Příprava dat pro online adaptaci LM

Jan Lehečka[1]

1 Úvod

V úloze automatického rozpoznávání mluvené řeči (automatic speech recognition, ASR) je jednou z nejdůležitějších komponent celého systému tzv. jazykový model (language model, LM), který definuje slovní zásobu rozpoznávače a matematicky popisuje vztahy mezi jednotlivými slovy. Pokud je v rozpoznávané promluvě řečeno slovo, které ve slovní zásobě ASR chybí, rozpozná se nějaké akusticky podobné slovo (či více slov), které rozpoznávač zná, ale jehož význam je zpravidla zcela jiný a rozpoznaný text pak nedává smysl.

Slova, která nejsou správně rozpoznána, protože chybí ve slovní zásobě ASR, se nazývají OOV (out-of-vocabulary). Nízký počet OOV slov je proto přirozený a důležitý požadavek na kvalitní ASR. V praxi ale rozpoznávaná promluva není předem známá, často dokonce není známé ani téma promluvy či její doména (obor). Příkladem takové úlohy, ve které se téma promluvy rychle mění a nelze jej předvídat, je rozpoznávání živě vysílaného televizního zpravodajství, jehož výstup (automaticky generované titulky k aktuální promluvě) může sloužit např. neslyšícím divákům.

Každá doména má svá specifická jazyková pravidla a specifickou slovní zásobou. Hovoří-li někdo o událostech v poslanecké sněmovně, je zřejmé, že bude používat zcela jinou slovní zásobu, než moderátor popisující konflikt na Ukrajině. Aby se mohl rozpoznávač přizpůsobit aktuálnímu obsahu promluvy, provádí se tzv. adaptace LM na určitou doménu. Doména se odhaduje z dosud rozpoznaného textu a adaptace je zpravidla realizována mícháním několika jazykových modelů, obvykle obecného LM (popisuje pravidla běžného jazyka) a tématického LM (popisuje specifická pravidla jazyka domény a obsahuje i příslušnou slovní zásobu).

Postup pro získání takových tématických jazykových modelů, konkrétně modelů, které se snaží co nejlépe pokrýt slovní zásobu náležící k určité světové geografické oblasti, popisuje tato práce.

2 Zdrojové texty

Jazykové modely jsou generovány z textu. Zdrojové texty pro každý LM tedy musí odpovídat zvolenému regionu. Celý svět byl pro tyto účely rozdělen na regiony, které budou vnímány každý jako jedna samostatná doména. Každý region sestává z jednoho nebo více států. Při rozdělování byly státy shlukovány podle významnosti vzhledem k poloze či historii České republiky (např. sousední státy jsou z tohoto hlediska významnější než vzdálenější státy, proto tvoří většina z blízkých států samostatný region), dále podle společného jazyka, polohy či historie. Celý svět byl takto rozdělen do 32 regionů a každý region představuje jednu doménu, pro kterou bude vytvořen samostatný jazykový model.

Aby výsledný jazykový model nebyl příliš velký, ale zároveň co nejlépe pokrýval slovní zásobu mluvčích hovořících o daném regionu, musí být zdrojové texty pečlivě vybírány. Vhodné texty pro tuto úlohu jsou především knihy, konkrétně cestovatelské průvodce, cestopisy a knihy popisující historii daného regionu. Z hlediska jazykového modelování je

[1] student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, e-mail: jlehecka@kky.zcu.cz

důležité, aby byly významné pojmy použity ve větách a ne jen heslovitě, jak tomu bývá v mnoha cestovatelských průvodcích. Jen velmi málo těchto textů je k dispozici v elektronické podobě, proto je většina knih nejprve skenována a text knih je z naskenovaných obrázků generován pomocí OCR (optical character recognition).

V knihách ale chybí aktuální významné pojmy (jména současných politiků, sportovců, míst, o kterých se začalo mluvit teprve v nedávné době atd.). Proto jsou jako další zdroje textů využívány články z českých zpravodajských serverů týkající se daného regionu (téma každého článku je detekováno automaticky).

3 Zpracování textu

Zdrojové texty nemohou být přímo použity pro vytvoření jazykových modelů, protože obsahují značné množství chyb (chyby OCR, překlepy ve zpravodajských článcích, …). Zároveň je v textech velké množství cizích slov s cizí výslovností (zejména vlastní jména osob, geografické údaje, …). Cizí výslovnost je nutné do rozpoznávače dodat, aby slova mohla být správně rozpoznána. Zdrojové texty jsou proto zpracovávány následujícím způsobem:

1. **vyčištění textů** – vymazání nežádoucích znaků,
2. **kontrola celistvosti vět** – odstraní nekompletní věty, nadpisy, tabulky … (delší úseky textu jsou před zahozením ručně kontrolovány, zda v nich není nějaký použitelný text)
3. **tokenizace** – oddělení interpunkce od slov,
4. **normalizace** – převod číslovek na slova, rozvoj některých zkratek atd.,
5. **truecasing** – zrušení velkého písmena na začátku vět, ponechání velkých písmen pouze tam, kde máme jistotu, že se slovo má psát vždy s velkým písmenem,
6. **náhrada podle slovníků** – nahradí některá definovaná slova za požadované tvary,
7. **vytvoření slovníku** – seznam všech slov v textu,
8. **vytvoření rozdílového slovníku** – obsahuje jen slova, která nejsou v žádném již zkontrolovaném slovníku,
9. **předání rozdílového slovníku anotátorům**, kteří ve slovníku ručně opraví chyby a doplní výslovnost k cizím slovům,
10. **promítnutí oprav zpět do zdrojových textů**,
11. **vytvoření výslovnostního slovníku a jazykového modelu**.

V současné době je práce ve fázi anotace slovníků, která představuje časově nejnáročnější část.

Vytvoření jazykového modelu pro každý region a jejich online přimíchávání do rozpoznávače podle obsahu aktuální promluvy by mělo výrazným způsobem snížit počet OOV a zlepšit kvalitu rozpoznávání tématicky nehomogenních promluv.

Tato práce se omezuje pouze na geografické oblasti. Pro další zlepšování rozpoznávače je v plánu vytvořit další jazykové modely, které budou pokrývat slovní zásobu různých profesí, vědních oborů atd.

Literatura

Pražák, A., Müller, L., Psutka Josef V.. and Psutka, J. : LIVE TV SUBTITLING - Fast 2-pass LVCSR System for Online Subtitling . *SIGMAP 2007*, p. 139-142, INSTICC PRESS, Lisabon, 2007.

Psutka, J., Müller, L., Matoušek, J. and Radová, V. : *Mluvíme s počítačem česky*. Academia, Prague, 2006.

Stolcke, A.: SRILM - an extensible language modeling toolkit. *INTERSPEECH* 2002.