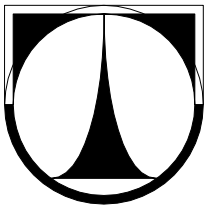
TECHNICKÁ UNIVERZITA V LIBERCI

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:

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

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

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

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:

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

Pravděpodobnost pro vyhlazený unigram se vypočítá podle:

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

Obdobně platí pro nevyhlazený bigram:

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

Pravděpodobnost pro vyhlazený bigram se vypočítá podle:

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

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:

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

Analogický vztah platí pro bigramy:

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

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 $p(W_H)$:

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

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.

$$G : c_n = G(w_n) \tag{2.15}$$

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]:

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

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:

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

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

Pro bigramový jazykový model založený na třídách potom platí:

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

Tento jazykový model sestává ze dvou složek:

- bigram značek

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

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

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

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).

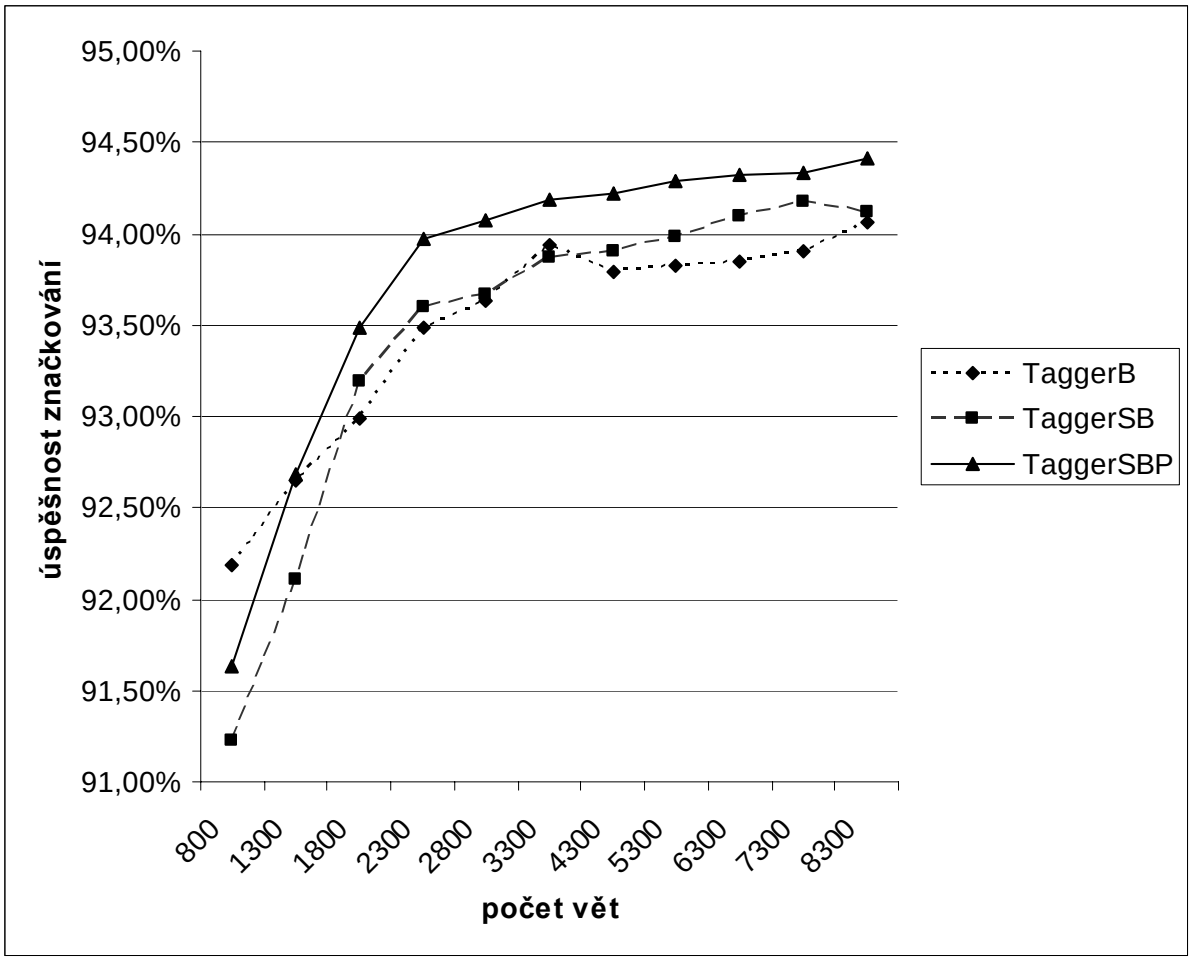
Stochastický značkovač s nevyhlazeným jazykovým modelem značek (TaggerB) používá k ohodnocení hran grafu prvky nevyhlazené bigramové matice.

