Použití mluvených korpusů ve vývoji systému pro rozpoznávání českých přednášek*

Tomáš Mikolov, Ilya Oparin, Ondřej Glembek, Lukáš Burget, Martin Karafiát, Jan Černocký

Speech@FIT, Ústav počítačové grafiky a multimédií FIT VUT v Brně,
Božetěchova 2, 61266 Brno
speech@fit.vutbr.cz

Abstrakt

Skupina automatického zpracování mluvené řeči na Fakultě informačních technologií VUT v Brně – Speech@FIT – je aktivní v mnoha oblastech automatického zpracování mluvené řeči jako je přepis na text, detekce klíčových slov, ověřování mluvčího a identifikace jazyka. V poslední době se zabývá rozpoznáváním spontánní mluvené řeči v přednáškách. Tento příspěvek se zabývá použitím českých mluvených korpusů pro trénování jazykového modelu pro přednášky. Ukázali jsme, že jazykové modely trénované na mluvených korpusech předčí modely trénované čistě na textových datech. Ještě lepších výsledků bylo ovšem dosaženo s daty přímo z cílové domény rozpoznávání. Experimentální výsledky jsou ukázány na perplexitě jazykového modelu na cílových datech a na konečné úspěšnosti rozpoznávání.

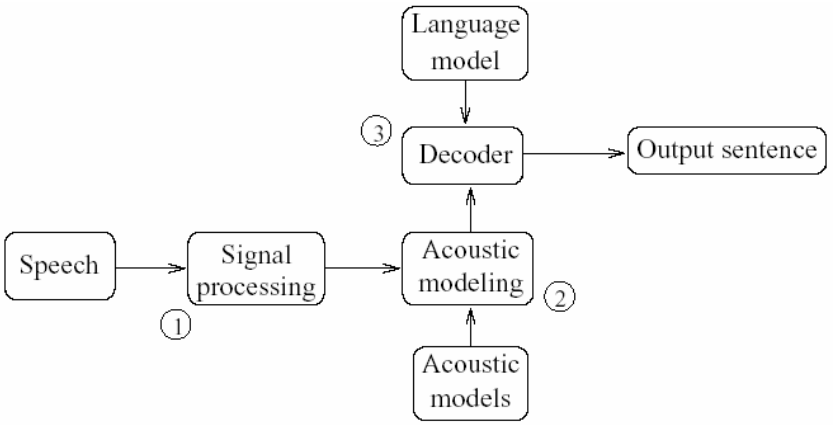
1. Úvod

E-learning se šíří napříč vzdělávacími a odbornými institucemi všech druhů [IEEE] a podle odhadů [Bates2004] se bude v několika následujících letech dále rozšiřovat. E-learning je však často mylně vykládán jako pouhé nahrávání přednášek, seminářů a jejich ukládání případně streamování společně s výukovými materiály. Jiný, neméně důležitý aspekt e-learningu, kterým je rychlá orientace studenta v dostupných audiovizuálních (AV) studijních materiálech. Pro tuto rychlou orientaci je nezbytné, aby byly k AV materiálům k disposici metainformace. V případech "osvíceného" pořizování takových AV materiálu, kdy jsou k disposici časově označkované slajdy, učitelovy poznámky k nim, studentské blogy, či jiné doprovodné texty, disponujeme již množstvím meta-informace. To nejdůležitější – co vyučující ve skutečnosti říká a na co ukazuje – však stále chybí. V mnoha případech jsou přitom učitelův výklad a jeho chování při výuce zcela nezbytné pro pochopení látky a jejích širších souvislostí. Při výkladu jednoho slajdu může učitel například poukázat slovně na zajímavou aplikaci, spojitost s jinou přednáškou, slovně popsat numerický příklad, či upozornit na vztah s jinou disciplínou. Proto pokládáme vyhledávání v audiu a videu za velmi důležité a zabýváme se rozpoznáváním přednášek, a jejich indexací. Tato práce spadá pod rozpoznávání spojité řeči s velkým slovníkem (LVCSR – large vocabulary continuous speech recognition) [Woodland2002].

