AUTOMATICKÁ TRANSKRIPCIA SLOVENČINY V POČÍTAČOVOM ROZPOZNÁVANÍ REČI

Jozef Ivanecký

IBM Voice Systems, European Speech Research,
Vangerowstr. 18, D-69115 Heidelberg
ivanecky@de.ibm.com

Technická Univerzita Košice
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie

ABSTRAKT

The paper reflects on one of the important tasks in the computer speech recognition - phonetic transcription - whose successful management has a significant impact on the process of both training and the recognition itself. We present the problems of the automatic transcription, procedures used in the formation of the system capable to transcribe a common text from an orthographic to orthoepical form. Furthermore we deal with the need to cope with a various (not only standard) pronunciation in the automatic transcription.

1. ÚVOD

Medzi najrozšírenejšie techniky pre počítačové rozpoznávanie reči patria v dnešnej dobe skryté markovovské modely (HMM) a neurónové siete (NN). V oboch prípadoch je na vytvorenie systému potrebné väčšie množstvo dát (niekoľko tisíc viet). V prípade kvalitnejších systémov ide o niekoľko desiatok až stoviek tisíc viet. Vstupom na trénovanie takýchto systémov je okrem nahratých viet aj ich fonetická transkripcia. Od jej kvality značne závisí kvalita výsledného systému.

Transkripciu môžeme vytvoriť manuálne, čo je v prípade väčšieho množstva viet značne pracné a môže viesť k väčšiemu množstvu chýb. V prípade použitia takýchto údajov pri trénovaní sa znižuje úspešnosť rozpoznávania. Ešte väčší vplyv má na druhej strane nepoužitie takýchto údajov v procese samotného rozpoznávania.

Druhý prístup je automatická transkripcia. Na rozdiel od prvej je pri nej možné vylúčiť mnohé chyby. Na druhej strane je vytvorenie systému na automatickú transkripciu značne zložitá záležitosť. V niektorých jazykoch natoľko zložitá, že sa od nej celkom upúšťa. O automatickej transkripcii slovenčiny, avšak za účelom syntézy, pojednáva [2]. Pre automatickú transkripciu je potrebné nájsť všeobecné pravidlá, pomocou ktorých by bolo možné fonetický, resp. fonematický prepis automaticky vytvárať. Tieto pravidlá môžu byť formulované ako produkčné pravidlá a nazývajú sa transkripčné pravidlá.

Ďalším problémom, ktorým sa zaoberáme v tomto príspevku, je problematika spisovnej výslovnosti v rozpoznávaní reči. Použitie fonologických pravidiel zadefinovaných na základe [3] síce bude vytvárať spisovný prepis, avšak ten nemusí a pravdepodobne ani nebude presne zodpovedať skutočnej výslovnosti. O to výraznejšie to bude v prípade, keď nahrávky pokrývajú všetky nárečové skupiny.

2. TRANSKRIPCIA VŠEOBECNE

Úlohy obsiahnuté v systéme pre automatickú fonetickú transkripciu môžeme rozdeliť na dve základné podskupiny:

- jazykovo nezávislé úlohy
- jazykovo závislé úlohy

V tejto časti sa budeme venovať prvej skupine. Zjednodušene môžeme povedať, že sem patria všetky úlohy, ktoré sú nad vstupným textom vykonané predtým, než sú aplikované vlastné jazykovo závislé transkripčné pravidlá.

Segmentácia: Prvým krokom je konverzia všetkých veľkých písmen na malé a rozdelenie vstupného textu na segmenty vhodné pre transkripciu. Na tomto mieste si musíme uvedomiť, čo je minimálny segment. V prípade, že počas transkripcie nechceme brať do úvahy ortoepické javy vznikajúce na hranici slov, realizujeme segmentáciu na jednotlivé slová. V opačnom prípade musíme vstupný text rozdeliť na jednoduché vety s vyznačením hraníc slov.

