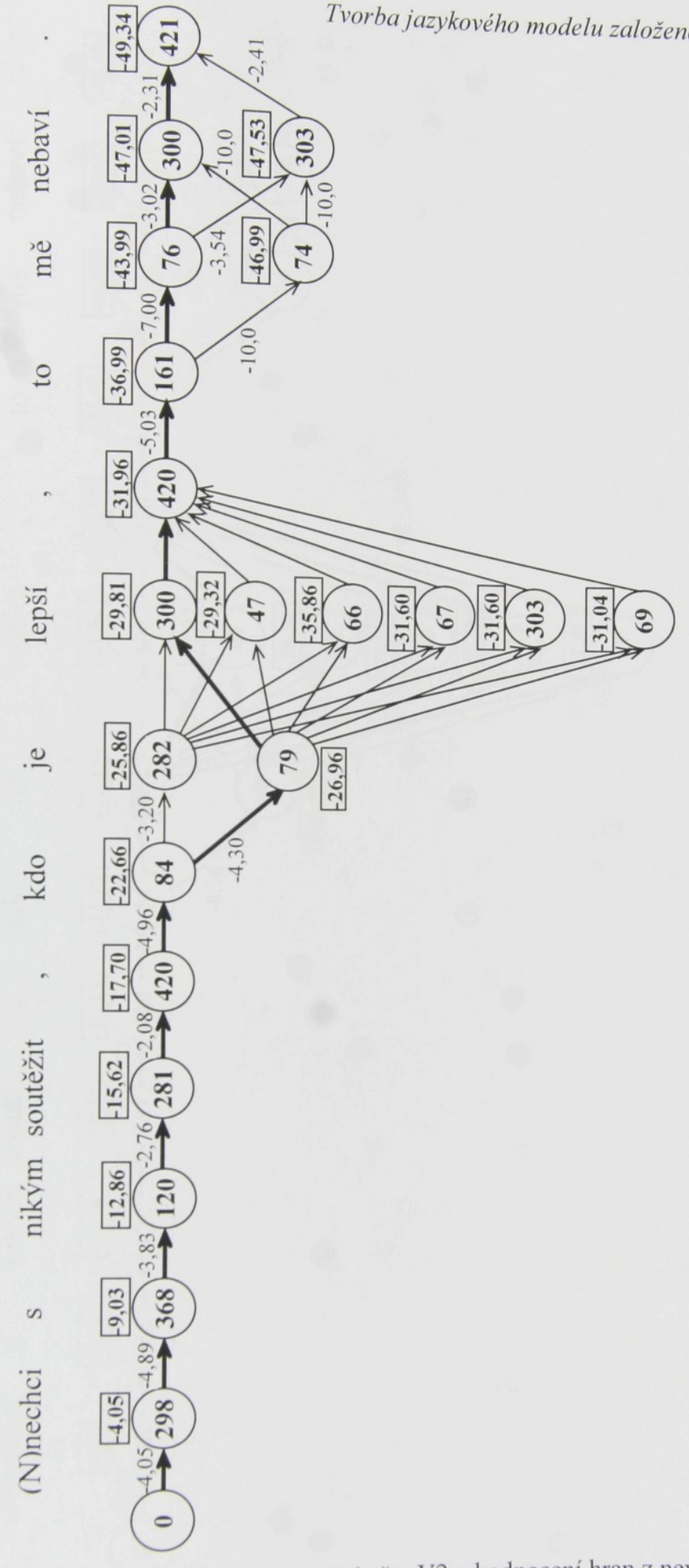


Obr. 8.2: Graf automatického značkování věty V1, ohodnocení hran z vyhlazené bigramové matice

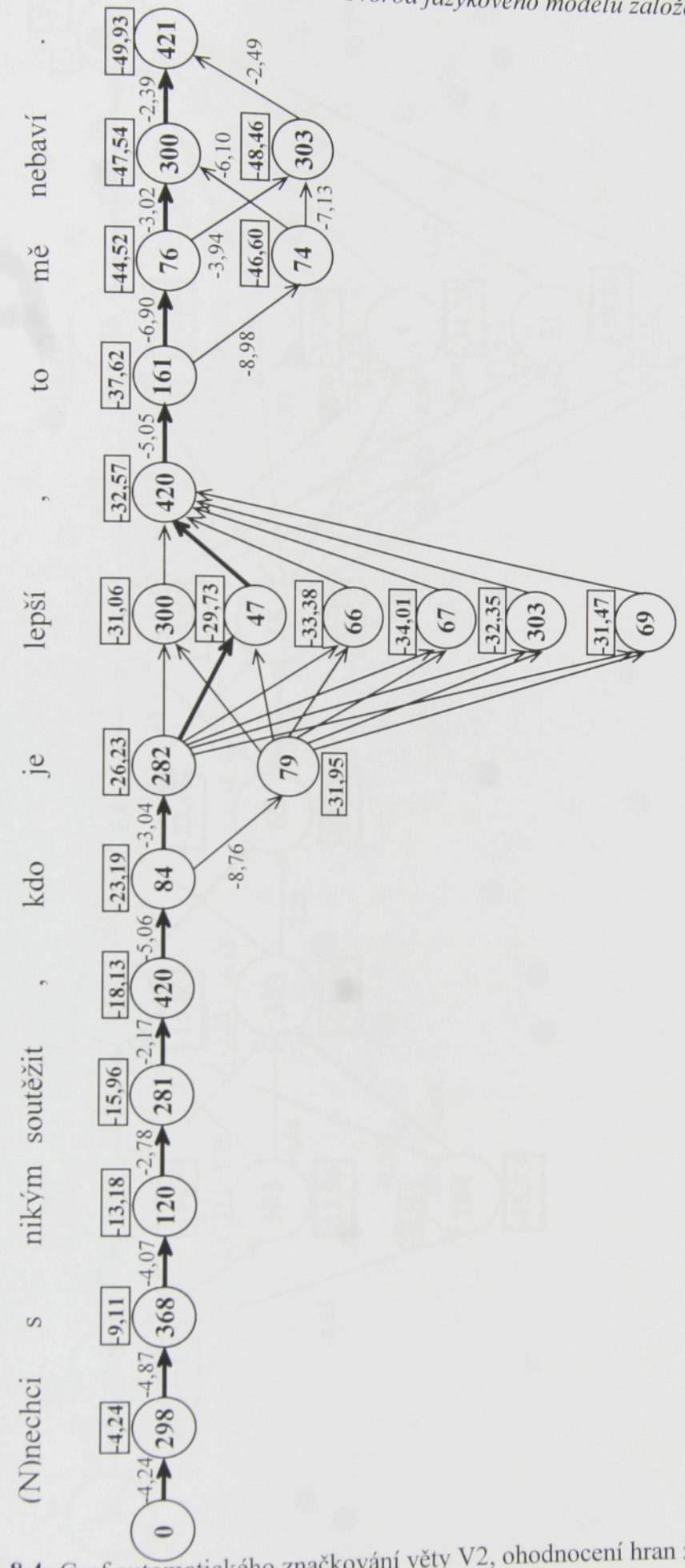
v korpusu neobjevila. Podle Bellmanova principu optimality je vypočítána nejlepší cesta grafem, která je v grafu označena silně a vede uzly 0, 325, 286, 19, 59, 24 a 421. Výběr nejlepších reprezentantů je jak pro vyhlazený tak pro nevyhlazený bigramový jazykový model značek shodný s ručně označkovanou větou.

Obrázky 8.3 a 8.4 znázorňují ohodnocené orientované grafy, které reprezentují větu „Nechci s nikým soutěžit, kdo je lepší, to mě nebaví." (dále věta V2). Na obrázku 8.3 je graf ohodnocen nevyhlazeným bigramovým jazykovým modelem značek, na obrázku 8.4 vyhlazeným jazykovým modelem značek. Z důvodu přehlednosti nejsou uvedena všechna ohodnocení hran. Uzly reprezentují kódy značek, které odpovídají jednotlivým slovům. Hrany jsou ohodnoceny prvky bigramové matice značek a nad uzly je v rámečku uvedena průběžná informace o nejlepším výběru. Většině slov je přiřazena značka jednoznačně. Slovu „je" jsou přiřazeny dvě značky – sloveso (79) a zájmeno (282). Slovu „lepší" je přiřazeno šest značek. Tři značky odpovídají přídavnému jménu ženského rodu jednotného čísla 2. pádu (66), 3. nebo 6. pádu (67) a 7. pádu (69). Slovo „lepší" může být také přídavné jméno v 1., 4, nebo 5. pádě skloňované podle vzoru „jarní", čemuž odpovídá značka 47. Další dvě značky odpovídají slovesu v přítomném čase jednotného (300) a množného čísla (303). Slovo „mě" může být osobní zájmeno já ve 2. (74) nebo 4. pádu (76). Slovesu „nebaví" jsou přiřazeny dvě značky pro sloveso v přítomném čase jednotného (300) nebo množného čísla (303). Výběr optimální cesty se u obou grafů liší ve slovech „je" a „lepší". U grafu ohodnoceném vyhlazeným jazykovým modelem tříd byl výběr proveden stejně jako u ručně označkovaných vět, „je" = sloveso, „lepší" = přídavné jméno, na rozdíl od druhého grafu, kde byla optimální cesta vedena přes „je" = zájmeno a „lepší" = sloveso.

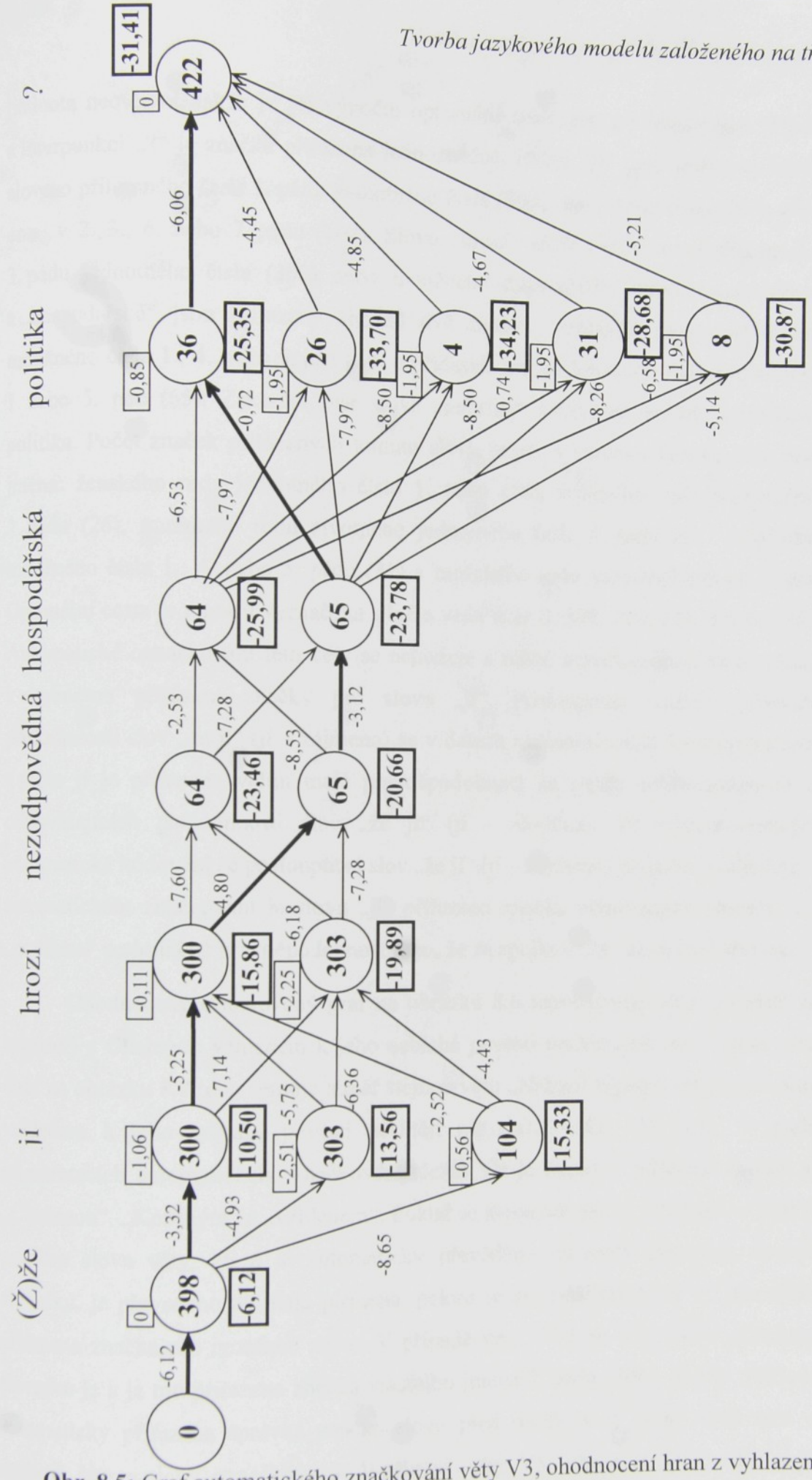
Na obrázku 8.5 je ohodnocený orientovaný graf, který reprezentuje větu „Že jí hrozí nezodpovědná hospodářská politika?" (dále věta V3). V tomto případě jsou ohodnoceny jak hrany (podmíněná pravděpodobnost z vyhlazeného bigramového jazykového modelu značek) tak uzly (pravděpodobnost, s jakou je k danému slovu přiřazena daná značka). Uzlům, které reprezentují značky slov „nezodpovědná" a „hospodářská", nejsou přiřazeny pravděpodobnosti, s jakými jsou k nim přiřazeny dané značky. Tato slova jsou ve slovníku, ale v trénovacích datech se neobjevila. Oběma je proto možné přiřadit pravděpodobnost odpovídající rovnoměrnému rozdělení (viz kapitola 7.3) – v našem případě přidat k oběma značkám přirozený logaritmus pravděpodobnosti 0,5, protože oběma slovům lze přiřadit dvě značky. Výsledné označkování věty však tato



Obr. 8.3: Graf automatického značkování věty V2, ohodnocení hran z nevyhlazené bigramové matice



Obr. 8.4: Graf automatického značkování věty V2, ohodnocení hran z vyhlazené bigramové matice

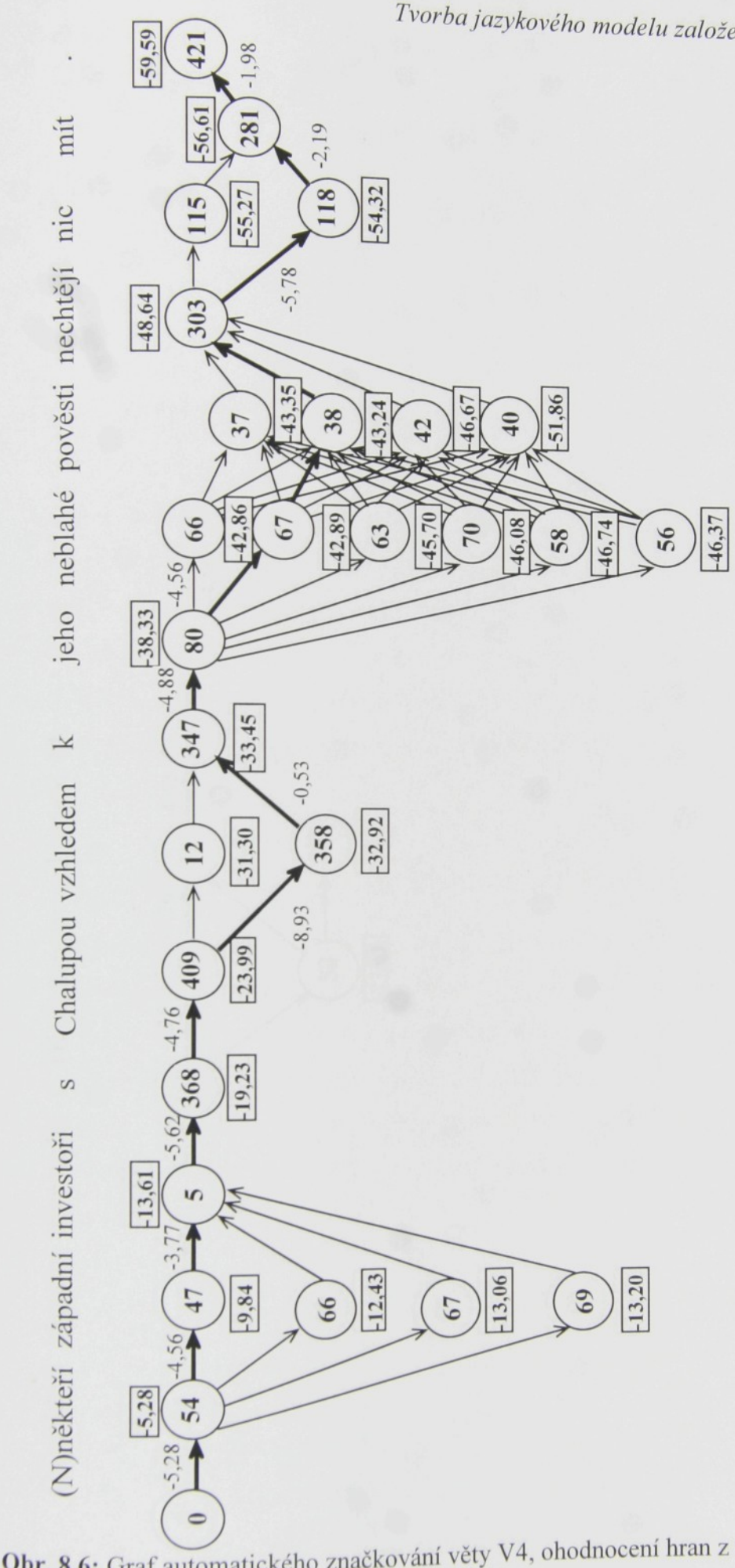


Obr. 8.5: Graf automatického značkování věty V3, ohodnocení hran z vyhlazené bigramové matice, ohodnocení uzlů

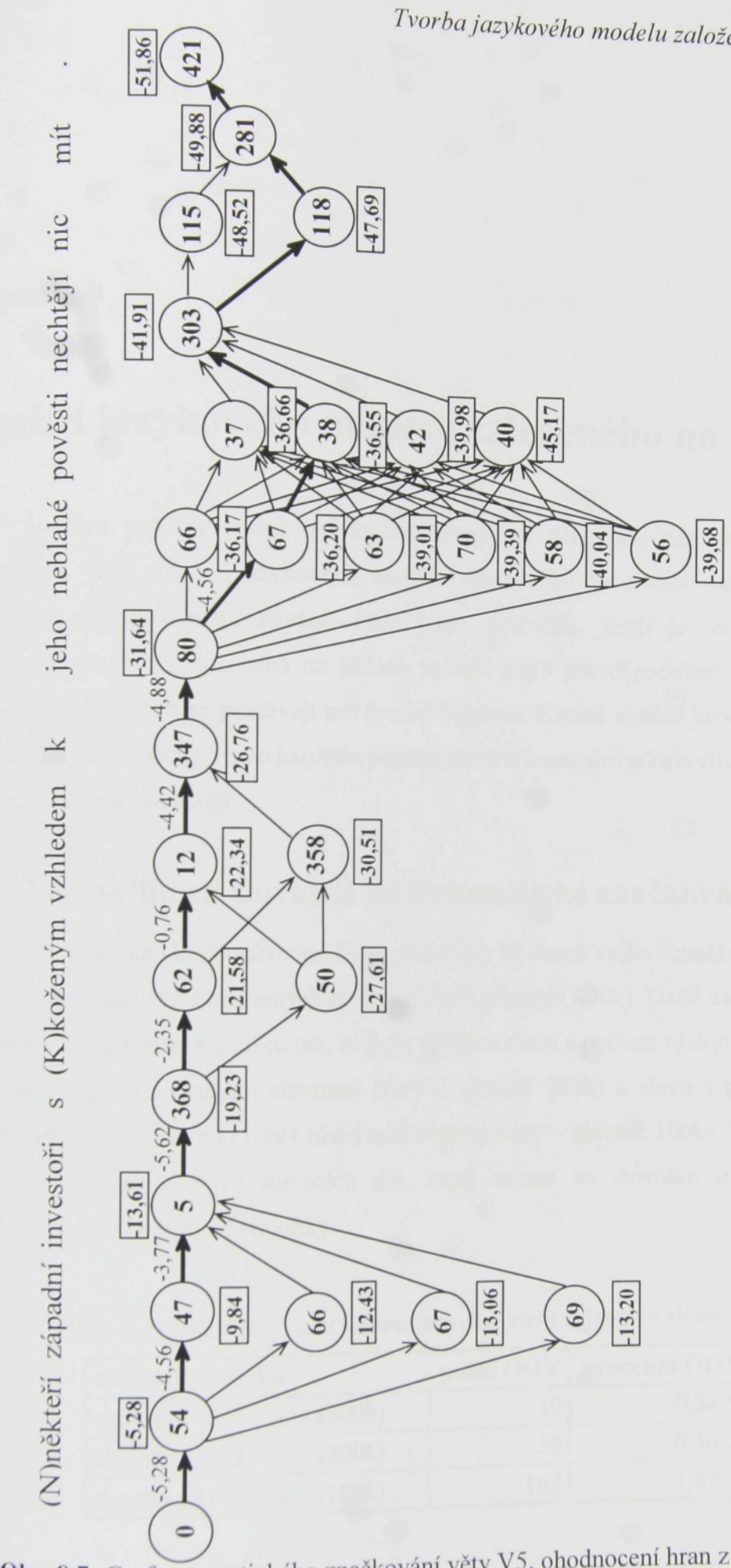
hodnota neovlivní, takže je při výpočtu optimální cesty grafem ignorována. Slovu „že" a interpunkci „?" je značka přiřazena jednoznačně. Slovu „jí" jsou přiřazeny tři značky, sloveso přítomného času 3. pádu jednotného čísla (300), množného čísla (303) a zájmeno ona v 2., 3., 6. nebo 7. pádu (104). Slovo „hrozí" může být sloveso přítomného času 3. pádu jednotného čísla (300) nebo množného čísla (303). Slovům „nezodpovědná" a „hospodářská" jsou přiřazeny shodně dvě značky, přídavné jméno středního rodu množného čísla 1., 4. nebo 5. pád (64) a přídavné jméno ženského rodu jednotného čísla 1. nebo 5. pád (65). Základní tvar slova „politika" může být politik, politikum nebo politika. Počet značek přiřazených tomuto slovu je pět a všechny značky jsou podstatná jména: ženského rodu jednotného číslo 1. pádu (36), středního rodu jednotného čísla 2. pádu (26), mužského rodu životného jednotného čísla 4. pádu (4), středního rodu množného čísla 1., 4. nebo 5. pádu (31) a mužského rodu jednotného čísla 2. pádu (8). Optimální cesta je v grafu vyznačena silně a vede uzly 0, 398, 300, 300, 65, 65, 36 a 422. Automatické označkování této věty se nehoduje s ručně označkovanou větou. Došlo zde k chybnému přiřazení značky pro slovo „jí". Posloupnost značek odpovídajících posloupnosti slov „že jí" (jí – zájmeno) se v datech manuálně označkovaných neobjevila, a proto jí je přiřazena velmi malá pravděpodobnost na rozdíl od posloupnosti značek odpovídajících posloupnosti slov „že jí" (jí – sloveso). Ve větách určených pro automatické kódování se posloupnost slov „že jí" (jí – zájmeno) již jednou objevuje, ale při automatickém značkování je slovu „jí" přiřazena značka odpovídající slovesu, a tak je „posílena" podmíněná pravděpodobnost toho, že za spojkou „že" následuje sloveso.