2. Rozpoznávání a jazykové modelování a na čem je trénovat

* Tato práce byla částečně podporována Grantovou agenturou České republiky, projekt č. 102/05/0278 a Výzkumným záměrem Ministerstva školství, mládeže a tělovýchovy ČR, č. MSM0021630528. Hardware použité v této práci bylo částečně poskytnuto CESNET projekty č. 119/2004, 162/2005 a 201/2006. Lukáš Burget je podporován Grantovou agenturou České republiky, post-doktorský projekt č. GP102/06/383.

LVCSR přepisuje audio na řetězec nebo orientovaný graf se slovy. Základní schéma rozpoznávacího systému je na Obr. 1. Vstupní řeč je v bloku zpracování signálu parametrizovaná vektory tzv. akustických parametrů (features). Akustický model kvantifikuje shodu vstupní řeči se základními řečovými jednotkami používanými pro modelování akustické formy signálu – to mohou být fonémy nebo kontextově závislé fonémy. Dekodér pak za pomocí jazykového modelu (který kvantifikuje a-priorní pravděpodobnosti sekvencí slov) prohledává prostor všech možných vět a generuje nejlepší řetězec slov nebo nejlepší orientovaný graf s různými variantami slov.
Akustický model je trénován na řečových datech s odpovídajícími slovními přepisy. Ty jsou pomocí výslovnostního slovníku převedeny na sekvence fonémů a na jich je pak natrénována sada akustických modelů pro jednotlivé základní zvuky.



Obr. 1: Rozpoznávání řeči LVCSR

Jazykový model si klade za cíl určit pravděpodobnost určité promluvy v daném jazyce. Statistické jazykové modely definují pravděpodobnost promluvy *w* jako

$$P(w) = \prod_{i=1}^{K} P(w_i \mid w_1, w_2, ..., w_{i-1})$$

kde *K* je počet slov promluvy a $P(w_i \mid w_1, w_2, ..., w_{i-1})$ je pravděpodobnost *i*-tého slova promluvy *w*. Nejpoužívanějšími technikami pro odhad této pravděpodobnosti jsou N-gramové modely, kde slovo w_i závisí pouze na N-1 předchozích slovech:

$$P(w_i \mid w_1, w_2, \ldots, w_{i-1}) = \frac{C(w_{i-N+1}, \ldots, w_i)}{C(w_{i-N+1}, \ldots, w_{i-1})}$$

kde *C(*)* označuje počet výskytů daných *N*-gramů v trénovacím korpusu. Pro krátké kontexty (např. 0 slov - unigramový model, 1 slovo - bigramový model) jsou tyto odhady příliš zjednodušené, a tím pádem nepřesné. Naopak, pro dlouhé kontexty (více než 3 slova) má většina *N*-gramů pouze jeden příklad, a jsou tedy nevhodné pro statistické modely. Běžně používané modely pracují s historií délky 2 - jedná se o trigramové modely.

Pro porovnání různých jazykových modelů je nejčastěji využívaným měřítkem *perplexita*, která určuje prediktivní schopnost daného modelu na daných testovacích datech:

$$PPL = \sqrt[K]{\prod_{i=1}^{K} \frac{1}{P(w_i \mid w_{1\ldots i-1})}}$$

Její hodnota určuje průměrnou míru větvení jazykového modelu - vyšší hodnota znamená, že jazykový model si je více nejistý následujícím slovem. S rostoucí perplexitou tedy klesá kvalita jazykového modelu. Nejnižší možnou hodnotou je 1 - nastane v případě, kdy si je jazykový model zcela jistý budoucností (prakticky nedosažitelné).
Ačkoliv z hlediska teorie informace je každé snížení perplexity užitečné, v praxi se ukazuje, že pro pochopení přirozeného jazyka je perplexita spíše pomocným kritériem. Podstatně větší vypovídací hodnotu o kvalitě jazykového modelu má jeho začlenění do systému, který již s přirozeným jazykem v nějaké formě pracuje - příkladem může být rozpoznávač řeči nebo systém pro automatický překlad.
Statistické jazykové modely vychází pouze z informace obsažené v trénovacích datech - *korpusech*. Je nutné si uvědomit, že kvalita výsledného rozpoznávače nezávisí pouze na *množství dat* v korpusu, ale především na jejich *shodě s cílovou doménou*. Množství užitečné informace obsažené v určitém korpusu můžeme tedy poměrně spolehlivě určit tak, že jej začleníme do rozpoznávače, pro který budeme mít vhodná testovací data.