Filtrácia: V tomto kroku odstraňujeme všetky nepotrebné znaky z textu.

Kódovanie: Vstupný text môže byť v jednom z N rôznych kódovaní. V prípade slovenčiny je $N=4$. V našom prípade je kódovanie vstupného textu zmenené do interného kódovania, ktoré obsahuje len 7 bitové ASCII znaky. Dôvodom pre toto riešenie bola požiadavka na ľahkú prenositeľnosť medzi jednotlivými operačnými systémami pri zaručenej čitateľnosti kódu. Celý systém je napísaný v jazyku Perl a vytváraný pod operačným systémom Linux.
Z hľadiska internej štruktúry údajov bolo potrebné zaručiť jednoznačnú identifikáciu už prepísaných a ešte neprepísaných častí, keďže výstupná abeceda pozostáva taktiež zo 7 bitové ASCII znakov, vzhľadom na to, že pre výstup zo systému bola zvolená SAMPA abeceda. Na tomto mieste je dôležité poznamenať, že SAMPA abeceda použitá ako výstupná abeceda v našom systéme zatiaľ nie je oficiálnou SAMPA abecedou pre slovenčinu.

Použitie všetkých vyššie opísaných krokov demonštrujeme na nasledujúcej vete:

Keby som bol vedel, ako je tam zle, nikam by som nešiel ...

Po aplikovaní prvého kroku sa dopracujeme k nasledujúcemu tvaru vstupnej vety:

keby som bol vedel; ako je tam zle; nikam by som nešiel

Bodkočiarka v tomto prípade oddeľuje jednotlivé segmenty. Ak by najmenším segmentom bolo slovo, nachádzala by sa za každým slovom. Ďalšími krokmi prechádza každý segment samostatne. Po aplikovaní filtrácie na tretí segment dostaneme:

nikam-by-som-nešiel

Výsledkom záverečného prekódovania je:

!n! !i! !k! !a! !m! - !b! !y! - !s! !o! !m! - !n! !e! !S! !i! !e! !l!

Znak "*!*" slúži na jasnú identifikáciu znakov patriacich k vstupnej abecede. Znak – označuje hranicu slov. Všetky ďalšie kroky spadajú do kategórie jazykovo závislých.

3. TRANSKRIPCIA SLOVENČINY

Pri vytváraní transkripčných pravidiel pre slovenčinu sme čerpali hlavne z [3]. Na každý segment sú postupne aplikované všetky pravidlá. Sadu pravidiel môžeme rozdeliť do skupín podľa rôznych kritérií, napríklad fonetických, avšak z nášho pohľadu sa pravidlá delia na dva základné skupiny:

- jednoznačné transkripčné pravidlá
- nejednoznačné transkripčné pravidlá

Medzi jednoznačné transkripčné pravidla zaraďujeme také pravidlá, pri ktorých je jednoznačný vzťah medzi ortografickou a ortoepickou reprezentáciou, a teda transkripcia je možná bez potreby akýchkoľvek ďalších znalostí. Typickým príkladom pre túto skupinu pravidiel sú pravidlá pre samohlásky. Pre samohlásky sme preto zadefinovali nasledujúce transkripčné pravidlá:

Krátke samohlásky:			Dlhé samohlásky:		
a	→	a	á	→	a:
e	→	E	é	→	e:
i	→	I	í	→	i:
y	→	i	ý	→	i:
o	→	o	ó	→	o:
u	→	u	ú	→	u:
ä	→	E			

Vyššie uvedené pravidlá sú na samohlásku aplikované vtedy, ak sa samohláska nachádza medzi dvoma spoluhláskami alebo ak sa nachádza na hranici slova. Znak *ä* prepisujeme v našom systéme ako *E* z nasledujúcich dôvodov:

- Výskyt spisovnej výslovnosti znaku *ä* v hovorovej výslovnosti je v súčasnosti minimálny. Len tento dôvod by však bol v rozpore s našimi ďalšími tvrdeniami uvedenými v nasledujúcom texte.
- V slovenčine vieme len o jednom prípade, kde znak *ä* má význam pri rozlíšeni významu dvoch slov. Ak však aj v tomto jedinom prípade berieme do úvahy predchádzajúce tvrdenie, domnievame sa, že prepis znaku *ä* ako *E* je viac než opodstatnený.