Ohodnocený orientovaný graf na obrázku 8.6 reprezentuje větu „Někteří západní investoři s Chalupou vzhledem k jeho neblahé pověsti nechtějí nic mít." (dále věta V4). Graf na obrázku 8.7 reprezentuje téměř stejnou větu „Někteří západní investoři s Koženým vzhledem k jeho neblahé pověsti nechtějí nic mít." (dále věta V5). V grafických interpretacích automatického značkování těchto vět je rozdíl v přiřazení značek slovům „Chalupou", „Koženým" a „vzhledem". Pokud se slovo tak, jak je uvedeno ve větě (kromě prvního slova věty, které je automaticky převáděno na malá písmena), neobjeví ve slovníku, je převedeno na malá písmena, pokud se ani poté neobjeví ve slovníku je mu přiřazena značka pro neznámé slovo. V případě věty V4 (obr. 8.6) slovo „Chalupou" ve slovníku je a je mu přiřazena značka vlastního jména 7. pádu (409). Slovu „vzhledem" je automaticky přiřazena správná značka slovo před předložkou (358). Celá tato věta je správně automaticky označkována. V případě věty V5 (obr. 8.7) slovo „Koženým" ve

slovníku není, takže je toto slovo převedeno na malá písmena. Slovo „koženým“ už ve slovníku obsaženo je a je mu přiřazena značka přídavného jména jednotného čísla mužského nebo středního rodu 7. pádu (62) a značka přídavného jména množného čísla 3. pádu (50). V tomto případě je větší pravděpodobnost, že se za přídavným jménem jednotného čísla mužského nebo středního rodu 7. pádu vyskytne podstatné jméno mužského rodu jednotného čísla 7. pádu (12) než slovo před předložkou (358). Tato věta je označkována ve srovnání s ručně označkovanými daty chybně.



Obr. 8.6: Graf automatického značkování věty V4, ohodnocení hran z vyhlazené bigramové matice



Obr. 8.7: Graf automatického značkování věty V5, ohodnocení hran z vyhlazené bigramové matice

Kapitola 9

Využití jazykového modelu založeného na třídách

V části práce popsané v této kapitole jsme zjišťovali, zda je možné použít ohodnocení věty pomocí jazykového modelu založeného na třídách jako metriku pro porovnání přesnosti rozpoznávání. Dále jsme zjišťovali, jestli je možné s pomocí jazykového modelu založeného na třídách vyřadit málo pravděpodobné dvojice slov ze seznamu dvojic, které se používají pro tvorbu bigramů. Kromě využití jazykového modelu založeného na třídách je v této kapitole popsán vliv velikosti slovníku a vliv interpunkce na automatické značkování.

9.1 Vliv velikosti slovníku na automatické značkování

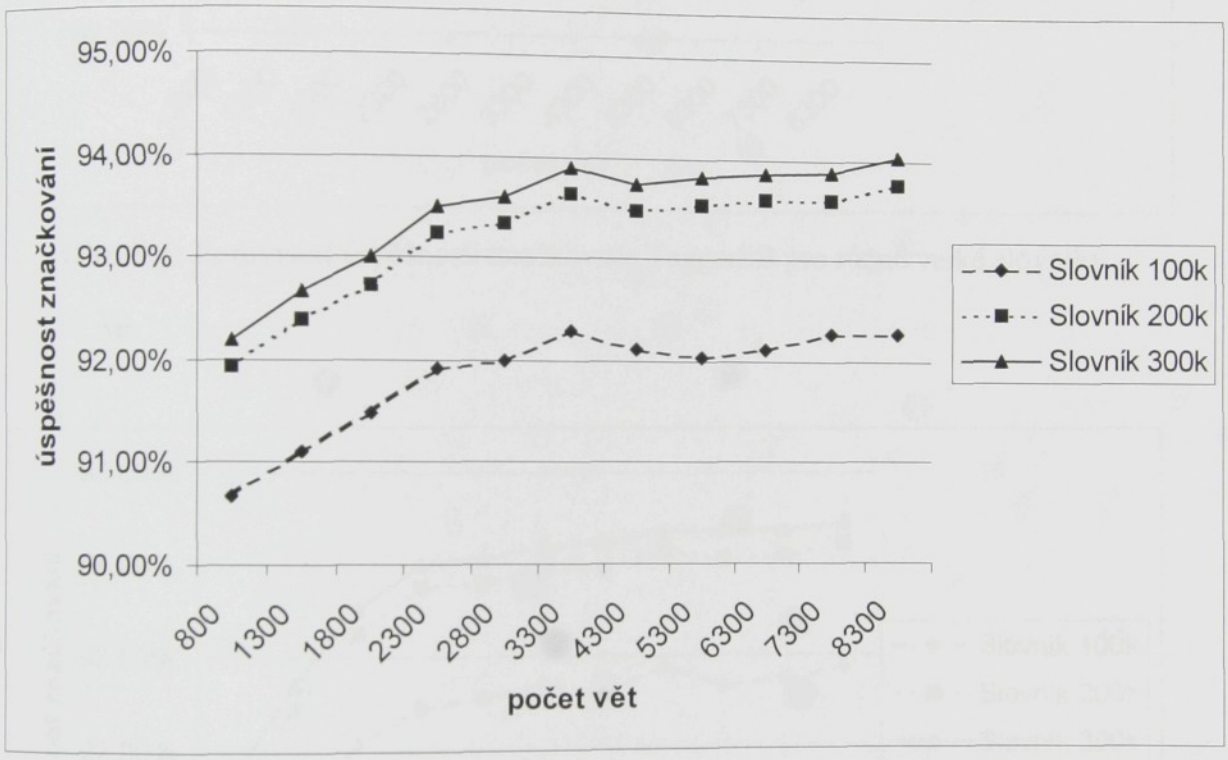
Pro automatické značkování byly používány tři různě velké označkované slovníky. Největší obsahuje 313 217 různých slovních tvarů (slovník 300k). Další dva slovníky byly vytvořeny z největšího slovníku tak, že byla vybrána slova s počtem výskytů větším než 10 (slovník s 213 412 různými slovními tvary – slovník 200k) a slova s počtem výskytů větším než 50 (slovník s 111 294 různými slovními tvary – slovník 100k). V tabulce 9.1 je uvedeno procento slov z testovacích dat, která nejsou ve slovníku (OOV – out of vocabulary), pro jednotlivé slovníky.

Tab. 9.1: Procento slov z testovacích dat, která nejsou ve slovnících

velikost slovníku	počet OOV	procento OOV
slovník 313217 slov (300k)	30	0,34 %
slovník 213412 slov (200k)	50	0,56 %
slovník 111294 slov (100k)	162	1,83 %

Automatické značkování s vyhlazeným a nevyhlazeným bigramovým jazykovým modelem značek, které byly vytvořeny z vět s interpunkcí, bylo prováděno pro všechny tři slovníky. Pro automatické značkování byly postupně použity všechny tři značkovače, které jsou popsány v 7. kapitole.

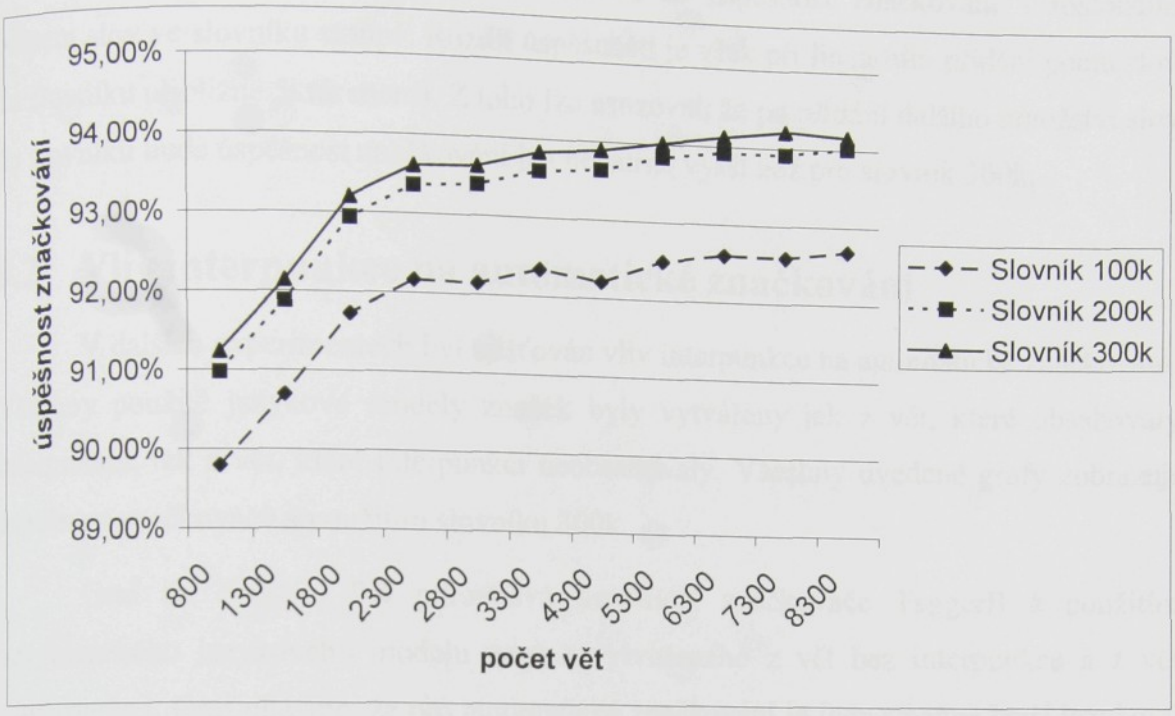
Graf na obrázku 9.1 ukazuje porovnání úspěšnosti značkovače TaggerB v závislosti na velikosti slovníku. Z grafu je zřejmé, že rozdíl úspěšnosti u slovníků 100k a 300k se pohybuje okolo 1,6 %. U slovníků 200k a 300k už je rozdíl úspěšnosti podstatně menší – přibližně 0,26 %.



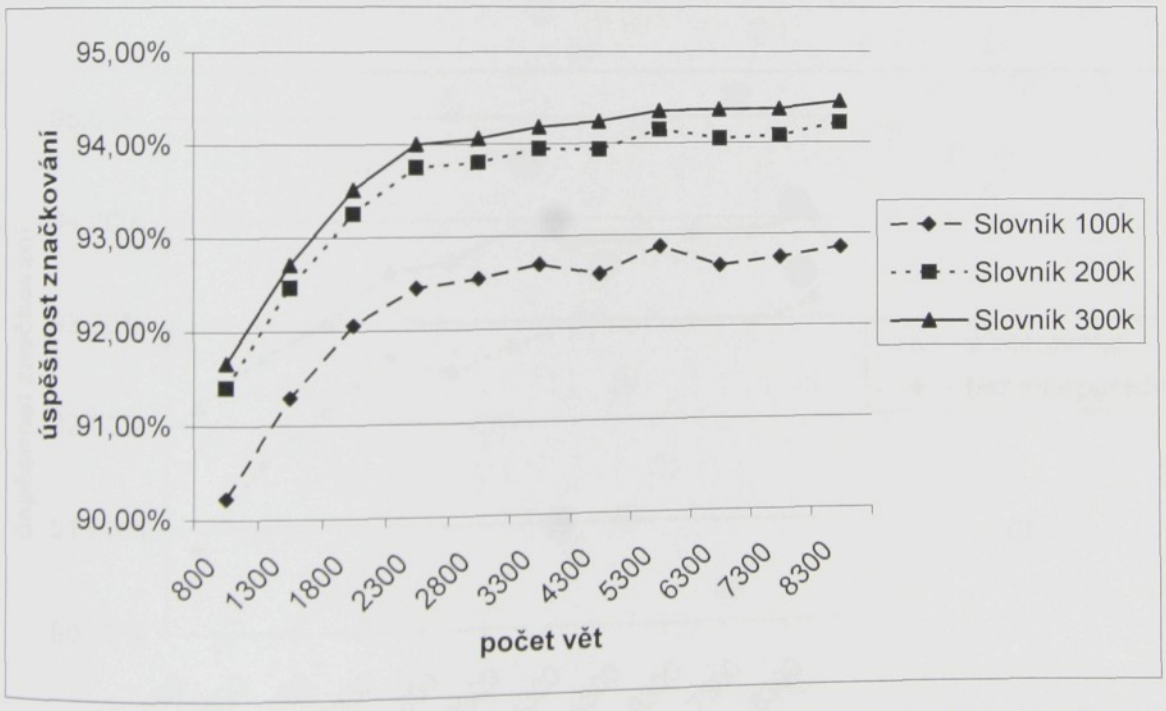
Obr. 9.1: Porovnání úspěšnosti značkovače TaggerB pro různě velké slovníky

Graf na obrázku 9.2 ukazuje porovnání úspěšnosti značkovače TaggerSB v závislosti na velikosti slovníku. Z grafu je zřekmé, že rozdíl úspěšnosti u slovníků 100k a 300k se pohybuje okolo 1,5 %. U slovníku 200k a 300k je rozdíl úspěšnosti přibližně 0,23 %.

Graf na obrázku 9.3 ukazuje porovnání úspěšnosti značkovače TaggerSBP v závislosti na velikosti slovníku. Průběh grafu a rozdíly v úspěšnosti značkování mezi jednotlivými slovníky jsou velice podobné předchozím dvěma grafům. Rozdíl úspěšnosti u slovníků 100k a 300k je přibližně 1,5 %. U slovníků 200k a 300k je rozdíl úspěšnosti 0,26 %.



Obr. 9.2: Porovnání úspěšnosti značkovače TaggerSB pro různě velké slovníky



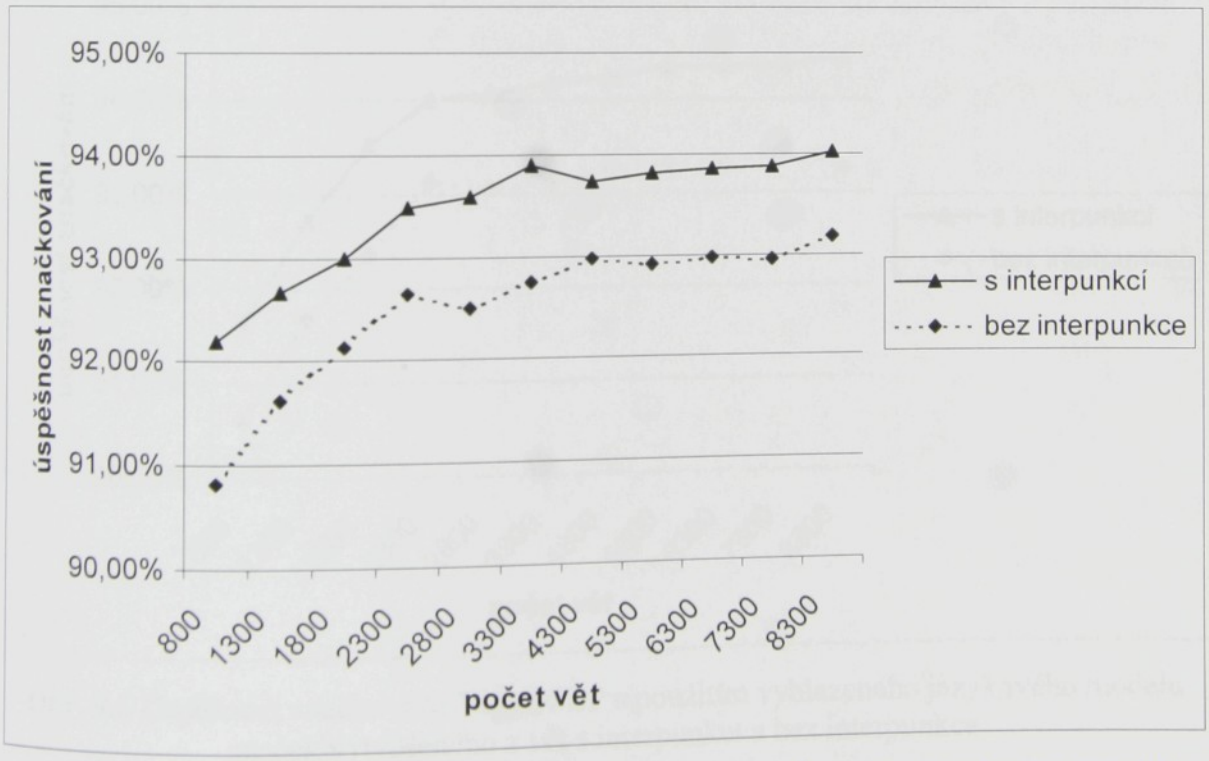
Obr. 9.3: Porovnání úspěšnosti značkovače TaggerSBP pro různě velké slovníky

Z uvedených výsledků lze konstatovat, že úspěšnost značkování s rostoucím počtem slov ve slovníku stoupá. Rozdíl úspěšnosti je však při lineárním přidání počtu slov do slovníku přibližně 5krát menší. Z toho lze usuzovat, že po přidání dalšího množství slov do slovníku bude úspěšnost značkování jen nepatrně vyšší než pro slovník 300k.

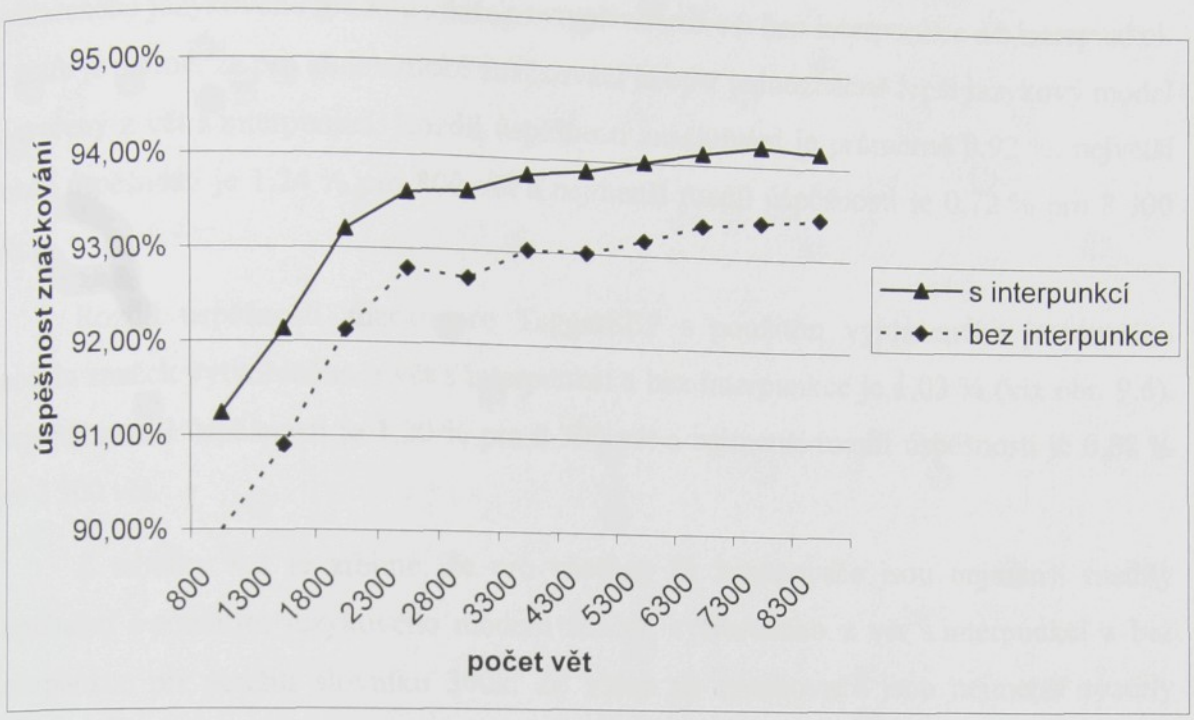
9.2 Vliv interpunkce na automatické značkování

V dalších experimentech byl zjišťován vliv interpunkce na automatické značkování. Všechny použité jazykové modely značek byly vytvářeny jak z vět, které obsahovaly interpunkci, tak z vět, které interpunkci neobsahovaly. Všechny uvedené grafy zobrazují úspěšnost značkovačů s použitím slovníku 300k.