3. Český LVSCR systém

Akustické modely byly trénovány na 56 hodinách foneticky vyvážených vět z databází SpeeCon a Temic (16 kHz vzorkovací frekvence, čtená data). Akustickými parametry je 13 PLP (Perceptual Linear Prediction) cepstrálních koeficientů s odhadem prvních a druhých derivací v čase, celkem 39 koeficientů. Je provedena normalizace střední hodnoty a rozptylu (cepstral mean and variance normalization). Akustické modely jsou kontextově závislé skryté Markovoy modely (hidden Markov models - HMM) se třemi vysílacími stavy. Parametry stavů jsou navzájem provázány pomocí rozhodovacího stromu založeného na fonetických otázkách. Systém byl trénován pomocí toolkitu HTK [HTK] s kritériem maximalizace věrohodnosti (maximum likelihood) na trénovacích datech. Výsledné modely mají 34 gaussovských komponentů v jednotlivých stavech (Gaussian Mixture models – GMM) a stavy fonémů závislých na kontextu sdílí v systému 2604 takových směsí Gaussových rozdělení.
Jazykové modely jsou sestaveny pomocí běžně využívaného toolkitu SRILM [SRILM]. Použité jsou trigramové modely s Good-Turing vyhlazováním. Parametry modelů jsou optimalizovány na held-out datech (nejsou obsažena ani v trénovacích, ani v testovacích datech).
Pro účely rozpoznávání řeči je nutné sestavit *výslovnostní slovník*, který každému slovu přiřadí jeho nejpravděpodobnější výslovnostní variantu. Pro tyto účely je využíván software `transc` z ČVUT v Praze [Pollak2002].

4. Data a experimenty

Testovací data tvoří záznam jedné přednášky předmětu Signály a systémy z Fakulty informačních technologií VUT v Brně. Celkem je v těchto datech obsaženo 8953 slov rozdělených do 873 úseků - promluv, které zhruba odpovídají větám. Cílem rozpoznávání je tedy spontánní čeština s odbornými výrazy z prostředí vysokých škol.
Trénovací data jsou tvořena těmito běžně dostupnými korpusy: velký český textový korpus „all“ z Fakulty informatiky Masarykovy univerzity v Brně (FIMU-all) a Pražský a Brněnský mluvený korpus [PMK2001, BMK2002]. Data z části FIMU-all byla využita pro vytvoření

základního jazykového modelu (dále FIMU-all korpus). Tento korpus obsahuje psané texty, jako jsou novinové články, knihy ap. Data z mluvených korpusů naopak vhodně reprezentují spontánní mluvenou řeč (dále PBM korpus).
Pro účely rozpoznávání přednášek byly vytvořeny korpusy obsahující *doménová data*. První je sestaven s textových materiálů, které byly poskytnuty Masarykovou univerzitou (dále MU korpus). Obsahuje poznámky a textové materiály používané pro výuku odborných předmětů. Druhý byl vytvořen na FIT VUT Brno a sestává se z textů studijních opor (dále FIT korpus).

Tabulka 1: Trénovací data. Slovník FIMU-all je omezen na slova, která se vyskytla alespoň 30x.

	počet slov	**slovník**
FIMU-all	500 miliónů	360 900
PBM	1 170 000	69 300
MU	184 000	23 300
FIT	1 732 000	70 100

Pro určení úspěšnosti se definuje jednoduché kritérium - Word Error Rate (WER), které lze spočítat jako

$$WER = \frac{i + s + d}{n}$$

kde *n* je počet slov v testovacích datech a *i, s* a *d* představují počty slov která byla ve výsledku rozpoznávání navíc (Insertion), byla zaměněna (Substitution) a smazána (Deletion). Dále lze jednoduše definovat přesnost rozpoznávání (accuracy, ACC) jako *1–WER*. Protože je čeština ohebný jazyk, je vhodné počítat chybu nejen pro slova, ale i pro znaky - chyba v koncovce je tak méně dramaticky hodnocena, stejně jako různé synonymní varianty slov s podobnou výslovností a stejným významem.