Pre tretí segment z vyššie uvedeného príkladu po aplikácii pravidiel pre samohlásky dostaneme výsledok:

!n! I !k! a !m! - !b! I - !s! o !m! - !n! E !S! !i! !e! !l!

Na poslednom slove si môžeme všimnúť, že dvojhlásky ostali nezmenené. Podobné jednoznačné pravidlá, ako sme zadefinovali pre samohlásky, môžeme zadefinovať aj pre niektoré spoluhlásky. Napríklad: "*p, r, l, m* " a iné. Na tomto mieste je potrebné zdôrazniť, že jednoznačnosť pravidiel neznamená aj ich jednoduchosť. Pre mnohé spoluhlásky existuje viacero transkripčných pravidiel. Výber pravidla závisí na okolí znaku, na ktorý aplikujeme dané transkripčné pravidlo. Ako jednoduchý príklad nám môže poslúžiť spoluhláska "*m*". V [3] sa o výslovnosti znaku "*m*" môžeme dočítať nasledujúce: "*Pred spoluhláskami 'v, f' sa vyslovuje namiesto pernoperného 'm' pernozubné 'm–'. ... Spoluhláska 'm–' sa v spisovnej slovenčine vyslovuje vtedy, keď po 'm' vnútri slova bezprostredne nasleduje v, f ako v slovách 'domvedúca, komfort' ...*". Pre znak "*m*" sme preto zadefinovali nasledujúce pravidlá:

!m! !v!	$\rightarrow$	F v
!m! !f!	$\rightarrow$	F f
!m!	$\rightarrow$	m

Znak "*F* " je SAMPA znakom pre nosové "*m* ". Ak dodržíme pri aplikovaní transkripčných pravidiel pre "*m*" vyššie uvedené poradie, dosiahneme presné aplikovanie pravidiel citovaných z [3]. V tomto prípade sme pravidlá uviedli v internom kódovaní, aby sme ukázali princíp činnosti transkripčného algoritmu. Z uvedeného príkladu vyplýva aj to, že pravidlá pre "*v*" a "*f*" musia nasledovať za pravidlom pre "*m*".

V prípade nejednoznačných pravidiel nie je možné zadefinovať jednoduché transkripčné pravidlá, pomocou ktorých by bolo možné len na základe okolia jednoduchým algoritmom zaistiť správny prepis. V takomto prípade sú na správnu transkripciu potrebné ďalšie informácie. Ako jednoduchý príklad nám tentokrát poslúžia dvojhlásky.

V spisovnej slovenčine sú štyri dvojhlásky: "*ia, ie, iu, ô*". V prípade, že po samohláske "*i*" bezprostredne nasleduje ďalšia samohláska, nemusí ísť o dvojhlásku. Príkladom sú predponové slová (priučiť, vyučiť) a prevzaté slová (Ázia). Prepis dvojhlásky "*ô*" a slov tvorených predponou "*vy*" a základom začínajúcim samohláskou je jednoznačný, preto sa ním nebudeme zaoberať. Problémy vznikajú až pri slovách vytvorených predponou "*pri*" a slovným základom začínajúcim samohláskou. V tomto prípade už nie je možné len na základe jednoduchých pravidiel určiť, či dvojice samohlások "*ia, ie, iu*" tvoria samohláskovú postupnosť, alebo dvojhlásku. Prístup k riešeniu tohto problému rozoberieme pre každú dvojicu samostatne:

ia: Ak sa na začiatku slova nachádza postupnosť znakov "*pria*", na určenie toho, či ide o dvojhlásku alebo samohláskovú skupinu, sme využili skutočnosť, že slovenčina má len 23 slov [3], v ktorých základnom tvare sa na začiatku vyskytuje postupnosť znakov "*pria*" a znaky "*ia*" tvoria dvojhlásku. V prípade, že slovo patrí do tejto skupiny slov, je použité pravidlo

$$\hat{}\,!p!\ !r!\ !i!\ !a! \quad \rightarrow \quad \text{p r i\_\^{}a}$$

v opačnom prípade

$$\hat{}\,!p!\ !r!\ !i!\ !a! \quad \rightarrow \quad \text{p r I a}$$

Zoznam slov bez ohýbacej prípony bol vytvorený podľa počítačového korpusu slovenčiny a literatúry [3] a začlenený do pomocnej databázy. Databáza obsahuje výnimky vzťahujúce sa na konkrétne pravidlá. V tomto prípade všetky slová začínajúce na "*pria*" kde "*ia*" tvorí dvojhlásku a nie samohláskovú skupinu.

ie: V prípade postupnosti znakov "*prie*", sme zvolili opačný prístup, keďže slov, v ktorých postupnosť "*pri*" netvorí predponu ako v predchádzajúcom prípade, je podstatne viac. V sledovanom korpuse sme našli 593 rôznych slov začínajúcich na "*prie*", pričom vo všetkých prípadoch išlo o dvojhlásku a nie o spoluhláskovú postupnosť. Ani literatúra [3] neuvádza slová, v ktorých by "*prie*" obsahovalo spoluhláskovú postupnosť, teoreticky však môžu takéto slová vzniknúť spojením prevzatého, resp. cudzieho slova začínajúceho na "*e*" a predpony "*pri–*" ako napríklad "*prielegantný*". Pre tento prípad je za predpokladu, že dané slovo sa nachádza v pomocnom slovníku, použité nasledujúce pravidlo: V prípade, že v slovníku existuje výnimka pre predponu "*prie*" ktorá je aplikovateľná na dané slovo, použije sa pravidlo

ˆ!p! !r! !i! !e! → p r I E

v opačnom prípade

ˆ!p! !r! !i! !e! → p r i_ˆE

iu: Posledná z trojice je postupnosť "*iu*". Keďže slovenčina nemá bezpredponové slovo začínajúce s "*priu*", kde "*iu*" predstavuje dvojhlásku, je možné na každé slovo aplikovať nasledujúce pravidlo

ˆ!p! !r! !i! !u! → p r I u

V sledovanom korpuse boli jedinými slovami s postupnosťou "*priu–*" slová odvodené od slova "*priučiť*". Celkový výskyt dvojhlásky "*iu*" je v slovenčine veľmi nízky, podľa [5] je to len okolo 0.016%.

Postupy uvedené pre prípad nejednoznačnosti pravidiel pri dvojhláskach používame aj pre spoluhlásky, rovnako je to pri jednoznačných pravidlách. Po aplikovaní všetkých pravidiel v systéme na náš tretí segment uvedený na začiatku, získame nasledujúci prepis:

J I k a m - b I - s o m - J E S i_ˆE l

4. SPRÁVNE VERZUS SPISOVNE

V predchádzajúcej časti sme predpokladali, že text, ku ktorému potrebujeme vygenerovať fonetickú transkripciu, je vyslovovaný spisovne a zodpovedá pravidlám v [3], a teda rovnako aj nášmu výstupu, ktorý bol podľa týchto pravidiel vytvorený. Hovorená forma jazyka sa v rôznych oblastiach viac či menej líši od formy spisovnej a zanedbanie tejto skutočnosti pre účely počítačového rozpoznávania reči môže viesť k chybnému označeniu niektorých segmentov reči v prípade, že tento proces je vykonávaný automaticky a v zozname možných výslovností sa výslovnosť, ktorá presne zodpovedá nahovorenému textu, nevyskytuje.