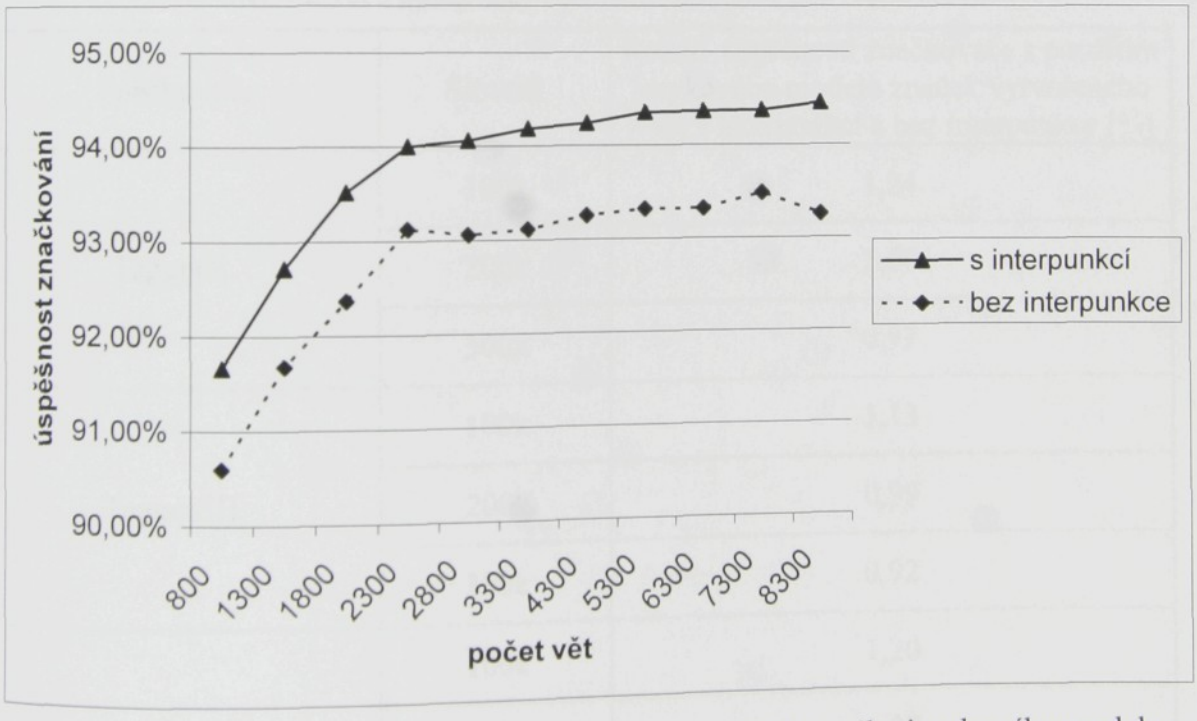
Graf na obrázku 9.4 porovnává úspěšnost značkovače TaggerB s použitím nevyhlazeného jazykového modelu značek vytvořeného z vět bez interpunkce a z vět s interpunkcí. Graf ukazuje, že pro automatické značkování je jednoznačně lepší jazykový model vytvořený z vět s interpunkcí. Rozdíl úspěšnosti značkování je průměrně 0,97 %, největší rozdíl úspěšnosti je 1,37 % pro 800 vět a nejmenší rozdíl úspěšnosti je 0,84 % pro 2 300 vět. Pro 8 300 vět je rozdíl úspěšnosti 0,85 %.



Obr. 9.4: Úspěšnost značkovače TaggerB s použitím nevyhlazeného jazykového modelu značek vytvořeného z vět s interpunkcí a bez interpunkce



Obr. 9.5: Úspěšnost značkovače TaggerSB s použitím vyhlazeného jazykového modelu značek vytvořeného z vět s interpunkcí a bez interpunkce



Obr. 9.6 Úspěšnost značkovače TaggerSBP s použitím vyhlazeného jazykového modelu značek vytvořeného z vět s interpunkcí a bez interpunkce

Graf na obrázku 9.5 porovnává úspěšnost značkovače TaggerSB s použitím vyhlazeného jazykového modelu značek vytvořeného z vět bez interpunkce a s interpunkcí. Z grafu je patrné, že pro automatické značkování je opět jednoznačně lepší jazykový model vytvořený z vět s interpunkcí. Rozdíl úspěšnosti značkování je průměrně 0,92 %, největší rozdíl úspěšnosti je 1,24 % pro 800 vět a nejmenší rozdíl úspěšnosti je 0,72 % pro 8 300 vět.

Rozdíl úspěšnosti značkovače TaggerSBP s použitím vyhlazeného jazykového modelu značek vytvořeného z vět s interpunkcí a bez interpunkce je 1,03 % (viz obr. 9.6). Největší rozdíl úspěšnosti je 1,20 % pro 8 300 vět a nejmenší rozdíl úspěšnosti je 0,88 % pro 2 300 vět.

Z tabulky 9.2 je zřejmé, že pro všechny tři značkovače jsou nejmenší rozdíly úspěšnosti s použitím jazykového modelu značek vytvořeného z vět s interpunkcí a bez interpunkce při použití slovníku 300k. Ze všech tří značkovačů jsou nejmenší rozdíly úspěšnosti u značkovače TaggerSB.

Tab. 9.2: Rozdíl úspěšnosti značkovačů s použitím jazykového modelu značek vytvořeného z vět s interpunkcí a bez interpunkce pro různé slovníky

Značkovač	Slovník	Rozdíl úspěšnosti značkovače s použitím jazykového modelu značek vytvořeného z vět s interpunkcí a bez interpunkce [%]
TaggerB	100k	1,24
	200k	1,00
	300k	0,97
TaggerSB	100k	1,13
	200k	0,99
	300k	0,92
TaggerSBP	100k	1,20
	200k	1,08
	300k	1,03

Přestože jsou rozdíly úspěšnosti značkovačů s použitím jazykového modelu značek vytvořeného z vět s interpunkcí a bez interpunkce poměrně malé, lze tvrdit, že jednoznačně úspěšnější je automatické značkování, které pro ohodnocení používá jazykový model značek vytvořený z vět s interpunkcí.

9.3 Experimenty s jazykovým modelem založeným na třídách

Ohodnocení vět pomocí jazykového modelu založeného na třídách bylo použito pro porovnání a výběr nejlepšího jazykového modelu, který je součástí rozpoznávače. Pro rozpoznávání parlamentní řeči bylo použito několik jazykových modelů. Cílem bylo vybrat ze všech jazykových modelů ten nejlepší. Dále jsme zjišťovali, zda je možné jazykový model založený na třídách použít pro vyřazení málo pravděpodobných dvojic ze souboru dvojic slov, který slouží k tvorbě jazykového modelu pro rozpoznávač řeči. Pro rozpoznávání řeči byl použit systém, vyvinutý na Technické univerzitě v Liberci v letech 2002 až 2005 [NOUZA 2004a].

9.3.1 Popis systému pro rozpoznávání řeči

Vstupem do rozpoznávacího systému je akustický signál, který je digitalizován do šestnáctibitových vzorků při vzorkovací frekvenci 16 kHz. Digitalizovaný signál pak prochází standardním preemfázovým filtrem a je segmentován do framů o délce 20 ms s polovičním překryvem. Data v rámci jednoho framu jsou nejprve předzpracována Hammingovým okénkem a následně jsou vypočteny kepstrální koeficienty. Pro další zpracování se používá prvních 13 kepstrálních koeficientů, které jsou dále doplněny o stejný počet dynamických parametrů odpovídajících první a druhé derivaci koeficientů (tzv. delta a delta-delta koeficienty).

Akustický model rozpoznávacího systému je tvořen inventářem modelů odpovídajících jednotlivým českým hláskám a nejčastěji se vyskytujícím hlukům. Jde o spojité Markovovy modely s multimodálním rozložením. V současné verzi systému se primárně používají kontextově nezávislé hláskové modely tzv. monofony, jejichž výstupní funkce jsou tvořeny váženou směsí až 100 gaussovských rozložení. Základní inventář obsahuje třístavové modely 40 českých hlásek a 7 šumových a ruchových jevů (včetně ticha) [NOUZA 2004a]. Tyto modely byly natrénovány na vlastní databázi obsahující cca 45 hodin anotovaných nahrávek řeči z různých zdrojů (zejména z různých televizních a rozhlasových pořadů a dále z nahrávek pořízených mikrofonem přímo do počítače).

Současná verze rozpoznávacího systému je schopna pracovat se slovníkem o velikosti až 500 000 položek, přičemž dosud největší užívaný slovník obsahuje přibližně 312 000 nejčastějších českých slov (slovník 312k). Systém podporuje více variant výslovností u každého slova, což je využito u asi 20 000 položek. Proces efektivního rozpoznávání spojité řeči je podporován bigramovým jazykovým modelem. Jeho hodnoty byly určeny na rozsáhlém korpusu českých (zejména novinových, zpravodajských a částečně i beletristických) textů o celkovém objemu přibližně 3,5 GB. K vyhlazení jazykového modelu byla použita modifikace [NOUZA 2002] metody navržené původně Wittenem a Bellem [JURAFSKY 2000]. Dekódování je založeno na využití jednoprůchodového časově synchronního Viterbiho algoritmu [NEY 1999], který je výrazně optimalizován s ohledem na práci s velmi rozsáhlými slovníky.

Výstupem z rozpoznávacího systému je textový soubor obsahující:

- informace o parametrech rozpoznávače,
- textovou verzi vyslovené posloupnosti slov bez interpunkce psanou malými písmeny,
- rozpoznanou posloupnost slov včetně označení pro šumy a ruchy,
- u každé věty počet slov, která nejsou ve slovníku,
- chyby, které jsou potřebné k výpočtu přesnosti rozpoznávání, pro každou větu a pro celý soubor,
- vypočtenou přesnost rozpoznání označovanou *Acc* (Accuracy) – viz vzorec (2.49) – pro každou větu i celý soubor.

Výsledky dosahované systémem závisí na obtížnosti úlohy, zejména na stylu řeči, na způsobu výslovnosti mluvící osoby a na kvalitě zaznamenaného signálu. Nejvyšší úspěšnost rozpoznávání (nad 90 %) je u řeči čtené v prostředí s minimálním šumem a hlukem, nižších hodnot úspěšnosti rozpoznávání je dosahováno např. u přepisu záznamů televizních a rozhlasových zpravodajských pořadů (v průměru cca 80 %). U spontánní řeči snímané navíc ve velmi rušném prostředí mohou být výsledky výrazně zhoršeny, a to i pod hranici 50 % správně rozpoznaných slov.

9.3.2 Porovnání přesnosti rozpoznávání s ohodnocením věty

Jedním z úkolů, jak prakticky vyzkoušet jazykový model založený na třídách, bylo ohodnotit rozpoznané věty, porovnat je s přesností rozpoznávání *Acc* a zjistit, zda by ohodnocení $P(W)$ mohlo sloužit jako metrika pro přesnost rozpoznávání. Pro výpočet

ohodnocení vět $P(W)$ byl použit vzorec (2.31), ve kterém byl součin podmíněných pravděpodobností $p(c_i \mid c_{i-1})$ a $p(w_i \mid c_i)$ nahrazen součtem přirozených logaritmů. Podmíněná pravděpodobnost $p(w_1 \mid c_1)$ pro první slovo ve větě je v našem případě rovna jedné, protože každá věta vždy začíná značkou pro začátek věty.

Věty, které byly zpracovány, pocházejí z parlamentních debat a jsou uvedeny bez interpunkce. Parlamentní řeč je specifická tím, že obsahuje hodně dlouhé věty (průměrně 62 slov ve větě), takže často je věta bez interpunkce dost nepřehledná. Tyty věty byly rozpoznány výše uvedeným rozpoznávačem se sedmi různými jazykovými modely (LM): obecným LM (obecny), LM vytvořeným jen z parlamentní řeči (parl only), jejich kombinací (parl x1) a několikerým započítáním LM vytvořeného z parlamentní řeči k obecnému LM (parl x2, parl x4, parl x8, parl x16). Tak vzniklo sedm souborů s různě rozpoznanými 120 větami.

Pro výpočet $P(W)$ je třeba nejdříve věty označkovat. Pro označkování jsme použili TaggerSBP s jazykovým modelem značek BigCSLM+ vytvořeným z velkého množství dat (3,1 GB dat) – viz kapitola 7. Pro každou větu byl proveden výpočet přirozeného logaritmu pravděpodobnosti $P(\mathrm{W})$ a pro každý soubor byl vypočten součet těchto logaritmů $\Sigma P(\mathrm{W})$. Počet slov, které se nevyskytly ve slovníku 300k z jednotlivých souborů (OOV), je uveden v tabulce 9.3.

Tab. 9.3: Počet OOV v souborech se 120 větami rozpoznanými s použitím různých jazykových modelů

	obecny	parl only	parl x1	parl x2	parl x4	parl x8	parl x16
OOV [%]	1,39	1,54	1,21	1,12	1,05	0,97	0,96

Ze všech sedmi souborů byla vybrána pro každou větu nejlepší úspěšnost Acc_{best} rozpoznávání a největší ohodnocení $P(W)_{max}$. Tyto hodnoty byly použity pro výpočet korelačního faktoru:

$$R = \frac{\sum_{\forall N} i \otimes j}{N} \tag{9.1}$$

kde i odpovídá rovnosti $P(W) = P(W)_{max}$,

j odpovídá rovnosti $Acc = Acc_{best}$,

$\otimes$ je operace XOR (jestliže současně platí i a j nebo současně neplatí i a j výsledek je 1, jindy je výsledek 0),

N je počet vět,

R je korelační faktor v procentech.

V tabulce 9.4 jsou uvedeny korelační faktory R, přesnosti rozpoznávání Acc a celkové ohodnocení $\Sigma P(W)$ 120 vět pro všech sedm souborů. Tučně jsou uvedeny maximální hodnoty pro R, Acc a $\Sigma P(W)$. Podle přesnosti rozpoznávání je nejlepší použít jazykový model „parl x8" a jako druhý model „parl x16". Podle korelačního faktoru je nejlepší použít jazykový model „parl x16" a jako druhý model vytvořený jen z parlamentní řeči. V případě, že porovnáváme ohodnocení $\Sigma P(W)$ s přesností rozpoznávání Acc, je třeba si uvědomit, že rozpoznaná věta může obsahovat více resp. méně slov než věta vyslovená a potom i $P(W)$ je větší resp. menší než u správně rozpoznané věty. Např. věta „děkuji za pozornost" byla rozpoznána rozpoznávačem s obecným jazykovým modelem nesprávně („jejich pozornost") s přesností rozpoznávání $Acc = 33$ % a s ohodnocením $P(W) = -14{,}2$. S ostatními jazykovými modely byla tato věta rozpoznána správně s přesností $Acc = 100$ % a ohodnocením $P(W) = -18{,}8$. Objektivně proto nelze porovnávat ohodnocení $\Sigma P(W)$ s přesností, protože při výpočtu $\Sigma P(W)$ záleží na počtu slov ve větě a je potřeba toto ohodnocení nějakým způsobem normovat např. hodnotu $\Sigma P(W)$ dělit celkovým počtem slov. V posledním řádku tabulky 9.4 je toto normované ohodnocení uvedeno. Tučně je opět vyznačena maximální hodnota, která říká, že podle normovaného ohodnocení je nejlepší použít jazykový model „parl x16". Z uvedených výsledků vyplývá, že jako metriku pro porovnání rozpoznaných textů z rozpoznávače s různými parametry je možné kromě přesnosti rozpoznávání použít také korelační faktor popřípadě normované ohodnocení.

Tab. 9.4: Korelační faktor, přesnost rozpoznávání a celkové ohodnocení všech vět pro jednotlivé jazykové modely

jazykový model	obecny	parl only	parl x1	parl x2	parl x4	parl x8	parl x16
R [%]	62,50	66,67	65,00	63,33	63,33	65,00	**71,67**
Acc [%]	67,06	67,15	69,67	70,03	70,43	**70,62**	70,45
$\Sigma P(W)$	-61 347	-62 138	**-61 321**	-61 439	-61 450	-61 529	-61 590
$\Sigma P(W)/N$	-8,14	-8,24	-8,12	-8,11	-8,10	-8,09	**-8,07**

Normovaná ohodnocení rozpoznaných textů uvedená v tabulce jsme porovnali s ohodnocením správně přepsaného textu. V případě, že byl správně přepsaný text uveden včetně přeřeknutí, bylo procento OOV 1,9 % a normované ohodnocení –8,06. V případě, že jsme text upravili bez přeřeknutí, procento OOV bylo nižší (1,3 %) a normované

ohodnocení bylo –8,07. Z těchto výsledků lze opět konstatovat, že normované ohodnocení lze použít jako metriku pro porovnání úspěšnosti rozpoznávání.

Při výpočtu ohodnocení správně rozpoznaných vět hraje velkou roli i počet OOV, což lze ukázat na následujícím příkladě. Vzhledem k tomu, že parlamentní řeč je specifická, byl výpočet ohodnocení $P(W)$ opakován pro 841 vět z denního tisku. Průměrný počet slov ve větě byl přibližně 20. Celková přesnost rozpoznávání podle vzorce (2.49) byla rovna 80,79 %. Počet OOV v přepsaném textu byl 2,76 %. U rozpoznaných vět byl počet OOV 1,56 %. Pro rozpoznávání byl použit rozpoznávač s obecným jazykovým modelem. Hodnota normovaného ohodnocení $\Sigma P(W)/N$ pro rozpoznané věty byla rovna –8,17. Byl proveden výpočet normovaného ohodnocení také pro přepsaný text, přičemž hodnota $\Sigma P(W)/N$ byla rovna –8,09. V tomto případě neodpovídá hodnota normovaného ohodnocení výsledkům z předchozího experimentu, kdy při přesnosti rozpoznávání 70,45 % bylo normované ohodnocení rovno –8,07.

Na základě uvedených výsledků nelze nahradit přesnost rozpoznávání normovaným ohodnocením. Normované ohodnocení a korelační faktor však mohou být dobrým měřítkem pro porovnání výsledků z různých rozpoznávačů nebo z rozpoznávače s různými jazykovými modely. Velkou roli při výpočtu normovaného ohodnocení hraje také počet slov, která se nevyskytují ve slovníku (neznámých slov), což jsou většinou ohebná slova. Největší část neznámých slov tvoří podstatná a přídavná jména. V případě, že jsou tato slova nahrazena značkou pro neznámé slovo, dojde ke ztrátě informace o pádu, rodu, čísle apod., takže s tím souvisí i chyba ve značkování okolních slov a ohodnocení je pak zkreslené.

9.3.3 Odhad četnosti dvojic slov

Jazykový model založený na třídách byl použit také pro odhad četnosti dvojic slov. Soubor s dvojicemi slov včetně počtu výskytů dvojic byl vytvořen z korpusu o velikosti 3,5 GB. Ve dvojicích jsou jen ta slova, která se vyskytla ve slovníku 312k (viz odstavec 9.3.1). Tyto dvojice se používají k vytvoření bigramového jazykového modelu, který je součástí rozpoznávače. Kromě dvojic slov jsou v souboru uloženy i takové dvojice, kde místo jednoho nebo obou slov může být sousloví (kolokace) jako např. „a když“ „addis abeba“ „v letošním“. Počet všech dvojic slov resp. sousloví je 64 351 852. Dvojice slov, ve kterých byla obsažena sousloví, byly při dalším zpracování ignorovány.

Jednou z úloh, kde je možné odhad četnosti dvojic slov využít, je nalezení dvojic s malým odhadem četnosti a jejich vyřazení ze seznamu dvojic. Pro odhad četnosti dvojic slov – úpravou z (2.32) – platí:

$$C(w_{n-1}, w_n) = C(w_{n-1}) \cdot p(w_n \mid c_n) \cdot p(c_n \mid c_{n-1}) \tag{9.2}$$

kde $C(w_{n-1})$ je počet výskytů slova w_{n-1},

$p(w_n \mid c_n)$ je podmíněná pravděpodobnost, s jakou bude k dané značce přiřazeno dané slovo,

$p(c_n \mid c_{n-1})$ je bigram značek.