Tabulka 2: Perplexita, počet slov mimo slovník (OOV) a úspěšnost na slovech a znacích

	Perplexita	**OOV**	**ACC (slova)**	**ACC (znaky)**
CNK	1035.9	7.8%	60.13%	82.71%
CNK+PBM	690.6	7.2%	60.98%	83.78%
CNK+MU	932.5	4.6%	62.58%	84.16%
CNK+FIT	967.7	3.4%	63.18%	84.18%
Všechno	**606.9**	**2.6%**	**65.99%**	**86.09%**

Výsledky ve formě perplexity, procenta slov mimo slovník (OOV-rate) a úspěšnosti rozpoznávání na slovech a znacích jsou uvedeny v Tabulce 2. Je zřejmé, že použití mluvených korpusů je pro stavbu slovníku a trénování jazykové modelu vhodné, avšak podstatné snížení perplexity není koherentní s relativně malým posunem v úspěšnosti rozpoznávání. Použití dat cílové domény (byť textových) nenabízí spektakulární snížení perplexity, avšak citelně se projevuje na úspěšnosti rozpoznávání. Data z FIT překonávají data z MU díky velikosti korpusu, a pravděpodobně i větší blízkostí k doméně přednášky. Není překvapující, že nejlepší výsledky získáme kombinací **všech** dat.

5. Závěr

Rozpoznávače řeči využívají praktickým způsobem lingvistické zdroje. Naše výzkumná skupina je potěšena faktem, že v České republice vznikají kvalitní korpusy mluvené češtiny, protože lingvistickou informaci v nich obsaženou nelze najít v korpusech psaného textu.

Provedené experimenty prokázaly užitečnost těchto zdrojů, které ve spojení s daty z cílové domény vedou k podstatnému zlepšení přesnosti rozpoznávače.
Zároveň jsme přesvědčeni, že automatického rozpoznávače jsou velmi dobrou pomůckou při studiu korpusů:

1. lze s nimi provést podrobnější analýzu zvukového signálu včetně zarovnání na slova, fonémy či fonémové stavy, věrohodnost při zarovnání lze použít při kontrole kvality anotací.
2. posun přesnosti rozpoznávače lze použít jako kritérium užitečnosti korpusu pro danou doménu.

Praktický výsledek – rozpoznávač mluvené řeči – je použitelný ve dvou směrech: v pokračování základního výzkumu rozpoznávání spontánní mluvené řeči, ale i v praktických aplikacích použitelných v mnoha odvětvích (bezpečnost, obrana, analýza chování zákazníků call-center, atd.).

LITERATURA

[Bates2004] T. Bates, D. Segarra: "e-Learning should be used strategically and not just as a tool that everybody uses", interview at elearningeuropa.info, http://www.elearningeuropa.info/index.php?page=doc&doc_id=5943&doclng=6&menuzone=1&focus=1

[BMK2002] *Český národní korpus – Brněnský mluvený korpus BMK*, 2002. Filozofická fakulta Masarykovy univerzity v Brně, http://ucnk.ff.cuni.cz .

[HTK] S. Young et al.: The HTK Book, Cambridge University Engineering Department, version 3.4, 2006, http://htk.eng.cam.ac.uk/

[IEEE] IEEE Computer Society Distance Learning Campus. http://bell.computer.org/distancelearning/index.htm

[Woodland2002] P. C. Woodland: The development of the HTK Broadcast News transcription system: an overview, Speech Comm. Vol 37, No. 1-2, 2002, ISSN 0167-6393, pp. 47—67.

[PMK2001] *Český národní korpus – Pražský mluvený korpus PMK*, 2001. Ústav Českého národního korpusu FF UK, Praha. http://ucnk.ff.cuni.cz .

[Pollak2002] P. Pollák and V. Hanžl: Tool for Czech pronunciation generation combining fixed rules with pronunciation lexicon and lexicon management tool. In Proc. of LREC'02, Third International Conference on Language Resources and Evaluation. Las Palmas, Canary Islands - Spain.

[SRILM] Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. International Conference on Speech and Language Processing, 2002, pp. II: 901–904.