Z tohto dôvodu, aby sme získali správny prepis výslovnosti, museli sme do nášho systému pridať možnosť generovať aj takzvanú "nespisovnú" výslovnosť. Pre tento druh pravidiel sme čerpali z vlastných skúseností v oblasti rozpoznávania reči rovnako ako aj z [3]. Na základe vlastných skúseností sme vyrobili pravidlá na prepis dlhých samohlások na krátke, čo sa nám osvedčilo v prípade hovoriacich z východných oblastí Slovenska. Ako príklad na druhý zdroj nám môže opäť poslúžiť hláska "*m*". "*V neutrálnom a vyššom štýle výslovnosti sa v slovenčine nepripúšťa výslovnosť pernozubnej spoluhlásky 'm–' namiesto 'n, J'. V slovách ako 'dezinfekcia, fanfáry, infekcia, informácia, kanva, konfekcia, konferencia, konflikt, konvergencia, konvoj, panvica, panvička, skonfiškovať' a v slovných spojeniach ako 'ten váš, len volaj, jeleň vyskočil*". Pre nás bolo toto pravidlo indikáciou toho, že takáto výslovnosť sa môže vyskytnúť, a pridali sme ho medzi pravidlá pre generovanie nespisovnej výslovnosti.

5. ZÁVER

Cieľom tohto príspevku bolo poukázať na problematiku automatizácie fonetickej transkripcie a prezentovať nami zvolený spôsob riešenia tohto problému. Doterajšie výsledky dosiahnuté v budovaní systému sú porovnateľné s výsledkami podobných systémov v iných jazykoch.

Naším ďalším cieľom je zvýšenie robustnosti systému zameraním sa na frekventované prevzaté slová v slovenčine, a tým aj rozšírenie databázy výnimiek. Z dlhodobejšieho hľadiska sa chceme zamerať na kvalitu generovania nespisovnej výslovnosti, čo je však už podmienené analýzou výsledkov reálnych systémov.

Poďakovanie Chceli by sme na tomto mieste poďakovať pracovníkom Jazykovedného ústavu Ľudovíta Štúra v Bratislave za ústretový prístup a možnosť použitia korpusu slovenčiny, bez ktorého by sme neboli schopní dopracovať sa k mnohým z našich záverov.

LITERATÚRA

[1] Černý M., Matoušek V., Mautner P.: An automatic creation of the pronunciation dictionary. *Proceedings of 3rd Slovenian – German and 2nd SDRV Workshop*, April 24–26, 1996, Ljubljana, Slovenia

[2] Daržágín S., Franeková Ľ., Rusko M.: *Konverzia a rečová syntéza slovenčiny*, Jazykovedný časopis, 45, 1994, 1, s. 31–43

[3] Kráľ A.: *Pravidlá slovenskej výslovnosti*, Slovenské pedagogické nakladateľstvo, Bratislava 1983, 632 s.

[4] Kráľ A., Sabol J.: *Fonetika a fonológia*, Slovenské pedagogické nakladateľstvo, Bratislava 1987, 392 s.

[5] Mistrík J.: *Frekvencia tvarov a konštrukcií v slovenčine*, Vydavateľstvo VEDA, Bratislava 1985, 320 s.

[6] Ondruš Š., Sabol J.: *Úvod do štúdia jazykov*, Slovenské pedagogické nakladateľstvo, Bratislava 1987, 344 s.

[7] Pauliny E.: *Slovenská fonológia*, Slovenské pedagogické nakladateľstvo, Bratislava 1979, 213 s.

[8] Sabol J.: *Syntetická fonologická teória*, Jazykovedný Ústav Ľ. Štúra SAV, Bratislava 1988, 255 s.

[9] Wells J. C.: Computer–coded phonetic transcription. *Journal of the International Phonetic Association* 17, No. 2, s. 94–114, 1987