Při výpočtu odhadu četnosti dvojic slov jsme použili prvky nevyhlazené bigramové matice a podmíněné pravděpodobnosti $p(w_n \mid c_n)$ vyhlazené metodou Add-One Smoothing. Nevyhlazený bigramový jazykový model značek byl vytvořen z automaticky označkovaného korpusu o velikosti 3,1 GB. Interpunkce při tvorbě jazykových modelů byla ignorována, protože i při sestavování dvojic slov z korpusu byla interpunkce ignorována. Ke každému slovu v souboru dvojic byla přidána příslušná značka (značky). V případě, že bylo jednomu nebo oběma slovům ve dvojici přiřazeno více značek, byl vybrán maximální součin ze všech možných součinů dvojic značek pro danou dvojici slov.

V korpusu 3,1 GB je přibližně 4,5 % slov, kterým byla přiřazena značka pro neznámé slovo. Značka pro neznámé slovo se vyskytuje v kombinaci s 94 % značek na levé straně a s 97 % značek na pravé straně dvojice slov. Ve všech dvojicích slov jsou přibližně 4 milióny neznámých slov. Opět jde nejčastěji o podstatná jména. Tato skutečnost samozřejmě napomáhá ke zkreslení odhadů četností přiřazené dvojicím slov. Např. stejný odhad četnosti byl přiřazen jak dvojici slov „dva capart" tak dvojici „dva caparty", protože slovům „capart" a „caparty" je přiřazena značka pro neznámé slovo.

Vzhledem k tomu, že pravděpodobnosti v součinu (9.2) jsou velmi malá čísla, použili jsme váhový koeficient k, kterým jsme celý vztah násobili. Počet výskytů v korpusu je celé číslo, takže vynásobený výsledek byl převeden na celé číslo. Obě uvedené úpravy vzorce (9.2) napomohly k přehlednějším výsledkům. V případě, že byla hodnota váhového koeficientu k nastavena na 100, byla přiřazena nulová hodnota odhadu četnosti přibližně čtvrtině dvojic slov. Pro $k = 10\,000$ byla nulová hodnota odhadu četnosti přiřazena 5 484 448 dvojicím (téměř 10 %). V tabulce 9.5 jsou uvedeny konkrétní příklady dvojic slov s nulovou hodnotou odhadu četnosti pro $k = 10\,000$ včetně počtu výskytů dvojic v ČNK.

Tab. 9.5: Příklady dvojic slov s nulovým odhadem četnosti

slovo 1	skupina 1	slovo 2	skupina 2	četnost v korpusu 3,5 GB	počet výskytů v ČNK
bys	05bys	dva	04dva	1	0
mé	03me	než	08nez	1	0
ho	03ho	ní	03ona0	1	0
ke	07ke3	byl	05byl	1	0
kterého	03tazj2 03tazj4	tisíc	04tisic	1	0
obojí	04nas0	přirozených	02m2 02m6	1	0
ze	07ze	jedny	04nas4 04nas15	1	0

Podle předpokládaných výsledků by měly být ty dvojice slov, u kterých je odhad četnosti velmi malý, vyřazeny. Přestože některé dvojice splňovaly toto kritérium a přestože se tyto dvojice neobjevily v ČNK a v korpusu o velikosti 3,5 GB se objevily velice zřídka, mohou se tyto dvojice slov v textu vyskytnout. Z dvojic uvedených v tabulce 9.5 se ve větě můžeme setkat například s dvojicemi slov „bys dva", „mé než" nebo „kterého tisíc".

Z uvedených výsledků vyplývá, že námi vytvořený nevyhlazený jazykový model založený na třídách není vhodný pro vyřazení dvojic slov, které slouží k tvorbě bigramového modelu pro rozpoznávač.

Jedním z možných řešení, jak některé málo pravděpodobné dvojice slov vyřadit a jiné přidat, by mohlo být vytvoření souboru všech pravděpodobných dvojic ze slovníku 300k s odhadem četností podle (9.2), což je předmětem dalšího výzkumu.

Kapitola 10

Závěr

Dizertační práce předkládá návrh jazykového modelu založeného na třídách včetně jeho praktického použití.

V první řadě bylo třeba stanovit gramatické značky. K stanovení značek byly využity tři přístupy: statistický, gramatický a pravděpodobnostní. S použitím statistického přístupu byla vybrána nejfrekventovanější slova včetně interpunkce, která tvoří přibližně 30 % textu. Těm byly přiřazeny značky tak, že každé značce bylo přiřazeno právě jedno slovo. Pro stanovení dalších značek byl využit hladový algoritmus, gramatický a syntaktický přístup. Seznam všech značek je uveden v příloze.

K tvorbě jazykového modelu založeného na třídách byl vytvořen označkovaný korpus, jehož část byla označkována ručně a část automaticky. Věty, které jsou v korpusu použity, byly získány především z internetu a jde o texty ze zpravodajství, různých novinových článků, knížek apod. Do korpusu byly přidány také věty „uměle" vytvořené, aby byly v korpusu obsaženy i méně frekventované značky.

Pro tvorbu označkovaného slovníku byl využit slovník vytvořený na Technické univerzitě v Liberci obsahující přibližně 300 000 slov, do kterého byla ručně a na základě syntaktické metody přidána informace o příslušných gramatických značkách. Slovník byl použit pro automatické značkování vět. Na základě 3 300 ručně označkovaných vět byly vytvořeny dva bigramové modely značek – nevyhlazený a vyhlazený. Pro vyhlazení jazykového modelu značek byla použita metoda lineární interpolace. Na základě Bellmanova principu optimality bylo provedeno automatické značkování dalších 5 000 vět. Pro automatické značkování byly vytvořeny tři značkovače (taggery). TaggerB využívá pro automatické značkování nevyhlazenou bigramovou matici značek, TaggerSB bigramovou matici značek vyhlazenou metodou lineární interpolace a TaggerSBP

bigramovou matici značek vyhlazenou metodou lineární interpolace a pravděpodobnost výskytu daného slova v dané skupině vyhlazenou metodou Add-One Smoothing. Nejlepší výsledky automatického značkování vykázal značkovač TaggerSBP.

V dizertační práci jsou také shrnuty výsledky automatického značkování s různě velkým slovníkem. Z výsledků uvedených v kapitole 9 vyplývá, že zlepšení úspěšnosti při lineárním zvětšování slovníku stoupá logaritmicky. Při automatickém značkování byly použity jazykové modely značek, které byly vytvořeny z vět s interpunkcí a bez interpunkce. Výsledky jednoznačně prokázaly, že automatické značkování je úspěšnější (cca o 1 %) v případě, že bereme v úvahu interpunkci. Nejlepší výsledky automatického značkování (94,5 %) byly dosaženy při použití značkovače TaggerSBP s jazykovým modelem značek vytvořeným z vět s interpunkcí a s využitím slovníku o velikosti 300 tisíc slov.

Na základě experimentů s větami rozpoznanými systémem pro rozpoznávání spojité řeči jsme došli k závěru, že pomocí jazykového modelu lze rozpoznané věty ohodnotit a normované ohodnocení použít pro porovnání výsledků z různých rozpoznávačů nebo z rozpoznávače s různými jazykovými modely.

Přestože jsme předpokládali, že s pomocí nevyhlazeného jazykového modelu založeného na třídách bude možné vyřadit málo pravděpodobné dvojice slov ze seznamu dvojic, které se používají pro tvorbu bigramů, výsledky experimentů tuto hypotézu vyvrátily.

Na základě výsledků dizertační práce hodláme vytvořit soubor všech pravděpodobných dvojic slov ze slovníku o velikosti 300 tisíc slov s odhadem četností pomocí bigramového jazykového modelu založeného na třídách. Soubor dvojic se využívá pro tvorbu jazykového modelu, který je součástí rozpoznávače. Předpokládáme, že s takto vytvořeným jazykovým modelem, by mohlo dojít ke zlepšení přesnosti rozpoznávání. Předmětem dalšího výzkumu je vytvoření trigramového jazykového modelu založeného na třídách a jeho využití nejen ve finální fázi rozpoznávání pro korekturu již rozpoznaných vět, ale i v předzpracování textů sloužících pro tvorbu bigramového jazykového modelu, který je součástí rozpoznávače.

Literatura

[BASE] *The British Academic Spoken English Corpus*. University of Warwick and Reading. URL: <http://www.rdg.ac.uk/AcaDepts/ll/base_corpus>.

[BENGIO 1999] BENGIO, Yoshua. *Markovian Models for Sequential Data*. Neural Computing Surveys, Vol. 2, s. 129–162, 1999. ISSN 1093-7609. URL: <ftp://ftp.icsi.berkeley.edu/pub/ai/jagota/vol2_5.pdf>.

[BNC] *British National Corpus*. Oxford University Computing Services. URL: < http://www.natcorp.ox.ac.uk/index.html>.

[BOLDIŠ 2004a] BOLDIŠ, Petr. *Bibliografické citace dokumentů podle ČSN ISO 690 a ČSN ISO 690-2: Část 1 – Citace: metodika a obecná pravidla*. Verze 3.3. © 1999–2004, poslední aktualizace 11.11. 2004. URL: <http://www.boldis.cz/citace/citace1.pdf>.

[BOLDIŠ 2004b] BOLDIŠ, Petr. *Bibliografické citace dokumentů podle CSN ISO 690 a CSN ISO 690-2: Část 2 – Modely a příklady citací u jednotlivých typů dokumentů*. Verze 3.0 (2004). © 1999–2004, poslední aktualizace 11. 11. 2004. URL: <http://www.boldis.cz/citace/citace2.pdf>.

[BROWN 1992] BROWN, Peter F., et al. *Class-Based n-gram Models of Natural Language*. Computational Linguistics 18, No. 4, s. 467–479, 1992, ISSN 0891-2017.

[BROWNC] *The Brown Corpus*. URL: <http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html>.

[CHURCH 1988] CHURCH, Keneth Ward. *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text*. In Proceedings of the Second Conference on Applied Natural Language Processing (ANLP), Austin, Texas, s. 136–143, 1988.

[ČECHOVÁ 2000] ČECHOVÁ, M. a kolektiv. *Čeština – řeč a jazyk*, ISV nakladatelství, Praha, 2000, ISBN 80-85866-57-9.

[ČERMÁK 2004] ČERMÁK, František – SCHMIEDTOVÁ, Věra. *Český národní korpus – základní charakteristika a širší souvislosti*. Národní knihovna, 15, 2004, č. 3, s. 152–168, ISSN 1214-0678. URL: <http://full.nkp.cz/nkkr/nkkr0403/0403152.html>.

[ČNK 2000] *Český národní korpus – SYN2000*. Ústav Českého národního korpusu FF UK, Praha 2000. URL: <http://ucnk.ff.cuni.cz>.

[DICTIONARY] *Dictionary of Algorithms and Data Structures*. National Institute of Standards and Technology. URL: <http://www.nist.gov/dads>.

[DRÁBEK 1998] DRÁBEK, Jaromír. *Řízení složitých systémů metodami síťové analýzy*. Praha, 1998. Dizertační práce na Fakultě elektrotechnické Českého vysokého učení technického v Praze.

[DRÁBKOVÁ 2003] DRÁBKOVÁ, Jindra. *Formation of Classes for Continuous Speech Language Model and Building the Large Tagging Vocabulary for Czech Language*. In: Proc. of 13th Czech-German Workshop „Speech Processing", September 2003, Prague, Czech Republic, s. 121–125, ISBN 80-86269-10-8

[DRÁBKOVÁ 2005] DRÁBKOVÁ, Jindra. *Punctuation Effect on Class-Based Language for Czech Language*. In: Electronic Speech Signal Processing, September 2005, Prague, Czech Republic, s. 267–272, ISBN 80-86269-10-8

[FRANCIS 1979] FRANCIS, W. Nelson – KUCERA, Henry. *Brown Corpus Manual*. Brown University, 1979. URL: <http://khnt.hit.uib.no/icame/manuals/brown/index.htm>.

[HAJIČ 2000a] HAJIČ, Jan. *Introduction to NLP (Fall 2000)*. URL: <http://www.cs.jhu.edu/~hajic/courses/cs465/syllabus.html>.

[HAJIČ 2000b] HAJIČ, Jan. *Positional Tags*: *Quick Reference* (*Czech HM Morphology*). Praha, 2000. URL: <http://quest.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html>.

[HÁNOVÁ 2003] HÁNOVÁ, Kristina. *Blokování lexikálního faktoru*. Praha, 2003. Písemná práce z lingvistiky na Filozofické fakultě Univerzity Karlovy.

[HUANG 2001] HUANG, Xuedong – ACERO, Alex – HON Hsiao-Wuen. *Spoken Language Processing*. Prentice Hall PTR, Upper Saddle River, New Jersey, 2001, ISBN 0-13-022616-5.

[IRCING 2003a] IRCING, Pavel. *Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language* (*Czech*). Plzeň, 2003. Dizertační práce na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

[IRCING 2003b] IRCING, Pavel – PSUTKA, Josef. *Fitting Class-Based Language Models into Weighted Finite-State Transducer Framework*. In Proceedings of Eurospeech 2003, the 8th European Conference on Speech Communication and Technology, September 1–4, 2003, Geneva, Switzerland, s. 1873–1876, ISSN 1018-4074.

[JURAFSKY 2000] JURAFSKY, Daniel – MARTIN, James H. *Speech and Language Processing*. Prentice-Hall, Inc., New Jersey, 2000, ISBN 0-13-095069-6.

[JURČÍČEK 2003] JURČÍČEK, Filip. *Dekodér systému rozpoznávání souvislé mluvené řeči s velkým slovníkem (LVCSR) s n-gramovým jazykovým modelem.* Plzeň, 2003. Diplomová práce na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

[LEECH 1994] LEECH, G. – GARSIDE, R. – BRYANT, M. *CLAWS4: The tagging of the British National Corpus.* In Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) Kyoto, Japan, 1994, s. 622–628.

[LEECH 2000] LEECH, Geoffrey – SMITH, Nicholas. *The British National Corpus (Version 2) with Improved Word-class Tagging.* Manual to accompany, University Centre for Computer Corpus Research on Language, Lancaster University, UK, March 2000.

[LINARES 2004] LINARES, Diego – BENEDÍ, José-Miquel – SÁNCHEZ, Joan-Andreu. *A hybrid language model based on a combination of N-grams and stochastic context-free grammars*, ACM Transactions on Asian Language Information Processing (TALIP), Volume 3, Issue 2, June 2004, s. 113–127, ISSN 1530-0226.

[MARCUS 1993] MARCUS, Mitchell P. – SANTORINI, Beatrice – MARCINKIEWICZ, Mary Ann. *Building a large annotated corpus of English: the Penn Treebank.* Computational Linguistics 19, No. 2, s. 313–330, 1993, ISSN 0891-2017.

[MONZ 2004] MONZ, Christof. *Introduction to Computational Linguistics*, Fall 2004. Syllabus for CMSC723/LING645. URL: http://www.umiacs.umd.edu/~christof/courses/cmsc723-fall04/syllabus.html.

[MÍROVSKÝ 1998] MÍROVSKÝ, Jiří. *Morfologické značkování textu: automatická disambiguace.* Praha, 1998. Diplomová práce na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze.

[MRVA 2000] MRVA, David. *Jazykové modelování přirozeného jazyka založené na kořenech a koncovkách.* Praha, 2000. Diplomová práce na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze.

[NEJEDLOVÁ 2002] NEJEDLOVÁ, Dana. *Comparative Study on Bigram Language Models for Spoken Czech Recognition.* In: Proc. of TSD 2002, September 9–12, 2002, Brno, Czech Republic, s. 197–204, ISBN 3-540-44129-8.

[NEJEDLOVÁ 2004] NEJEDLOVÁ, Dana. *Building and Evaluation of a Large Vocabulary for a Czech Voice Dictation System.* Proc. of ECMS 2003, June 2003, Liberec, s. 74–78, ISBN 80-7083-708-X.

[NEY 1999] NEY, H. – ORTMANNS, S. *Dynamic Programming Search for Continuous Speech Recognition*, IEEE Signal Processing Magazine, vol.16, No.5, September 1999, s. 64–83

[NOUZA 2001] *Počítačové zpracování řeči: cíle, problémy, metody a aplikace.* Editor Jan Nouza. Liberec, 2001, ISBN 80-7083-551-6.

[NOUZA 2002] NOUZA, J. – DRÁBKOVÁ, J. *Combining Lexical and Morphological Knowledge in Language Model for Inflectional (Czech) Language*. In Proc. of 6th Int. Conference on Spoken Language Processing, September 2002, Denver USA, s. 705–708, ISBN 1876346418.

[NOUZA 2003] NOUZA, Jan. *Voice Dictation into a PC: Recent Research State at TUL*. Proc. of ECMS 2003, June 2003, Liberec, s. 69–73, ISBN 80-7083-708-X.

[NOUZA 2004a] NOUZA, Jan, et al. *Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs*. In: Proc. of ICSLP 2004, October 2004, Jeju Island, Korea, s. 409–412, ISSN 1225-441x.

[NOUZA 2004b] NOUZA, Jan – NOUZA, Tomáš. *A Voice Dictation System for a Million-Word Czech Vocabulary*. In: Proc. of ICCCT 2004, August 2004, Austin, USA, s. 149–152, ISBN 980-6560-17-5.

[OLIVA 2000] OLIVA, Karel, et al. *The Linguistic Basis of a Rule-Based Tagger of Czech*. In: Proceedings of the Third International Workshop on Text, Speech and Dialogue, Brno, Czech Republic, September 13–16, 2000, s. 3–8, ISBN 3-540-41042-2.

[OSOLSOBĚ 1995] OSOLSOBĚ, Klára – PALA, Karel – RYCHLÝ, Pavel. *Frekvence vzorů českých sloves*. Brno, 1995. URL: <http://nlp.fi.muni.cz/publications/sas1998_pala_osolsobe_pary/sas1998_pala_osolsobe_pary.htm>.

[PALA 1996] PALA, Karel. *Informační technologie a korpusová lingvistika (1)* [on-line]. Zpravodaj ÚVT MU, roč. VI, č. 3, Brno, leden 1996. URL: <http://www.ics.muni.cz/zpravodaj/zpravodaj.html>, ISSN 1212-0901.

[POTAMIANOS 1998] POTAMIANOS, G. – JELINEK, F. *A Study of n-gram and Decision Tree Letter Language Modeling Methods*. Speech Communication, Elsevier, Holland, Vol. 24, s. 171–192, June 1998.

[PDT 2005] *Pražský závislostní korpus 2.0*. Ústav formální a aplikované lingvistiky MFF UK, Praha 2005. URL: < http://ufal.mff.cuni.cz/pdt2.0/index-cz.html>.

[PSUTKA 1995] PSUTKA, J. *Komunikace s počítačem mluvenou řečí*. Academia, Praha, 1995, ISBN 80-200-0203-0.

[RABINER 1989] RABINER, Lawrence R. *A tutorial on Hidden Markov Models and selected applications in speech recognition*. In Proceedings of IEEE, volume 77, s. 257–286, February 1989.

[RIBAROV 2000] RIBAROV, Kiril. *Rule-Based Tagging: Morphological Tagsets versus Tagset of Analytical Functions*. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece, 2000, s. 1123–1125.

[RYCHLÝ 1997] RYCHLÝ, Pavel. *Korpusy textů na FI MU* [on-line]. Zpravodaj ÚVT MU, roč. VIII, č. 2, Brno, prosinec 1997. URL: <http://www.ics.muni.cz/zpravodaj/zpravodaj.html>, ISSN 1212-0901.

[SAMUELSSON 1997] SAMUELSSON, Christer – VOUTILAINEN, Atro. *Comparing a Linguistic and a Stochastic Tagger*. In Proceedings of 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, ACL, Madrid, 1997.

[SANTORINI 1990] SANTORINI, Beatrice. *Part-of-speech tagging guidelines for the Penn Treebank Project*. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, 1990.

[SEDLÁČEK 1999] SEDLÁČEK, Radek: *Morfologický analyzátor češtiny*. Brno, 1999. Diplomová práce na Fakultě informatiky Masarykovy Univerzity v Brně.

[VYMAZAL 2001] VYMAZAL, Petr. *Statistické charakteristiky češtiny*. Brno, 2001. Diplomová práce na Fakultě informatiky Masarykovy Univerzity.

[VOUTILAINEN 1993] VOUTILAINEN, Atro – HEIKKILÄ, Juha. *An English Constraint Grammar (ENGCG) a surface-syntactic parser of English*, Research Unit for Computational Linguistics University of Helsinki. In Udo Fries, Gunnel Tottie & Peter Schneider, eds. Creating and Using English Language Corpora. Amsterdam: Rodopi, 1993, s. 189–199.

[WARAKAGODA 1996] WARAKAGODA, Narada Dilp. *A Hybrid ANN-HMM ASR system with NN based adaptive preprocessing*. Transmisjonsteknikk, 1996. M.Sc. thesis, Norges Tekniske Høgskole, Institutt for Teleteknikk.

[ZEMAN 1998] ZEMAN, Daniel. *Pravděpodobnostní model významových zápisů vět*. Praha, 1998. Diplomová práce na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze.