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

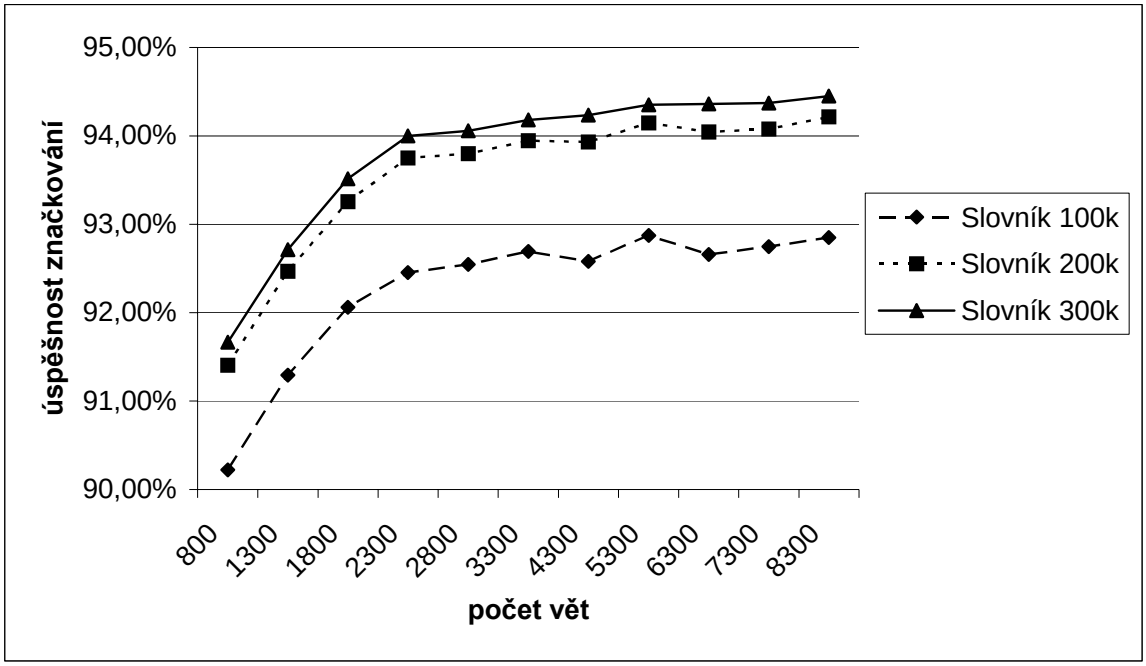
V části práce popsané v této kapitole jsme zjišťovali, zda je možné použít ohodnocení věty pomocí jazykového modelu založeného na třídách jako metriku pro porovnání přesnosti rozpoznávání. Dále jsme zjišťovali, jestli je možné s pomocí jazykového modelu založeného na třídách vyřadit málo pravděpodobné 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 ΣP(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:

$$R = \frac{\sum_{\forall N} i \otimes j}{N} \tag{7.1}$$

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 ΣP(W) 120 vět pro všech sedm souborů. Tučně jsou uvedeny maximální hodnoty pro R, *Acc* a ΣP(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 ΣP(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 ΣP(W) s přesností, protože při výpočtu ΣP(W) záleží na počtu slov ve větě a je potřeba toto ohodnocení nějakým způsobem normovat např. hodnotu ΣP(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
Σ*P*(W)	-61 347	-62 138	**-61 321**	-61 439	-61 450	-61 529	-61 590
Σ*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 Σ*P*(W)/*N* pro rozpoznané věty byla rovna –8,17. Byl proveden výpočet normovaného ohodnocení také pro přepsaný text, přičemž hodnota Σ*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:

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

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 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 (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

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