Seznam vlastních publikací

NEJEDLOVÁ, D. – DRÁBKOVÁ, J. – KOLORENČ, J. – NOUZA, J. *Lexical, Phonetic, and Grammatical Aspects of Very-Large-Vocabulary Continuous Speech Recognition of Czech Language*. In: Electronic Speech Signal Processing, September 2005, Prague, Czech Republic, s. 224–231. ISBN 80-86269-10-8

DRÁBKOVÁ, J. *Punctuation Effect on Class-Based Language for Czech Language*. In: Electronic Speech Signal Processing, September 2005, Prague, Czech Republic, s. 267–272. ISBN 80-86269-10-8

DRÁBKOVÁ, J. – HOLADA, M. – NOUZA, J. – HORÁK, P. – NOUZA, T. *New Version of Phone Dialogue Information System InfoCity*. In: Proc. of 14th Czech-German Workshop „Speech Processing", September 2004, Prague, Czech Republic, s. 66–71, ISBN 80-86269-11-6

DRÁBKOVÁ, J. *Formation of Classes for Continuous Speech Language Model and Building the Large Tagging Vocabulary for Czech Language*. In: Proc. of 13th Czech-German Workshop „Speech Processing", September 2003, Prague, Czech Republic, s. 121–125. ISBN 80-86269-10-8

DRÁBKOVÁ, J. *How Good is Speech Recognition Performed by Human and by Machine?* In Proc. of 6th International Workshop on Electronics, Control, Measurment and Signals-ECMS 2003. Liberec, June 2003. s. 79–83. ISBN 80-7083-708-X

DRÁBKOVÁ, J. *Language Model Based on the Czech Morphology*. In Proc. of 12th Czech-German Workshop „Speech Processing". Prague, September 2002, s. 70–73. ISBN 80-86269-09-4

NOUZA, J. – DRÁBKOVÁ, J. *Combining Lexical and Morphological Knowledge in Language Model for Inflectional (Czech) Language*. In Proc. of 6th Int. Conference on Spoken Language Processing. Denver USA, September 2002, s. 705–708. ISBN 1876346418

NOUZA, T. – NOUZA, J. – DRÁBKOVÁ, J. *An Efficient Graphic System for Developing Voice Operated Applications*. In Proc. of SCI 2002. Orlando USA, July 2002, Volume I, s. 239–244. ISBN 980-07-8150-1

CHALOUPKA, J. – NOUZA, J. – DRÁBKOVÁ, J. *Developing an Artificial Talking Head for Czech Language*. In Proc. of SCI 2002. Orlando USA, July 2002, Volume III, s. 232–236. ISBN 980-07-8150-1

Příloha

Seznam značek pro český jazyk

Podstatná jména:

jednotné číslo:

mužský rod:

01mj1	podstatné jméno, mužský rod, jednotné číslo, 1. pád
01mj2	podstatné jméno, mužský rod, jednotné číslo, 2. pád
01mj3	podstatné jméno, mužský rod, jednotné číslo, 3. pád
01maj4	podstatné jméno, mužský rod životný, jednotné číslo 4. pád
01mnj4	podstatné jméno, mužský rod neživotný, jednotné číslo 4. pád
01mj5	podstatné jméno, mužský rod, jednotné číslo, 5. pád
01mj6	podstatné jméno, mužský rod, jednotné číslo, 6. pád
01mj7	podstatné jméno, mužský rod, jednotné číslo, 7. pád

ženský rod:

01zj1	podstatné jméno, ženský rod, jednotné číslo, 1. pád
01zj2	podstatné jméno, ženský rod, jednotné číslo, 2. pád
01zj36	podstatné jméno, ženský rod, jednotné číslo, 3. nebo 6.pád
01zj4	podstatné jméno, ženský rod, jednotné číslo, 4. pád
01zj5	podstatné jméno, ženský rod, jednotné číslo, 5. pád
01zj7	podstatné jméno, ženský rod, jednotné číslo, 7. pád

střední rod:

01sj145	podstatné jméno, střední rod, jednotné číslo, 1. nebo 4. nebo 5. pád
01sj2	podstatné jméno, střední rod, jednotné číslo, 2. pád
01sj3	podstatné jméno, střední rod, jednotné číslo, 3. pád

01sj6	podstatné jméno, střední rod, jednotné číslo, 6. pád
01sj7	podstatné jméno, střední rod, jednotné číslo, 7. pád
01sj	podstatné jméno, střední rod, jednotné číslo

množné číslo:

mužský rod:

01mnm15	podstatné jméno, mužský rod neživotný, množné číslo, 1. nebo 5. pád
01mam15	podstatné jméno, mužský rod životný, množné číslo, 1. nebo 5. pád
01mm2	podstatné jméno, mužský rod, množné číslo, 2. pád
01mm3	podstatné jméno, mužský rod, množné číslo, 3. pád
01mm4	podstatné jméno, mužský rod, množné číslo, 4. pád
01mm6	podstatné jméno, mužský rod, množné číslo, 6. pád
01mm7	podstatné jméno, mužský rod, množné číslo, 7. pád

ženský rod:

01zm145	podstatné jméno, ženský rod, množné číslo, 1. nebo 4. nebo 5. pád
01zm2	podstatné jméno, ženský rod, množné číslo, 2. pád
01zm3	podstatné jméno, ženský rod, množné číslo, 3. pád
01zm6	podstatné jméno, ženský rod, množné číslo, 6. pád
01zm7	podstatné jméno, ženský rod, množné číslo, 7. pád

střední rod:

01sm145	podstatné jméno, střední rod, množné číslo, 1. nebo 4. nebo 5. pád
01sm2	podstatné jméno, střední rod, množné číslo, 2. pád
01sm3	podstatné jméno, střední rod, množné číslo, 3. pád
01sm6	podstatné jméno, střední rod, množné číslo, 6. pád
01sm7	podstatné jméno, střední rod, množné číslo, 7. pád
01sm	podstatné jméno, střední rod, množné číslo

nesklonná podstatná jména:

01neskl	podstatné jméno nesklonné

podstatná jména spadající do samostatných tříd:

01bez	podstatné jméno bez
01let1	podstatné jméno let 1. pád
01let2	podstatné jméno let 2. pád
01mezi	podstatné jméno mezi (1. pád mez)
01par	podstatné jméno pár
01pri	podstatné jméno při (1. pád pře)
01set	podstatní jméno set

Přídavná jména

jednotné číslo

mužský rod:

02mj15	přídavné jméno, mužský rod, jednotné číslo, 1. nebo 5. pád
02maj4	přídavné jméno, mužský rod životný, jednotné číslo, 4. pád
02mnj4	přídavné jméno, mužský rod neživotný, jednotné číslo, 4. pád

mužský a střední rod:

02msj2	přídavné jméno, mužský nebo střední rod, jednotné číslo, 2. pád
02msj3	přídavné jméno, mužský nebo střední rod, jednotné číslo, 3. pád
02msj6	přídavné jméno, mužský nebo střední rod, jednotné číslo, 6. pád
02msj7	přídavné jméno, mužský nebo střední rod, jednotné číslo, 7. pád

střední rod:

02sj145	přídavné jméno, střední rod, jednotné číslo, 1. nebo 4. nebo 5. pád

ženský rod:

02zj15	přídavné jméno, ženský rod, jednotné číslo, 1. nebo 5. pád
02zj2	přídavné jméno, ženský rod, jednotné číslo, 2. pád
02zj36	přídavné jméno, ženský rod, jednotné číslo, 3. nebo 6. pád
02zj4	přídavné jméno, ženský rod, jednotné číslo, 4. pád
02zj7	přídavné jméno, ženský rod, jednotné číslo, 7. pád

množné číslo:

02m2	přídavné jméno, množné číslo, 2. pád
02m3	přídavné jméno, množné číslo, 3. pád
02m6	přídavné jméno, množné číslo, 6. pád
02m7	přídavné jméno, množné číslo, 7. pád

mužský rod:

02mam15	přídavné jméno, mužský rod životný, množné číslo, 1. nebo 5. pád
02mnm15	přídavné jméno, mužský rod neživotný, množné číslo, 1. nebo 5. pád
02mm4	přídavné jméno, mužský rod, množné číslo, 4. pád

ženský rod:

02zm145	přídavné jméno, ženský rod, množné číslo, 1. nebo 4. nebo 5. pád

střední rod:

02sm145	přídavné jméno, střední rod, množné číslo, 1. nebo 4. nebo 5. pád

ostatní:

02jar145	přídavné jméno, vzor jarní, jednotné nebo množné číslo, 1. nebo 4. nebo 5. pád
02ko	slova končící na -ko (statisticko, hospodářsko), většinou následuje spojovník

Zájmena

Osobní:

03ja1	zájmeno já 1. pád
03ja2	zájmeno já 2. pád
03ja3	zájmeno já 3. pád
03ja4	zájmeno já 4. pád
03ja6	zájmeno já 6. pád
03ja7	zájmeno já 7. pád
03ty15	zájmeno ty 1. nebo 5. pád
03ty2	zájmeno ty 2. pád
03ty3	zájmeno ty 3. pád

03ty4	zájmeno ty 4. pád
03ty6	zájmeno ty 6. pád
03ty7	zájmeno ty 7. pád
03on1	zájmeno on 1. pád
03on2	zájmeno on 2. pád
03on3	zájmeno on 3. pád
03on4	zájmeno on 4. pád
03on6	zájmeno on 6. pád
03on7	zájmeno on 7. pád
03ona1	zájmeno ona 1. pád
03ona0	zájmeno ona 2. nebo 3. nebo 6. nebo 7. pád
03ona4	zájmeno ona 4. pád
03ono1	zájmeno ono 1. pád
03my1	zájmeno my 1. pád
03my3	zájmeno my 3. pád
03my7	zájmeno my 7. pád
03vy15	zájmeno vy 1. nebo 5. pád
03vy2	zájmeno vy 2. pád
03vy3	zájmeno vy 3. pád
03vy4	zájmeno vy 4. pád
03vy6	zájmeno vy 6. pád
03vy7	zájmeno vy 7. pád
03oni1	zájmeno oni 1. pád
03oni2	zájmeno oni 2. pád
03oni3	zájmeno oni 3. pád
03oni4	zájmeno oni 4. pád
03oni7	zájmeno oni 7. pád
03ony1	zájmeno ony 1. pád

Přivlastňovací:

03prij2	zájmeno přivlastňovací, jednotné číslo, 2. pád
03prij3	zájmeno přivlastňovací, jednotné číslo, 3. pád
03prij6	zájmeno přivlastňovací, jednotné číslo, 6. pád
03prij7	zájmeno přivlastňovací, jednotné číslo, 7. pád
03prim145	zájmeno přivlastňovací, množné číslo, 1. nebo 4. nebo 5. pád
03prim2	zájmeno přivlastňovací, množné číslo, 2. pád

03prim3	zájmeno přivlastňovací, množné číslo, 3. pád
03prim6	zájmeno přivlastňovací, množné číslo, 6. pád
03prim7	zájmeno přivlastňovací, množné číslo, 7. pád
03primj15	zájmeno přivlastňovací, jednotné číslo, mužský rod, 1. nebo 5. pád
03primaj4	zájmeno přivlastňovací, jednotné číslo, mužský rod životný, 4. pád
03primnj4	zájmeno přivlastňovací, jednotné číslo, mužský rod neživotný, 4. pád
03prizj15	zájmeno přivlastňovací, jednotné číslo, ženský rod, 1. nebo 5. pád
03prizj236	zájmeno přivlastňovací, jednotné číslo, ženský rod, 2. nebo 3. nebo 6. pád
03prizj4	zájmeno přivlastňovací, jednotné číslo, ženský rod, 4. pád
03prizj7	zájmeno přivlastňovací, jednotné číslo, ženský rod, 7. pád
03prisj15	zájmeno přivlastňovací, jednotné číslo, střední rod, 1. nebo 5. pád
03prisj4	zájmeno přivlastňovací, jednotné číslo, střední rod, 4. pád
03primam15	zájmeno přivlastňovací, množné číslo, mužský rod životný, 1. nebo 5. pád
03prism145	zájmeno přivlastňovací, množné číslo, střední rod, 1. nebo 4. nebo 5. pád

Ukazovací:

03ukaj15	zájmeno ukazovací, jednotné číslo, 1. nebo 5. pád
03ukaj2	zájmeno ukazovací, jednotné číslo, 2.pád
03ukaj3	zájmeno ukazovací, jednotné číslo, 3. pád
03ukaj4	zájmeno ukazovací, jednotné číslo, 4. pád
03ukaj6	zájmeno ukazovací, jednotné číslo, 6. pád
03ukaj7	zájmeno ukazovací, jednotné číslo, 7. pád
03ukam15	zájmeno ukazovací, množné číslo, 1. nebo 5. pád
03ukam2	zájmeno ukazovací, množné číslo, 2. pád
03ukam3	zájmeno ukazovací, množné číslo, 3. pád
03ukam4	zájmeno ukazovací, množné číslo, 4. pád
03ukam6	zájmeno ukazovací, množné číslo, 6. pád
03ukam7	zájmeno ukazovací, množné číslo, 7. pád

Tázací:

03taz	zájmeno tázací
03tazj1	zájmeno tázací, jednotné číslo, 1. pád
03tazj2	zájmeno tázací, jednotné číslo, 2. pád
03tazj3	zájmeno tázací, jednotné číslo, 3. pád
03tazj4	zájmeno tázací, jednotné číslo, 4. pád

03tazj6	zájmeno tázací, jednotné číslo, 6. pád
03tazj7	zájmeno tázací, jednotné číslo, 7. pád
03tazm1	zájmeno tázací, množné číslo, 1. pád
03tazm2	zájmeno tázací, množné číslo, 2. pád
03tazm3	zájmeno tázací, množné číslo, 3. pád
03tazm4	zájmeno tázací, množné číslo, 4. pád
03tazm6	zájmeno tázací, množné číslo, 6. pád
03tazm7	zájmeno tázací, množné číslo, 7. pád

Vztažná

03vzt	zájmeno vztažné
03vzt1	zájmeno vztažné, 1. pád
03vzt2	zájmeno vztažné, 2. pád
03vzt3	zájmeno vztažné, 3. pád
03vzt4	zájmeno vztažné, 4. pád
03vzt6	zájmeno vztažné, 6. pád
03vzt7	zájmeno vztažné, 7. pád

Ostatní:

03ost	zájmeno ostatní
03ost15	zájmeno ostatní, 1. nebo 5. pád
03ost2	zájmeno ostatní, 2. pád
03ost3	zájmeno ostatní, 3. pád
03ost4	zájmeno ostatní, 4. pád
03ost6	zájmeno ostatní, 6. pád
03ost7	zájmeno ostatní, 7. pád
03pom	zájmeno pomocné (např. nič)

Zájmena spadající do samostatných tříd:

03co	zájmeno co
03ho	zájmeno ho
03je	zájmeno je
03jeho	zájmeno jeho
03jeji	zájmeno její
03jejich	zájmeno jejich

03jiz	zájmeno již
03kdo	zájmeno kdo
03ktera	zájmeno která
03ktere	zájmeno které
03kteri	zájmeno kteří
03ktery	zájmeno který
03ma	zájmeno má
03me	zájmeno mé
03mi	zájmeno mi
03mu	zájmeno mu
03nas	zájmeno nás
03nich	zájmeno nich
03se	zájmeno se, ses
03si	zájmeno si, sis
03sve	zájmeno své
03take	zájmeno také
03ten	zájmeno ten
03tim	zájmeno tím
03to	zájmeno to
03toho	zájmeno toho
03tom	zájmeno tom
03tu	zájmeno tu
03tve	zájmeno tvé
03vsechny	zájmeno všechny

Číslovky

Základní – jeden, jedna, jedno

04jeden15	číslovka jeden, 1. nebo 5. pád
04jeden2	číslovka jeden, 2. pád
04jeden3	číslovka jeden, 3. pád
04jeden6	číslovka jeden, 6. pád
04jeden7	číslovka jeden, 7. pád
04jedena4	číslovka jeden, rod životný, 4. pád
04jedenn4	číslovka jeden, rod neživotný, 4. pád
04jedna15	číslovka jedna, 1. nebo 5. pád
04jedna2	číslovka jedna, 2. pád

04jedna3	číslovka jedna, 3. pád
04jedna4	číslovka jedna, 4. pád
04jedna6	číslovka jedna, 6. pád
04jedna7	číslovka jedna, 7. pád
04jedno145	číslovka jedno, 1. nebo 4. nebo 5. pád

Základní – dva, dvě

04dva15	číslovka dva, 1. nebo 5. pád
04dva2	číslovka dva, 2. pád
04dva3	číslovka dva, 3. pád
04dva4	číslovka dva, 4. pád
04dva6	číslovka dva, 6. pád
04dva7	číslovka dva, 7. pád
04dve145	číslovka dvě, 1. nebo 4. nebo 5. pád

Základní – tři, čtyři

04tricty15	číslovka tři nebo čtyři, 1. nebo 5. pád
04tricty2	číslovka tři nebo čtyři, 2. pád
04tricty3	číslovka tři nebo čtyři, 3. pád
04tricty4	číslovka tři nebo čtyři, 4. pád
04tricty6	číslovka tři nebo čtyři, 6. pád
04tricty7	číslovka tři nebo čtyři, 7. pád

Základní – ostatní

04zak15	číslovka základní, 1. nebo 5. pád
04zak2	číslovka základní, 2. pád
04zak3	číslovka základní, 3. pád
04zak4	číslovka základní, 4. pád
04zak6	číslovka základní, 6. pád
04zak7	číslovka základní, 7. pád

Násobné

04nas0	číslovka násobná (např. obojí)
04nas15	číslovka násobná, 1. nebo 5. pád (např. jedni, jedny)

04nas2	číslovka násobná, 2. pád (např. dvojího, dvojích)
04nas3	číslovka násobná, 3. pád (např. dvojímu, dvojím)
04nas4	číslovka násobná, 4. pád (např. oboje, trojího)
04nas6	číslovka násobná, 6. pád (např. obojích, obojím)
04nas7	číslovka násobná, 7. pád (např. obojími, dvojím)

Číslovky spadající do samostatných tříd:

04ctvrt	číslovka čtvrt
04dva	číslovka dva
04par	číslovka pár
04prvni	číslovka první
04pul	číslovka půl
04set	číslovka set
04sta	číslovka sta
04ste	číslovka stě
04sto	číslovka sto
04tisic	číslovka tisíc
04tri	číslovka tři

Ostatní číslovky:

04era	číslovka *era
04erem	číslovka *erem
04ero2	číslovka *ero s 2. pádem (např. devatero pohádek)
04ero7	číslovka *ero se 7. pádem (např. devatero pohádkami)
04erou4	číslovka *erou se 4. pádem
04erou7	číslovka *erou se 7. pádem
04eru	číslovka *eru
04ery1	číslovka *ery s 1. pádem (např. devatery kalhoty)
04ery4	číslovka *ery se 4. pádem
04krat	číslovka násobná (např. dvakrát)
04ost1	číslovka ostatní, 1. pád (např. několik)
04ost0	číslovka ostatní (např. několika)
04ot	číslovka – otázka
04ot0	číslovka – otázka, 2. nebo 3. nebo 6. nebo 7. pád
04po	číslovka s předponou po (např. podruhé)

Slovesa

Tvary slovesa být

05budu	sloveso bude nebo nebude
05budes	sloveso budeš nebo nebudeš
05bude	sloveso bude nebo nebude
05budeme	sloveso budeme nebo nebudeme
05budete	sloveso budete nebo nebudete
05budou	sloveso budou nebo nebudou
05bych	sloveso bych
05bys	sloveso bys
05by	sloveso by
05bychom	sloveso bychom
05byste	sloveso byste
05byl	sloveso byl nebo nebyl
05byla	sloveso byla nebo nebyla
05bylo	sloveso bylo nebo nebylo
05byli	sloveso byli nebo nebyli
05byly	sloveso byly nebo nebyly
05byt	sloveso být
05jsem	sloveso jsem nebo nejsem
05jsi	sloveso jsi nebo nejsi
05je	sloveso je
05jsme	sloveso jsme nebo nejsme
05jste	sloveso jste nebo nejste
05jsou	sloveso jsou nebo nejsou

Přítomný čas

05inf	sloveso, infinitiv
05pj1	sloveso, 1. osoba, přítomný čas, jednotné číslo
05pj2	sloveso, 2. osoba, přítomný čas, jednotné číslo
05pj3	sloveso, 3. osoba, přítomný čas, jednotné číslo
05pm1	sloveso, 1. osoba, přítomný čas, množné číslo
05pm2	sloveso, 2. osoba, přítomný čas, množné číslo
05pm3	sloveso, 3. osoba, přítomný čas, množné číslo

Minulý čas

05mjm	sloveso, minulý čas, jednotné číslo, mužský rod
05mjs	sloveso, minulý čas, jednotné číslo, střední rod
05mjz	sloveso, minulý čas, jednotné číslo, ženský rod, nebo množné číslo střední rod
05mmm	sloveso, minulý čas, množné číslo, mužský rod
05mmz	sloveso, minulý čas, množné číslo, ženský rod

Příčestí trpné

05tjm	sloveso, příčestí trpné, jednotné číslo, mužský rod
05tjs	sloveso, příčestí trpné, jednotné číslo, střední rod
05tjz	sloveso, příčestí trpné, jednotné číslo, ženský rod nebo množné číslo, střední rod
05tjz4	sloveso, příčestí trpné, jednotné číslo ženský rod, 4. pád
05tmm	sloveso, příčestí trpné, množné číslo, mužský rod
05tmz	sloveso, příčestí trpné, množné číslo, ženský rod

Rozkazovací způsob a přechodník

05rj2	sloveso, rozkazovací způsob, jednotné číslo, 2. osoba
05rm1	sloveso, rozkazovací způsob, množné číslo, 1. osoba
05rm2	sloveso, rozkazovací způsob, množné číslo, 2. osoba
05prech	sloveso, přechodník

Slovesa spadající do samostatných tříd

05ma	sloveso má
05maji	sloveso mají
05mel	sloveso měl
05muze	sloveso může
05neni	sloveso není
05pri	sloveso při
05set	sloveso set
05tri	sloveso tři

Příslovce

Druh

06cas	příslovce času
06mist	příslovce místa
06zp	příslovce způsobu
06poc	příslovce – počet (např. hodně)
06ot	příslovce otázky
06s2	příslovce 2. stupeň
06s3	příslovce 3. stupeň
06	příslovce ostatní

Příslovce spadající do samostatných tříd

06dnes	příslovce dnes
06jak	příslovce jak
06jeste	příslovce ještě
06jiz	příslovce již
06kde	příslovce kde
06kdy	příslovce kdy
06pak	příslovce pak
06podle	příslovce podle
06tak	příslovce tak
06take	příslovce také
06tedy	příslovce tedy
06treba	příslovce třeba
06tu	příslovce tu
06uz	příslovce už
06vice	příslovce více

Předložky

Rozdělení předložek podle pádu, se kterým se pojí

071	předložka s 1. pádem
072	předložka s 2. pádem
073	předložka se 3. pádem
074	předložka se 4. pádem

077	předložka se 7. pádem
07p	slovo před předložkou (např. vzhledem)

Předložky spadající do samostatných tříd

07bez2	předložka bez s 2. pádem
07do2	předložka do s 2. pádem
07k3	předložka k se 3. pádem
07ke3	předložka ke se 3. pádem
07mezi4	předložka mezi se 4. pádem
07mezi7	předložka mezi se 7. pádem
07na4	předložka na se 4. pádem
07na6	předložka na se 6. pádem
07nad4	předložka nad se 4. pádem
07nad7	předložka nad se 7. pádem
07o4	předložka o se 4. pádem
07o6	předložka o se 6. pádem
07od2	předložka od s 2. pádem
07po4	předložka po se 4. pádem
07po6	předložka po se 6. pádem
07pod4	předložka pod se 4. pádem
07pod7	předožka pod se 7. pádem
07podle2	předložka podle s 2. pádem
07pred4	předložka před se 4. pádem
07pred7	předložka před se 7. pádem
07pri6	předložka při se 6. pádem
07pro4	předložka pro se 4. pádem
07s7	předložka s se 7. pádem
07se7	předložka se se 7. pádem
07u2	předložka u s 2. pádem
07v4	předložka v se 4. pádem
07v6	předložka v se 6. pádem
07ve4	předložka ve se 4. pádem
07ve6	předložka ve se 6. pádem
07z2	předložka z s 2. pádem
07za2	předložka za s 2. pádem
07za4	předložka za se 4. pádem

07za7	předložka za se 7. pádem
07ze2	předložka ze s 2. pádem

Spojky

08	spojka
08byj	spoka *by jednotné číslo (např. abych)
08bym	spoka *by množné číslo (např. kdybyste)

Spojky spadající do samostatných tříd

08a	spojka a
08aby	spojka aby
08ale	spojka ale
08ani	spojka ani
08az	spojka až
08i	spojka i
08jako	spojka jako
08kdyby	spojka kdyby
08kdyz	spojka když
08krat	spojka – krát
08nebo	spojka nebo
08nez	spojka než
08proto	spojka proto
08protoze	spojka protože
08vsak	spojka však
08ze	spojka že

Citoslovce 09

Částice

10	částice
10jen	částice jen
10li	částice li

Vlastní jména

11j1	vlastní jméno, jméno, 1. pád
11j2	vlastní jméno, jméno, 2. pád
11j3	vlastní jméno, jméno, 3. pád
11j4	vlastní jméno, jméno, 4. pád
11j5	vlastní jméno, jméno, 5. pád
11j6	vlastní jméno, jméno, 6. pád
11j7	vlastní jméno, jméno, 7. pád
11n1	vlastní jméno, název, 1. pád
11n2	vlastní jméno, název, 2. pád
11n3	vlastní jméno, název, 3. pád
11n4	vlastní jméno, název, 4. pád
11n5	vlastní jméno, název, 5. pád
11n6	vlastní jméno, název, 6. pád
11n7	vlastní jméno, název, 7. pád
11neskl	vlastní jméno nesklonné

Interpunkce

12!	vykřičník
12,	čárka
12.	tečka
12?	otazník
12ost	interpunkce – ostatní

Neznámá slova 13nez

Začátek věty 00bol

Jindra Drábková

Tvorba jazykového modelu založeného na třídách
Autoreferát dizertační práce

Technická univerzita v Liberci
Fakulta mechatroniky a mezioborových inženýrských studií

20 stran
Náklad 20 výtisků

Liberec 2005

TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových inženýrských studií

TVORBA JAZYKOVÉHO MODELU ZALOŽENÉHO NA TŘÍDÁCH

Autoreferát dizertační práce

Jindra Drábková

Liberec 2005

Tvorba jazykového modelu založeného na třídách

Autoreferát dizertační práce

Ing. Jindra Drábková

Studijní program: P2612 Elektrotechnika a informatika
Studijní obor: 2612V045 Technická kybernetika

Pracoviště: Katedra elektroniky a zpracování signálů
Fakulta mechatroniky a mezioborových inženýrských studií
Technická univerzita v Liberci
Hálkova 6, 461 17 Liberec

Školitel: Prof. Ing. Jan Nouza, CSc.

Rozsah dizertační práce a příloh

Počet stran: 123
Počet obrázků: 25
Počet tabulek: 21
Počet vzorců: 61
Počet příloh: 1

Abstrakt

Rozpoznávání spojité řeči je komplexní problém sestávající z několika úloh. Jednou z těchto úloh je tvorba jazykového modelu. Český jazyk patří mezi jazyky ohebné, což s sebou nese řadu nevýhod. Jednou z nich je velké množství slov, které tvorbu jazykového modelu komplikuje. Dizertační práce předkládá řešení, ve kterém se v jazykovém modelu použijí gramatické značky místo slov. Ke stanovení značek byly využity tři přístupy – pravděpodobnostní, statistický a gramatický.

Téma dizertační práce zasahuje do několika oblastí od počítačové lingvistiky, morfologické analýzy po tvorbu korpusu a slovníku. Všechny vyjmenované oblasti by mohly být zahrnuty pod společný název počítačové zpracování jazyka. Tato disciplína je základem pro řadu dalších odvětví, jako je např. strojový překlad, větný rozbor nebo rozpoznávání řeči.

K tvorbě bigramového jazykového modelu značek byl vytvořen vlastní označkovaný korpus. Ten byl označkován částečně ručně a z větší části automaticky. Pro automatické značkování byl navržen stochastický značkovač, který využívá označkovaný slovník obsahující přibližně 300 tisíc různých slovních tvarů. Značky byly do slovníku přidány jednak ručně a jednak na základě syntaktické metody.

Z označkovaných dat bylo vytvořeno několik bigramových jazykových modelů značek vytvořených z vět s interpunkcí i bez interpunkce. Všechny modely značek byly testovány v závislosti na velikosti slovníku a výsledky testování byly zhodnoceny. Nejlepší jazykový model značek byl použit pro experimenty se systémem pro rozpoznávání spojité řeči.

Abstract

Speech recognition is a complex challenge which consists of several tasks. Language modeling is one of these tasks. The Czech language belongs to a group of languages which can be termed as flexible languages. One of the greatest disadvantages of such flexible languages is the large number of words. This PhD thesis submits a solution to this disadvantage. It is to use the grammatical tags instead of the words. To determine these tags three different approaches were used – statistical, grammatical and stochastic.

PhD thesis theme includes several branches – computational linguistics, morphologic analyses, corpora building and vocabulary building. All of these branches could be called natural language processing. This creates a disciplinary foundation, for example, for machine translation, parsing or speech recognition.

A bigram class-based language model was built using tagged corpus. The tagging of the corpus was completed in part manually, and in part automatically. A stochastic tagger was devised to automatic tagging using tagged vocabulary which includes some 300,000 items. Tags were added to this vocabulary both manually and with using syntactic method.

Using tagged corpus it was possible to design a number of bigram class-based language models: unsmoothed, smoothed by linear interpolation, made from sentences both with and without punctuation. Evaluations were made of the effects of both punctuation and the size of vocabulary. The best bigram class-based language model was then used in experiments with the continuous speech recognizer.

Obsah

1 Úvod

Cílem dizertační práce je vytvoření jazykového modelu založeného na třídách pro rozpoznávač spojité řeči a jeho praktické využití.

V současné době jsou nejrozšířenější jazykové modely založené na statistických informacích získaných z korpusu. Na základě takových jazykových modelů je možno s určitou pravděpodobností předpovídat následující slovo, čehož se využívá nejen v rozpoznávání řeči, ale i v rozpoznávání rukou psaného textu, v detekci pravopisných chyb apod.

Práce je rozdělena na dvě části, teoretickou a praktickou. Teoretická část popisuje úlohu rozpoznávání řeči, tvorbu akustického a jazykového modelu. V této části jsou uvedeny také metody vyhlazování jazykového modelu, metriky používané k ohodnocení systémů rozpoznávání řeči, jsou zde vysvětleny pojmy korpus, slovník a značkování.

Praktická část se zabývá stanovením gramatických značek pro český jazyk, tvorbou označkovaného korpusu a slovníku. Jsou zde uvedeny návrhy stochastických značkovačů, které byly použity pro automatické označkování velkého množství dat, a prezentovány výsledky automatického značkování s bigramovým jazykovým modelem založeným na třídách.

Cíle práce lze shrnout do těchto bodů:

1. Stanovení gramatických značek pro český jazyk.
2. Vytvoření vlastního označkovaného slovníku.
3. Návrh a realizace různých automatických značkovačů.
4. Vyhodnocení vytvořených značkovačů na testovacích datech.
5. Automatické značkování velkého množství dat pomocí nejlepšího značkovače.
6. Tvorba různých bigramových jazykových modelů založených na třídách.
7. Testování jazykových modelů v závislosti na interpunkci a velikosti slovníku.
8. Využití nejlepšího bigramového jazykového modelu založeného na třídách při rozpoznávání spojité řeči.

2 Statistický přístup k rozpoznávání souvislé řeči

Rozpoznání řeči je proces, při kterém akustický signál snímaný např. mikrofonem generuje posloupnost slov.

Při řešení úlohy rozpoznávání spojité řeči se v současné době nejčastěji využívá statistický přístup. Předpokládejme, že $W = \{w_1, w_2, w_3, ..., w_N\}$ je posloupnost N slov a $O = \{o_1, o_2, o_3, ..., o_M\}$ je akustická informace odvozená z řečového signálu. Cílem je nalézt nejpravděpodobnější posloupnost slov $\hat{W}$ pro danou akustickou informaci O:

$$\hat{W} = \arg\max_W p(W \mid O) \tag{2.1}$$

kde $p(W \mid O)$ je podmíněná pravděpodobnost, že vyslovená posloupnosti W odpovídá akustické informaci O,
funkce arg max v tomto vztahu znamená nalezení posloupnosti W takové, pro kterou je $p(W \mid O)$ maximální.

V případě, že použijeme Bayesovo pravidlo, platí:

$$\hat{W} = \arg\max_W \frac{p(W)p(O \mid W)}{p(O)} \tag{2.2}$$

kde $p(W)$ je pravděpodobnost vyslovené posloupnosti W,
$p(O)$ je pravděpodobnost akustické informace O,
$p(O \mid W)$ je podmíněná pravděpodobnost, že akustická informace odpovídá vyslovené posloupnosti.

Pro výpočet nejpravděpodobnější posloupnosti slov $\hat{W}$ pro danou akustickou informaci O se používá vztah (2.2). Vzhledem k tomu, že se provádí maximalizace přes všechna slova a $p(O)$ není funkcí W, lze tuto pravděpodobnost při hledání maxima ignorovat. Potom:

$$\hat{W} = \arg\max_W p(W \mid O) = \arg\max_W p(W)p(O \mid W) \tag{2.3}$$

Úloha rozpoznávání spojité řeči může být tedy rozdělena do čtyř dílčích úloh [PSUTKA 1995]:

1. akustické zpracování řečového signálu,
2. vytvoření akustického modelu $p(O \mid W)$,
3. vytvoření jazykového modelu $p(W)$,
4. nalezení nejpravděpodobnější posloupnosti slov.

2.1 Tvorba jazykového modelu

Úkolem jazykového modelu je stanovit jistá omezení a nalézt určitá pravidla, pomocí nichž můžeme ze slov vytvořit větu. Omezení a pravidla vycházejí z vlastností konkrétního jazyka a mohou být modelována jak stochastickými tak i nestochastickými metodami. Stochastické jazykové modely (SLM) používají pro jazykové modelování pravděpodobnostní přístup. Tyto jazykové modely přiřazují každé posloupnosti slov $W = \{w_1, w_2, w_3, ..., w_n\}$ pravděpodobnost $p(W)$, kterou je třeba odhadnout z dat. Data, která se používají pro tvorbu modelu, se nazývají trénovací data. Nejrozšířenější SLM je n-gramový jazykový model. Podle „řetězového pravidla" pravděpodobnosti platí:

$$p(W) = p(w_1, w_2, w_3, ..., w_n) = p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2)...p(w_n \mid w_1, w_2, ..., w_{n-1}) =$$
$$= \prod_{i=1}^{n} p(w_i \mid w_1, w_2, ..., w_{i-1}) \tag{2.4}$$

Obecně lze pro odhad pravděpodobnosti výskytu slova použít n-gramový model slov, tj. $p(w_n \mid w_1, w_2, w_3, ..., w_{n-1})$. V praxi je nemožné tuto pravděpodobnost vypočítat, protože pro slovník o rozměru V a pro n-té slovo ve větě existuje V^{n-1} různých možností historií, což znamená, že ještě před samotným výpočtem posloupnosti je třeba zjistit celkem V^n různých pravděpodobností. Prvky n-gramového modelu jsou podmíněné pravděpodobnosti, které se rovnají pravděpodobnosti toho, že bude následovat jisté slovo w_n v případě, že nastala vstupní kombinace $w_1, w_2, w_3, ..., w_{n-1}$. V praxi by to znamenalo generovat obrovské množství dat, a proto se tento problém řeší aproximací. U n-gramového jazykového modelu předpokládáme, že je pravděpodobnost slova daná všemi předchozími slovy $p(w_n|w_1, w_2, w_3, ..., w_{n-1})$. Jestliže slovo závisí na předchozích dvou slovech, mluvíme o trigramu $p(w_n|w_{n-2}, w_{n-1})$, podobně o bigramu, kdy dané slovo závisí pouze na předchozím slově $p(w_n|w_{n-1})$ nebo unigramu $p(w_n)$.

Prvky n-gramové matice jsou podmíněné pravděpodobnosti:

$$p(w_n \mid w_1, w_2, w_3, ..., w_{n-1}) = \frac{p(w_1, w_2, w_3, ..., w_n)}{p(w_1)p(w_2 \mid w_1)...(p(w_{n-1} \mid w_1, w_2, w_3, ..., w_{n-2})} \quad (2.5)$$

V praxi se nejčastěji používá bigramový nebo trigramový jazykový model. Jestliže pravděpodobnosti vyjádříme pomocí četností, které získáme z trénovacích dat, pak pro bigram platí:

$$p(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)}{C(w_{n-1})} \quad (2.6)$$

kde $C(w_{n-1})$ je počet výskytů slova w_{n-1},
$C(w_{n-1}, w_n)$ je počet výskytů dvojic slov w_{n-1}, w_n.

Hodnoty pravděpodobností jsou z definice menší než jedna a pro velké množství dat jsou velice malé. Z tohoto důvodu mnoho programů používá pro výpočet podmíněné pravděpodobnosti bigramů logaritmy těchto pravděpodobností.

Základní (nevyhlazený) bigramový jazykový model je matice, jejíž prvky jsou podmíněné pravděpodobnosti určené pro všechny možné dvojice sousedních slov, které se objeví v trénovacích datech. Posloupnosti slov, které se v trénovacích datech neobjeví, mají hodnotu pravděpodobnosti rovnu nule. Od takto vytvořeného jazykového modelu se odvozují všechny vyhlazovací metody.

2.2 Vyhlazování jazykového modelu

Vyhlazování jazykového modelu se používá z důvodu velkého počtu nulových hodnot, které se vyskytují v bigramové matici. Každý trénovací korpus je konečný a nemůže obsahovat všechny dvojice slov. Nulová hodnota se v matici může objevit v případě, že se daná posloupnost slov nebo dané slovo v trénovacím korpusu neobjevily. V testovacím korpusu se ale objevit může. Proto se používají vyhlazovací algoritmy, které nulovým hodnotám v matici přiřadí malé nenulové pravděpodobnosti.

V dizertační práci se využívají dva typy vyhlazování Add-One Smoothing a Linear Interpolation Smoothing.

- **Add-One Smoothing**

Tento typ vyhlazování je nejjednodušší vyhlazovací technikou. V případě unigramů se k počtu všech slov včetně těch, které se v textu nevyskytly, přičte 1. Nechť N je počet všech slovních tvarů, V je počet slov ve slovníku, $C(w)$ je počet výskytů slova w a $C(w_{n-1}, w_n)$ je počet výskytů dvojic slov w_{n-1}, w_n.

Potom pro nevyhlazený unigram platí:

$$p(w) = \frac{C(w)}{N} \quad (2.7)$$

Pravděpodobnost pro vyhlazený unigram se vypočítá podle:

$$p_{+1}(w) = \frac{C(w)+1}{N+V} \quad (2.8)$$

Obdobně platí pro nevyhlazený bigram:

$$p(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)}{C(w_{n-1})} \quad (2.9)$$

Pravděpodobnost pro vyhlazený bigram se vypočítá podle:

$$p_{+1}(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)+1}{C(w_{n-1})+V} \quad (2.10)$$

Nevýhodou je fakt, že pro viděné dvojice slov je pravděpodobnost „podhodnocená" a pro neviděné dvojice slov je pravděpodobnost „nadhodnocená".

- **Linear Interpolation Smoothing**

Vyhlazování nazvané lineární interpolace (Linear Interpolation Smoothing) používá pro odhad podmíněných pravděpodobností $\hat{p}$ vyšších úrovní vždy všech n-gramů nižších úrovní. Každý člen je vážen lineárním koeficientem λ_i. Počet slov ve slovníku je V.

V případě trigramů se používá vzorec:

$$\hat{p}(w_n \mid w_{n-1}, w_{n-2}) = \lambda_3 p(w_n \mid w_{n-1}, w_{n-2}) + \lambda_2 p(w_n \mid w_{n-1}) + \lambda_1 p(w_n) + \lambda_0 / V \quad (2.12)$$

Analogický vztah platí pro bigramy:

$$\hat{p}(w_n \mid w_{n-1}) = \lambda_2 p(w_n \mid w_{n-1}) + \lambda_1 p(w_n) + \lambda_0 / V \quad (2.13)$$

Hodnoty podmíněných pravděpodobností trigramů, bigramů a unigramů se stanoví z trénovacích dat. Pro odhad koeficientů λ_i se používají „odložená data" (data oddělená od hlavní trénovací množiny). Při použití trénovacích dat pro odhad koeficientů je koeficient u trigramů λ_3 (resp. bigramů λ_2) roven jedné a ostatní koeficienty jsou nulové. Koeficienty jsou stanoveny tak, aby maximalizovaly pravděpodobnost odložené části dat $p(W_H)$:

$$p(W_H) = \prod_{i=1..H} p(w_i \mid w_{i-1}) \quad (2.14)$$

kde H je počet slov v odložených datech

Pro výpočet koeficientů λ_i se používá EM algoritmus (Expectation – Maximization Algorithm), jehož postup pro bigramy je popsán např. v [MRVA 2000].

2.3 Jazykový model založený na třídách

Jazykový model založený na třídách (Class-Based Language Model) určuje závislosti mezi třídami (značkami) slov a mezi značkami a slovy místo závislostí mezi konkrétními slovy. Pro tvorbu takového jazykového modelu je třeba jednotlivým slovům přiřadit značky (slova zařadit do tříd). Značky jsou většinou stanoveny na základě syntaktických a sémantických vlastností slov.

Předpokládejme, že existuje mapovací funkce G, která přiřazuje každému slovu w_n v korpusu značku c_n.

$$G : c_n = G(w_n) \quad (2.15)$$

Trénovací množina slov (w_1, w_2, …, w_T) se tak rozšíří na množinu dvojic – slovo a příslušná značka: ($\langle w_1, G(w_1)\rangle$, $\langle w_2, G(w_2)\rangle$, …, $\langle w_T, G(w_T)\rangle$).

Jestliže je definována mapovací funkce, je možné nahradit slova značkami. Pak v případě bigramů pro pravděpodobnost $p(W)$, kde $W = \{w_1, w_2, w_3, ..., w_n\}$, platí [JURAFSKY 2000]:

$$p(W) = \prod_{i=1}^{n} p(w_i \mid c_i) \cdot p(c_i \mid c_{i-1}) \qquad (2.16)$$

kde $p(w_i \mid c_i)$ je podmíněná pravděpodobnost, s jakou bude k dané značce přiřazeno dané slovo,
$p(c_i \mid c_{i-1})$ je bigram značek.

Jazykový model lze přepsat následujícím způsobem:

$$p(w_n \mid w_1^{n-1}) = p(w_n \mid c_n) p(c_n \mid c_1^{n-1}) \qquad (2.17)$$

kde w_1^{n-1} je historie slov,
c_1^{n-1} je historie značek.

Pro bigramový jazykový model založený na třídách potom platí:

$$p(w_n \mid w_{n-1}) = p(w_n \mid c_n) p(c_n \mid c_{n-1}) \qquad (2.18)$$

Tento jazykový model sestává ze dvou složek:

- bigram značek

$$p(c_n \mid c_{n-1}) = \frac{C(c_n, c_{n-1})}{C(c_{n-1})} \qquad (2.19)$$

kde $C(c_n, c_{n-1})$ je počet výskytů dvojice značek c_n, c_{n-1},
$C(c_{n-1})$ je počet výskytů značky c_{n-1}.

- podmíněná pravděpodobnost, s jakou bude k dané značce přiřazeno dané slovo

$$p(w_n \mid c_n) = \frac{C(w_n, c_n)}{C(c_n)} \qquad (2.20)$$

kde $C(c_n)$ je počet výskytů značky c_n,
$C(w_n, c_n)$ je počet současného výskytu slova w_n se značkou c_n, pokud je jednomu slovu přiřazena jen jedna značka, pak $C(w_n, c_n) = C(w_n)$.

3 Tvorba slovníku

Pro zpracování řeči i textu je třeba mít k dispozici vhodný slovník. Slovník je seznam typů slov, což jsou odlišné položky slovníku. To znamená, že slova pán a pánovi jsou dvě odlišné položky slovníku se stejným jazykovým kmenem. Slovník, který je používán v této práci, byl vytvořen na Technické univerzitě v Liberci z internetové verze článků z Lidových novin, z internetových novin Neviditelný pes, ze 4 diplomových prací a 27 novel. Kromě typu slova obsahuje i jeho četnost. Pro rozpoznávání řeči je nutné přidat do slovníku transkripci, pro zpracování textu lze do slovníku přidat další informaci o typu slova. Vzhledem k tomu, že úkolem práce bylo vytvoření jazykového modelu založeného na třídách (Class-Based Language Model), je ke každému typu slova přidána do slovníku informace o příslušných gramatických značkách, které lze danému typu slova přiřadit.

Pro přiřazení příslušných značek ke každému slovu ve slovníku byly postupně použity tři metody (přístupy). Nejprve byly přidány do slovníku značky těch slov, které byly obsaženy ve větách označkovaných ručně. Tato slova tvořila jen velmi malou část slovníku (4 %). Proto byla použita syntetická metoda založená na generování všech možných slovních tvarů s jejich morfologickými značkami. Tvary byly generovány s použitím pravidel pro tvorbu slov a s použitím množiny kořenů, předpon, přípon a koncovek (poskytnuto Ústavem formální a aplikované lingvistiky Univerzity Karlovy v Praze). Touto metodou bylo označkováno

dalších asi 49 % slov ze slovníku. Třetí přístup byl částečně manuální. Pomocí filtrů s typickými předponami, příponami a koncovkami byly přiřazeny příslušné značky k dalším slovům ve slovníku. Některým slovům ve slovníku byly příslušné značky přiřazeny ručně.

3.1 Příprava dat

Pro tvorbu jazykového modelu byl vytvořen korpus z novinových článků, článků z internetu, z částí knížek apod. Vytvořený korpus neobsahuje odborné a vědecké články. Všechna data jsou převedena na prostý text a splňují níže uvedené požadavky.

- Každá věta je na jednom řádku.
- Jednotlivá slova včetně interpunkce jsou oddělena mezerou.
- V korpusu nejsou použita čísla a zkratky zakončené tečkou – obojí je přepsáno do slov.
- Spojovník je od slov oddělen mezerami.
- Velká a malá písmena jsou v korpusu ponechána beze změny stejně jako další zkratky (např. ODS).
- Každá věta je vždy ukončena tečkou, vykřičníkem, otazníkem nebo uvozovkami.
- Věty neobsahují nespisovná slova (např. hladovej) a gramaticky nesprávná spojení (např. ty děvčata).

Vytvořený korpus obsahuje celkem 8 800 vět (130 718 slov včetně interpunkce). Z toho 3 300 vět sloužilo jako trénovací data a 500 vět jako testovací data. Pro další experimenty bylo upraveno 5 000 vět. V tabulce 3.1 jsou uvedeny počty slov ve větách s interpunkcí a bez interpunkce, procento interpunkce a procento OOV (out of vocabulary – slov, která nejsou ve slovníku).

Tab. 3.1: Statistika korpusu

počet vět	trénovací data	testovací data	data pro automatické značkování	celkem
	3 300	500T	5 000	8 800
počet slov s interpunkcí	44 331	8 867	77 520	130 718
počet slov bez interpunkce	38 084	7 836	67 440	113 360
interpunkce [%]	14,09	11,63	13,00	13,28
OOV [%]	0	0,34	1,65	1,02

4 Značkování a značkovače

Značkování (Part-of-Speech Tagging) je přiřazení gramatické značky (tagu) každému slovu a obvykle i interpunkčnímu znaménku v korpusu. Značkovače se používají např. v rozpoznávání řeči a syntaktické analýze. Vstupem do značkovače je řetězec slov (věta) a množina značek a výstupem je posloupnost značek, kdy pro každé slovo je vybrána nejpravděpodobnější značka. Značky lze přiřazovat ručně nebo automaticky. Přiřazení značky není vždy jednoznačné. Desambiguace (zjednoznačnění) je velmi obtížný problém. Milióny slov nelze značkovat ručně a prakticky není možné se obejít bez chyb. Podle [ČERMÁK 2004] dosahují nejlepší programy pro desambiguaci aplikované na korpusy angličtiny 97–98% úspěšnosti. Úspěšnost morfologické desambiguace korpusu SYN2000 (Český národní korpus) dosahuje zhruba 94 % [ČERMÁK 2004]. Uvedený rozdíl vyplývá

zejména z odlišných typologických vlastností češtiny a angličtiny. Angličtina je jazyk s poměrně velmi pevným slovosledem, takže se jak pravděpodobnostními metodami založenými na četnosti posloupností slov a jejich značek tak i nepravděpodobnostními metodami založenými na pravidlech značkuje mnohem úspěšněji.

Existují různé přístupy ke značkování textu. Nejznámější značkovače (taggers) jsou značkovače založené na pravidlech (rule-based taggers), stochastické značkovače (stochastic taggers) a značkovače, které kombinují tyto dva způsoby (hybrid taggers).

V dizertační práci jsou předloženy návrhy stochastických značkovačů, které jsou založeny na Stochastic Parts Program [CHURCH 1988]. Ten maximalizuje pravděpodobnost *p(značka|slovo)* · *p(značka|předchozích n značek)*. Pro bigramový značkovač potom platí:

$$c_i = \arg\max_j p(c_i \mid w_j) \cdot p(c_j \mid c_{i-1}) \qquad (4.1)$$

Funkce arg max v rovnici (4.1) znamená nalezení takového *j*, pro které je součin uvedených podmíněných pravděpodobností maximální.

Pro automatické značkování vět jsme vytvořili v programovacím jazyku Perl stochastické značkovače, které pro nalezení nejlepší posloupnosti značek pro danou posloupnost slov využívají dynamické programování. Značky odpovídající jednotlivým slovům jsou zobrazeny jako uzly grafu. Hrany (spojnice uzlů) jsou ohodnoceny prvky nevyhlazené nebo vyhlazené bigramové matice. Cílem značkování je najít optimální cestu grafem.

5 Stanovení značek

Nevýhodou tvorby jazykového modelu i z velkého korpusu je nedostatek dat. Není možné, aby se v korpusu vyskytla všechna slova a slovní spojení. Jedním z řešení je seskupení podobně se chovajících slov do tříd. Tím získáme reálný odhad i pro slovní spojení, která se dosud v korpusu nevyskytla, ale ztratíme přesnost při nahrazení slova značkou (třídou).

Při stanovení značek byl využit přístup gramatický, statistický i pravděpodobnostní. První fáze zahrnovala stanovení značek podle slovních druhů a jejich gramatických kategorií. Při takto stanovených značkách může být k jednomu slovu přiřazeno až několik desítek značek. Jejich celkový počet se pohybuje okolo 500.

Vzhledem k tomu, že by značkování mělo být využito k tvorbě jazykového modelu založeného na třídách pro rozpoznávání spojité řeči, jsme se snažili dodržet tři zásady:

- stanovit co nejmenší počet značek,
- nejfrekventovanějším slovům přiřadit samostatné značky,
- dodržet, aby k jednomu slovu bylo přiřazeno nejvýše deset značek.

V druhé fázi byla vybrána slova s největší frekvencí a každému takovému slovu byla přiřazena samostatná značka. Tato slova se vyskytují v textu tak často, že tvoří přibližně 30 % jakéhokoli textu.

Poslední fází bylo seskupení některých značek stanovených v první fázi tak, aby byly seskupeny značky s podobnými gramatickými nebo syntaktickými vlastnostmi. Ke slučování byl použit program vytvořený v programovacím jazyku Perl, který realizuje hladový algoritmus (Greedy Algorithm). Program ze zadaného textu postupně seskupuje slova do jednotlivých tříd. V případě velkého počtu dat je možné stanovit minimální četnost slov pro slučování. Program končí, když jsou všechna slova seskupena do jedné třídy nebo když je

dosaženo předurčeného počtu tříd, který lze opět stanovit. Výstupem je binární strom postupu slučování a vzájemná informace vypočítaná na začátku slučování.

Při finálním stanovení značek jsme dodrželi všechny tři zásady uvedené v úvodu kapitoly, využili jsme gramatický i statistický přístup a vzali jsme v úvahu výsledky získané na základě hladového algoritmu. Seznam všech značek včetně jejich popisu je uveden v příloze dizertační práce.

6 Postup značkování korpusu

Pro experimenty bylo ručně nebo poloautomaticky označkováno 3 800 vět, z toho 500 vět bylo použito pro testování. Pro ruční značkování vět byl vytvořen program, který pro každé slovo ve větě nabídl ze slovníku příslušné gramatické značky. U slov, která ve slovníku chyběla, byla přiřazena značka ručně. Z 800 manuálně označkovaných vět byla vytvořena nevyhlazená bigramová matice, jejíž prvky byly přirozené logaritmy podmíněných pravděpodobností. Tato matice byla použita pro poloautomatické značkování, při kterém program označil u každého slova nejpravděpodobnější značku. U nesprávně určených značek byla provedena ruční oprava. U slov, která ve slovníku chyběla, byla přiřazena značka ručně. Na základě takto ručně označkovaných vět byl vytvořen bigramový jazykový model značek.

Všechny experimenty byly prováděny se třemi různě velkými slovníky a jazykové modely byly vytvářeny z vět s interpunkcí i bez interpunkce. Pro automatické značkování byly vytvořeny tři různé stochastické značkovače založené na Stochastic Parts Program (viz kapitola 4).

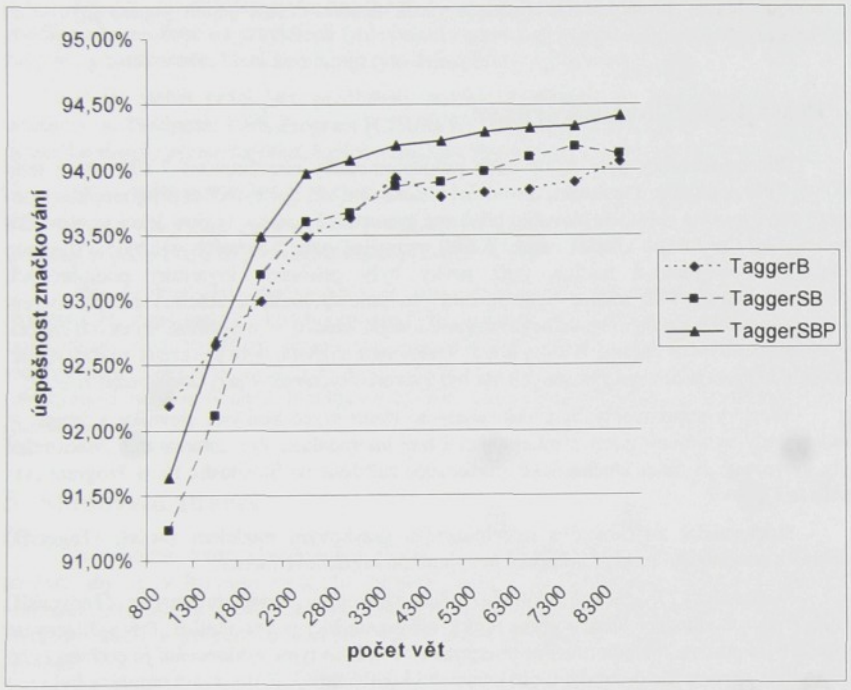
Stochastický značkovač s nevyhlazeným jazykovým modelem značek (TaggerB) používá k ohodnocení hran grafu prvky nevyhlazené bigramové matice.

Stochastický značkovač s vyhlazeným jazykovým modelem značek (TaggerSB) používá pro ohodnocení hran v grafu prvky vyhlazené bigramové matice. Pro vyhlazování modelu byla použita metoda lineární interpolace. U tohoto typu vyhlazování je potřeba držet nějaká data stranou (held-out data) pro stanovení koeficientů λ. Z trénovací množiny bylo pro tento účel vyřazeno 300 ručně označkovaných vět.

Stochastický značkovač TaggerSBP používá k ohodnocení hran grafu prvky vyhlazené bigramové matice a k ohodnocení uzlů pravděpodobnost výskytu daného slova v dané třídě $p(c_n \mid w_n)$. Označkované věty, ze kterých je tato pravděpodobnost vypočítána, jsou uloženy v souboru ve vertikálním formátu, což znamená, že každé slovo věty je na jednom řádku. Kromě slova je na řádku odpovídající značka, která je od slova oddělena tabulátorem. Pro výpočet podmíněné pravděpodobnosti $p(c_n \mid w_n)$ jsou všechna první slova ve větě převáděna na malá písmena. V případě, že se slovo w_n v trénovacích datech neobjeví, ale ve slovníku obsaženo je, může být slovu přiřazena každá značka, která je danému slovu přiřazena ve slovníku, se stejnou pravděpodobností. Tato hodnota však na celkové vyhledání optimální cesty nemá vliv. V tomto případě je možné pravděpodobnost $p(c_n \mid w_n)$ ignorovat. Pro vyhlazování bigramového jazykového modelu značek $p(c_n \mid c_{n-1})$ je opět použita metoda lineární interpolace. Podmíněná pravděpodobnost $p(c_n \mid w_n)$ je vyhlazena pomocí vyhlazovací techniky Add-One Smoothing.

Graf na obrázku 6.1 porovnává všechny tři způsoby automatického značkování. Z grafu je zřejmé, že TaggerB s nevyhlazeným bigramovým jazykovým modelem značek sestaveným z malého počtu vět vykazuje lepší výsledky než TaggerSB a TaggerSBP s vyhlazenými bigramovými jazykovými modely značek sestavenými ze stejného počtu vět.

Úspěšnost značkovačů TaggerSB a TaggerSBP však po přidání dalších vět rychle stoupá. Z uvedených výsledků lze usuzovat, že všechny tři značkovače vykazují znaky učícího se systému.



Obr. 6.1: Porovnání značkovačů TaggerB, TaggerSB a TaggerSBP

TaggerSBP byl použit pro automatické označkování korpusu velkého 3,1 GB dat (přibližně 558 miliónů slov ve 29 miliónech vět) vytvořeného na Technické univerzitě v Liberci [NOUZA 2004]. Všechna slova korpusu jsou převedena na malá písmena a věty v korpusu jsou uspořádány tak, že na každém řádku je jedna věta. Použité věty nejsou manuálně kontrolovány, takže se v nich objevují nespisovná slova, čísla, různé zkratky apod. V textu je přibližně 4,5 % slov, kterým je přiřazena značka pro neznámé slovo. Z takto označkovaných vět a z 8 300 označkovaných vět byly vytvořeny další dva vyhlazené bigramové jazykové modely značek – z vět s interpunkcí (BigCSLM+) a z vět, kde byla interpunkce vynechána (BigCSLM–). Oba modely byly otestovány na testovacích datech. Model vytvořený z vět s interpunkcí byl testován na testovacích datech s interpunkcí i bez interpunkce.

V tabulce 6.1 je uvedena úspěšnost značkování testovacích dat značkovačem TaggerSBP při použití obou jazykových modelů v porovnání s úspěšností značkování při použití vyhlazeného bigramového jazykového modelu značek z 8 300 vět (Big8300SLM+) vytvořeného z vět s interpunkcí. Z výsledků je zřejmé, že úspěšnost značkování vět bez

interpunkce při použití jazykového modelu z vět s interpunkcí je vyšší než při použití jazykového modelu bez interpunkce.

Tab. 6.1: Úspěšnost značkování pomocí jazykových modelů z velkého množství dat

jazykový model	úspěšnost značkování 500 testovacích vět	
	bez interpunkce	s interpunkcí
BigCSLM-	92,45 %	–
BigCSLM+	93,25 %	94,10 %
Big8300SLM+	93,54 %	94,45 %

7 Využití jazykového modelu založeného na třídách

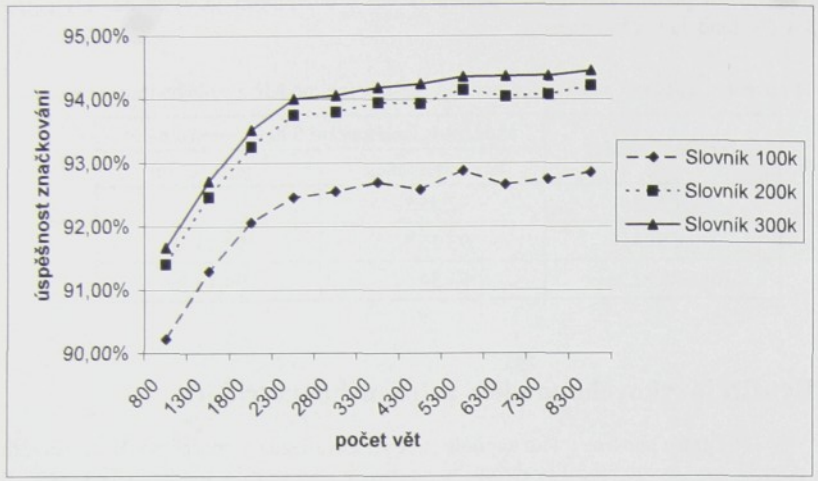
V části práce popsané v této kapitole jsme zjišťovali, zda je možné použít ohodnocení věty pomocí jazykového modelu založeného na třídách jako metriku pro porovnání přesnosti rozpoznávání. Dále jsme zjišťovali, jestli je možné s pomocí jazykového modelu založeného na třídách vyřadit málo pravděpodobné dvojice slov ze seznamu dvojic, které se používají pro tvorbu bigramů. Kromě využití jazykového modelu založeného na třídách je v této kapitole popsán vliv velikosti slovníku a vliv interpunkce na automatické značkování.

7.1 Vliv velikosti slovníku na automatické značkování

Pro automatické značkování byly používány tři různě velké označkované slovníky. Největší obsahuje 313 217 různých slovních tvarů (slovník 300k). Další dva slovníky byly vytvořeny z největšího slovníku tak, že byla vybrána slova s počtem výskytů větším než 10 (slovník s 213 412 různými slovními tvary – slovník 200k) a slova s počtem výskytů větším než 50 (slovník s 111 294 různými slovními tvary – slovník 100k).

Graf na obrázku 7.1 ukazuje porovnání úspěšnosti značkovače TaggerSBP v závislosti na velikosti slovníku. Z grafu je zřejmé, že rozdíl úspěšnosti u slovníků 100k a 300k je přibližně 1,5 %. U slovníků 200k a 300k je rozdíl úspěšnosti podstatně menší – 0,26 %.

Podobné výsledky vykazují také značkovače TaggerB a TaggerSB. Na základě uvedených výsledků lze konstatovat, že úspěšnost značkování s rostoucím počtem slov ve slovníku stoupá. Rozdíl úspěšnosti je však při lineárním přidání počtu slov do slovníku přibližně 5krát menší. Z toho lze usuzovat, že po přidání dalšího množství slov do slovníku bude úspěšnost značkování jen nepatrně vyšší než pro slovník 300k.



Obr. 7.1: Porovnání úspěšnosti značkovače TaggerSBP pro různě velké slovníky

7.2 Vliv interpunkce na automatické značkování

V dalších experimentech byl zjišťován vliv interpunkce na automatické značkování. Všechny použité jazykové modely značek byly vytvářeny jak z vět, které obsahovaly interpunkci, tak z vět, které interpunkci neobsahovaly. Rozdíly úspěšnosti všech značkovačů s použitím jazykového modelu značek vytvořeného z vět s interpunkcí a bez interpunkce se pohybují okolo 1 %. Lze tvrdit, že jednoznačně úspěšnější je automatické značkování, které pro ohodnocení používá jazykový model značek vytvořený z vět s interpunkcí.

7.3 Experimenty s jazykovým modelem založeným na třídách

Ohodnocení vět pomocí jazykového modelu založeného na třídách bylo použito pro porovnání a výběr nejlepšího jazykového modelu, který je součástí rozpoznávače. Pro rozpoznávání parlamentní řeči bylo použito několik jazykových modelů. Cílem bylo vybrat ze všech jazykových modelů ten nejlepší. Pro rozpoznávání řeči byl použit systém, vyvinutý na Technické univerzitě v Liberci v letech 2002 až 2005 [NOUZA 2004].

Pro výpočet ohodnocení vět $P(W)$ byl použit vzorec (2.18), ve kterém byl součin podmíněných pravděpodobností $p(c_n \mid c_{n-1})$ a $p(w_n \mid c_n)$ nahrazen součtem přirozených logaritmů. Podmíněná pravděpodobnost $p(w_1 \mid c_1)$ pro první slovo ve větě je v našem případě rovna jedné, protože každá věta vždy začíná značkou pro začátek věty.

Věty, které byly zpracovány, pocházejí z parlamentních debat a jsou uvedeny bez interpunkce. Parlamentní řeč je specifická tím, že obsahuje hodně dlouhé věty (průměrně 62 slov ve větě), takže často je věta bez interpunkce dost nepřehledná. Tyty věty byly rozpoznány výše uvedeným rozpoznávačem se sedmi různými jazykovými modely (LM): obecným LM (obecny), LM vytvořeným jen z parlamentní řeči (parl only), jejich kombinací (parl x1) a několikerým započítáním LM vytvořeného z parlamentní řeči k obecnému LM (parl x2, parl x4, parl x8, parl x16). Tak vzniklo sedm souborů s různě rozpoznanými 120

větami. Součástí těchto souborů je i přesnost rozpoznávání *Acc*, pro každou větu a pro celý soubor.

Pro výpočet $P(W)$ je třeba nejdříve věty označkovat. Pro označkování jsme použili TaggerSBP s jazykovým modelem značek BigCSLM+ vytvořeným z velkého množství dat (3,1 GB dat). Pro každou větu byl proveden výpočet přirozeného logaritmu pravděpodobnosti $P(W)$ a pro každý soubor byl vypočten součet těchto logaritmů $\Sigma P(W)$. Počet slov, které se nevyskytly ve slovníku 300k z jednotlivých souborů (OOV), je uveden v tabulce 7.1.

Tab. 7.1: Počet OOV v souborech se 120 větami rozpoznanými s použitím různých jazykových modelů

	obecny	parl only	parl x1	parl x2	parl x4	parl x8	parl x16
OOV [%]	1,39	1,54	1,21	1,12	1,05	0,97	0,96

Ze všech sedmi souborů byla vybrána pro každou větu nejlepší úspěšnost Acc_{best} rozpoznávání a největší ohodnocení $P(W)_{max}$. Tyto hodnoty byly použity pro výpočet korelačního faktoru:

$$R = \frac{\sum_{\forall N} i \otimes j}{N} \qquad (7.1)$$

kde i odpovídá rovnosti $P(W) = P(W)_{max}$,
j odpovídá rovnosti $Acc = Acc_{best}$,
$\otimes$ je operace XOR (jestliže současně platí i a j nebo současně neplatí i a j výsledek je 1, jindy je výsledek 0),
N je počet vět,
R je korelační faktor v procentech.

V tabulce 7.2 jsou uvedeny korelační faktory R, přesnosti rozpoznávání Acc a celkové ohodnocení $\Sigma P(W)$ 120 vět pro všech sedm souborů. Tučně jsou uvedeny maximální hodnoty pro R, Acc a $\Sigma P(W)$. Podle přesnosti rozpoznávání je nejlepší použít jazykový model „parl x8" a jako druhý model „parl x16". Podle korelačního faktoru je nejlepší použít jazykový model „parl x16" a jako druhý model vytvořený jen z parlamentní řeči. V případě, že porovnáváme ohodnocení $\Sigma P(W)$ s přesností rozpoznávání Acc, je třeba si uvědomit, že rozpoznaná věta může obsahovat více resp. méně slov než věta vyslovená a potom i $P(W)$ je větší resp. menší než u správně rozpoznané věty. Např. věta „děkuji za pozornost" byla rozpoznána rozpoznávačem s obecným jazykovým modelem nesprávně („jejich pozornost") s přesností rozpoznávání $Acc = 33$ % a s ohodnocením $P(W) = -14{,}2$. S ostatními jazykovými modely byla tato věta rozpoznána správně s přesností $Acc = 100$ % a ohodnocením $P(W) = -18{,}8$. Objektivně proto nelze porovnávat ohodnocení $\Sigma P(W)$ s přesností, protože při výpočtu $\Sigma P(W)$ záleží na počtu slov ve větě a je potřeba toto ohodnocení nějakým způsobem normovat např. hodnotu $\Sigma P(W)$ dělit celkovým počtem slov. V posledním řádku tabulky 7.2 je toto normované ohodnocení uvedeno. Tučně je opět vyznačena maximální hodnota, která říká, že podle normovaného ohodnocení je nejlepší použít jazykový model „parl x16". Z uvedených výsledků vyplývá, že jako metriku pro porovnání rozpoznaných textů z rozpoznávače s různými parametry je možné kromě přesnosti rozpoznávání použít také korelační faktor popřípadě normované ohodnocení.

Tab. 7.2: Korelační faktor, přesnost rozpoznávání a celkové ohodnocení všech vět pro jednotlivé jazykové modely

jazykový model	obecny	parl only	parl x1	parl x2	parl x4	parl x8	parl x16
R [%]	62,50	66,67	65,00	63,33	63,33	65,00	**71,67**
Acc [%]	67,06	67,15	69,67	70,03	70,43	**70,62**	70,45
$\Sigma P(W)$	-61 347	-62 138	**-61 321**	-61 439	-61 450	-61 529	-61 590
$\Sigma P(W)/N$	-8,14	-8,24	-8,12	-8,11	-8,10	-8,09	**-8,07**

Normovaná ohodnocení rozpoznaných textů uvedená v tabulce jsme porovnali s ohodnocením správně přepsaného textu. V případě, že byl správně přepsaný text uveden včetně přeřeknutí, bylo procento OOV 1,9 % a normované ohodnocení –8,06. V případě, že jsme text upravili bez přeřeknutí, procento OOV bylo nižší (1,3 %) a normované ohodnocení bylo –8,07. Z těchto výsledků lze opět konstatovat, že normované ohodnocení lze použít jako metriku pro porovnání úspěšnosti rozpoznávání.

Při výpočtu ohodnocení správně rozpoznaných vět hraje velkou roli i počet OOV, což lze ukázat na následujícím příkladě. Vzhledem k tomu, že parlamentní řeč je specifická, byl výpočet ohodnocení $P(W)$ opakován pro 841 vět z denního tisku. Průměrný počet slov ve větě byl přibližně 20. Celková přesnost rozpoznávání podle vzorce byla rovna 80,79 %. Počet OOV v přepsaném textu byl 2,76 %. U rozpoznaných vět byl počet OOV 1,56 %. Pro rozpoznávání byl použit rozpoznávač s obecným jazykovým modelem. Hodnota normovaného ohodnocení $\Sigma P(W)/N$ pro rozpoznané věty byla rovna –8,17. Byl proveden výpočet normovaného ohodnocení také pro přepsaný text, přičemž hodnota $\Sigma P(W)/N$ byla rovna –8,09. V tomto případě neodpovídá hodnota normovaného ohodnocení výsledkům z předchozího experimentu, kdy při přesnosti rozpoznávání 70,45 % bylo normované ohodnocení rovno –8,07.

Na základě uvedených výsledků nelze nahradit přesnost rozpoznávání normovaným ohodnocením. Normované ohodnocení a korelační faktor však mohou být dobrým měřítkem pro porovnání výsledků z různých rozpoznávačů nebo z rozpoznávače s různými jazykovými modely. Velkou roli při výpočtu normovaného ohodnocení hraje také počet slov, která se nevyskytují ve slovníku (neznámých slov), což jsou většinou ohebná slova. Největší část neznámých slov tvoří podstatná a přídavná jména. V případě, že jsou tato slova nahrazena značkou pro neznámé slovo, dojde ke ztrátě informace o pádu, rodu, čísle apod., takže s tím souvisí i chyba ve značkování okolních slov a ohodnocení je pak zkreslené.

7.4 Odhad četnosti dvojic slov

Jazykový model založený na třídách byl použit také pro odhad četnosti dvojic slov. Soubor s dvojicemi slov včetně počtu výskytů dvojic byl vytvořen z korpusu o velikosti 3,5 GB. Ve dvojicích jsou jen ta slova, která se vyskytla ve slovníku. Tyto dvojice se používají k vytvoření bigramového jazykového modelu, který je součástí rozpoznávače. Kromě dvojic slov jsou v souboru uloženy i takové dvojice, kde místo jednoho nebo obou slov může být sousloví (kolokace) jako např. „a když" „addis abeba" „v letošním". Počet všech dvojic slov resp. sousloví je 64 351 852. Dvojice slov, ve kterých byla obsažena sousloví, byly při dalším zpracování ignorovány.

Jednou z úloh, kde je možné odhad četnosti dvojic slov využít, je nalezení dvojic s malým odhadem četnosti a jejich vyřazení ze seznamu dvojic. Pro odhad četnosti dvojic slov – úpravou z (2.18) – platí:

$$C(w_{n-1}, w_n) = C(w_{n-1}) \cdot p(w_n \mid c_n) \cdot p(c_n \mid c_{n-1}) \qquad (7.2)$$

kde $C(w_{n-1})$ je počet výskytů slova w_{n-1},
$p(w_n \mid c_n)$ je podmíněná pravděpodobnost, s jakou bude k dané značce přiřazeno dané slovo,
$p(c_n \mid c_{n-1})$ je bigram značek.

Při výpočtu odhadu četnosti dvojic slov jsme použili prvky nevyhlazené bigramové matice a podmíněné pravděpodobnosti $p(w_n \mid c_n)$ vyhlazené metodou Add-One Smoothing. Nevyhlazený bigramový jazykový model značek byl vytvořen z automaticky označkovaného korpusu o velikosti 3,1 GB. Interpunkce při tvorbě jazykových modelů byla ignorována, protože i při sestavování dvojic slov z korpusu byla interpunkce ignorována. Ke každému slovu v souboru dvojic byla přidána příslušná značka (značky). V případě, že bylo jednomu nebo oběma slovům ve dvojici přiřazeno více značek, byl vybrán maximální součin ze všech možných součinů dvojic značek pro danou dvojici slov.

V korpusu 3,1 GB je přibližně 4,5 % slov, kterým byla přiřazena značka pro neznámé slovo. Značka pro neznámé slovo se vyskytuje v kombinaci s 94 % značek na levé straně a s 97 % značek na pravé straně dvojice slov. Ve všech dvojicích slov jsou přibližně 4 miliony neznámých slov. Opět jde nejčastěji o podstatná jména. Tato skutečnost samozřejmě napomáhá ke zkreslení odhadů četností přiřazené dvojicím slov. Např. stejný odhad četnosti byl přiřazen jak dvojici slov „dva capart" tak dvojici „dva caparty", protože slovům „capart" a „caparty" je přiřazena značka pro neznámé slovo.

Vzhledem k tomu, že pravděpodobnosti v součinu (7.2) jsou velmi malá čísla, použili jsme váhový koeficient k, kterým jsme celý vztah násobili. Počet výskytů v korpusu je celé číslo, takže vynásobený výsledek byl převeden na celé číslo. Obě uvedené úpravy vzorce (7.2) napomohly k přehlednějším výsledkům. V případě, že byla hodnota váhového koeficientu k nastavena na 100, byla přiřazena nulová hodnota odhadu četnosti přibližně čtvrtině dvojic slov. Pro $k = 10\,000$ byla nulová hodnota odhadu četnosti přiřazena 5 484 448 dvojicím (téměř 10 %). V tabulce 7.3 jsou uvedeny konkrétní příklady dvojic slov s nulovou hodnotou odhadu četnosti pro $k = 10\,000$ včetně počtu výskytů dvojic v ČNK.

Tab. 7.3: Příklady dvojic slov s nulovým odhadem četnosti

slovo 1	skupina 1	slovo 2	skupina 2	četnost v korpusu 3,5 GB	počet výskytů v ČNK
bys	05bys	dva	04dva	1	0
mé	03me	než	08nez	1	0
ho	03ho	ní	03ona0	1	0
ke	07ke3	byl	05byl	1	0
kterého	03tazj2 03tazj4	tisíc	04tisic	1	0

Podle předpokládaných výsledků by měly být ty dvojice slov, u kterých je odhad četnosti velmi malý, vyřazeny. Přestože některé dvojice splňovaly toto kritérium a přestože se tyto dvojice neobjevily v ČNK a v korpusu o velikosti 3,5 GB se objevily velice zřídka,

mohou se tyto dvojice slov v textu vyskytnout. Z dvojic uvedených v tabulce 7.3 se ve větě můžeme setkat například s dvojicemi slov „bys dva“, „mé než“ nebo „kterého tisíc“.

Z uvedených výsledků vyplývá, že námi vytvořený nevyhlazený jazykový model založený na třídách není vhodný pro vyřazení dvojic slov, které slouží k tvorbě bigramového modelu pro rozpoznávač.

Jedním z možných řešení, jak některé málo pravděpodobné dvojice slov vyřadit a jiné přidat, by mohlo být vytvoření souboru všech pravděpodobných dvojic ze slovníku 300k s odhadem četností podle (7.2), což je předmětem dalšího výzkumu.

8 Závěr

Dizertační práce předkládá návrh jazykového modelu založeného na třídách včetně jeho praktického použití.

V první řadě bylo třeba stanovit gramatické značky. K stanovení značek byly využity tři přístupy: statistický, gramatický a pravděpodobnostní. S použitím statistického přístupu byla vybrána nejfrekventovanější slova včetně interpunkce, která tvoří přibližně 30 % textu. Těm byly přiřazeny značky tak, že každé značce bylo přiřazeno právě jedno slovo. Pro stanovení dalších značek byl využit hladový algoritmus, gramatický a syntaktický přístup. Seznam všech značek je uveden v příloze dizertační práce.

K tvorbě jazykového modelu založeného na třídách byl vytvořen označkovaný korpus, jehož část byla označkována ručně a část automaticky. Věty, které jsou v korpusu použity, byly získány především z internetu a jde o texty ze zpravodajství, různých novinových článků, knížek apod. Do korpusu byly přidány také věty „uměle“ vytvořené, aby byly v korpusu obsaženy i méně frekventované značky.

Pro tvorbu označkovaného slovníku byl využit slovník vytvořený na Technické univerzitě v Liberci obsahující přibližně 300 000 slov, do kterého byla ručně a na základě syntaktické metody přidána informace o příslušných gramatických značkách. Slovník byl použit pro automatické značkování vět. Na základě 3 300 ručně označkovaných vět byly vytvořeny dva bigramové modely značek – nevyhlazený a vyhlazený. Pro vyhlazení jazykového modelu značek byla použita metoda lineární interpolace. Na základě Bellmanova principu optimality bylo provedeno automatické značkování dalších 5 000 vět. Pro automatické značkování byly vytvořeny tři značkovače (taggery). TaggerB využívá pro automatické značkování nevyhlazenou bigramovou matici značek, TaggerSB bigramovou matici značek vyhlazenou metodou lineární interpolace a TaggerSBP bigramovou matici značek vyhlazenou metodou lineární interpolace a pravděpodobnost výskytu daného slova v dané skupině vyhlazenou metodou Add-One Smoothing. Nejlepší výsledky automatického značkování vykázal značkovač TaggerSBP.

V dizertační práci jsou také shrnuty výsledky automatického značkování s různě velkým slovníkem. Z výsledků vyplývá, že zlepšení úspěšnosti při lineárním zvětšování slovníku stoupá logaritmicky. Při automatickém značkování byly použity jazykové modely značek, které byly vytvořeny z vět s interpunkcí a bez interpunkce. Výsledky jednoznačně prokázaly, že automatické značkování je úspěšnější (cca o 1 %) v případě, že bereme v úvahu interpunkci. Nejlepší výsledky automatického značkování (94,5 %) byly dosaženy při použití značkovače TaggerSBP s jazykovým modelem značek vytvořeným z vět s interpunkcí a s využitím slovníku o velikosti 300 tisíc slov.

Na základě experimentů s větami rozpoznanými systémem pro rozpoznávání spojité řeči jsme došli k závěru, že pomocí jazykového modelu lze rozpoznané věty ohodnotit a normované ohodnocení použít pro porovnání výsledků z různých rozpoznávačů nebo z rozpoznávače s různými jazykovými modely.

Přestože jsme předpokládali, že s pomocí nevyhlazeného jazykového modelu založeného na třídách bude možné vyřadit málo pravděpodobné dvojice slov ze seznamu dvojic, které se používají pro tvorbu bigramů, výsledky experimentů tuto hypotézu vyvrátily.

Na základě výsledků dizertační práce hodláme vytvořit soubor všech pravděpodobných dvojic slov ze slovníku o velikosti 300 tisíc slov s odhadem četností pomocí bigramového jazykového modelu založeného na třídách. Soubor dvojic se využívá pro tvorbu jazykového modelu, který je součástí rozpoznávače. Předpokládáme, že s takto vytvořeným jazykovým modelem, by mohlo dojít ke zlepšení přesnosti rozpoznávání. Předmětem dalšího výzkumu je vytvoření trigramového jazykového modelu založeného na třídách a jeho využití nejen ve finální fázi rozpoznávání pro korekturu již rozpoznaných vět, ale i v předzpracování textů sloužících pro tvorbu bigramového jazykového modelu, který je součástí rozpoznávače.

Literatura

[CHURCH 1988] CHURCH, Keneth Ward. *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text.* In Proceedings of the Second Conference on Applied Natural Language Processing (ANLP), Austin, Texas, s. 136–143, 1988.

[ČERMÁK 2004] ČERMÁK, František – SCHMIEDTOVÁ, Věra. *Český národní korpus – základní charakteristika a širší souvislosti.* Národní knihovna, 15, 2004, č. 3, s. 152–168, ISSN 1214-0678. URL: <http://full.nkp.cz/nkkr/nkkr0403/0403152.html>.

[JURAFSKY 2000] JURAFSKY, Daniel – MARTIN, James H. *Speech and Language Processing.* Prentice-Hall, Inc., New Jersey, 2000, ISBN 0-13-095069-6.

[MRVA 2000] MRVA, David. *Jazykové modelování přirozeného jazyka založené na kořenech a koncovkách.* Praha, 2000. Diplomová práce na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze.

[NOUZA 2004] NOUZA, Jan, et al. *Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs.* In: Proc. of ICSLP 2004, October 2004, Jeju Island, Korea, s. 409–412, ISSN 1225-441x.

[PSUTKA 1995] PSUTKA, J. *Komunikace s počítačem mluvenou řečí.* Academia, Praha, 1995, ISBN 80-200-0203-0.

Vlastní publikované práce

NEJEDLOVÁ, D. – DRÁBKOVÁ, J. – KOLORENČ, J. – NOUZA, J. *Lexical, Phonetic, and Grammatical Aspects of Very-Large-Vocabulary Continuous Speech Recognition of Czech Language.* In: Electronic Speech Signal Processing, September 2005, Prague, Czech Republic, s. 224–231. ISBN 80-86269-10-8

DRÁBKOVÁ, J. *Punctuation Effect on Class-Based Language for Czech Language*. In: Electronic Speech Signal Processing, September 2005, Prague, Czech Republic, s. 267–272. ISBN 80-86269-10-8

DRÁBKOVÁ, J. – HOLADA, M. – NOUZA, J. – HORÁK, P. – NOUZA, T. *New Version of Phone Dialogue Information System InfoCity*. In: Proc. of 14th Czech-German Workshop „Speech Processing", September 2004, Prague, Czech Republic, s. 66–71, ISBN 80-86269-11-6

DRÁBKOVÁ, J. *Formation of Classes for Continuous Speech Language Model and Building the Large Tagging Vocabulary for Czech Language*. In: Proc. of 13th Czech-German Workshop „Speech Processing", September 2003, Prague, Czech Republic, s. 121–125. ISBN 80-86269-10-8

DRÁBKOVÁ, J. *How Good is Speech Recognition Performed by Human and by Machine?* In Proc. of 6th International Workshop on Electronics, Control, Measurment and Signals-ECMS 2003. Liberec, June 2003. s. 79–83. ISBN 80-7083-708-X

DRÁBKOVÁ, J. *Language Model Based on the Czech Morphology*. In Proc. of 12th Czech-German Workshop „Speech Processing". Prague, September 2002, s. 70–73. ISBN 80-86269-09-4

NOUZA, J. – DRÁBKOVÁ, J. *Combining Lexical and Morphological Knowledge in Language Model for Inflectional (Czech) Language*. In Proc. of 6th Int. Conference on Spoken Language Processing. Denver USA, September 2002, s. 705–708. ISBN 1876346418

NOUZA, T. – NOUZA, J. – DRÁBKOVÁ, J. *An Efficient Graphic System for Developing Voice Operated Applications*. In Proc. of SCI 2002. Orlando USA, July 2002, Volume I, s. 239–244. ISBN 980-07-8150-1

CHALOUPKA, J. – NOUZA, J. – DRÁBKOVÁ, J. *Developing an Artificial Talking Head for Czech Language*. In Proc. of SCI 2002. Orlando USA, July 2002, Volume III, s. 232–236. ISBN 980-07-8150-1