ACQUIRING DATA FOR LANGUAGE MODEL TRAINING

Michal Jurzykowski
Bachelor Degree Programme (1), FIT BUT
E-mail: xjurzy01@stud.fit.vutbr.cz

Supervised by: Jan Černocký
E-mail: cernocky@fit.vutbr.cz

ABSTRACT

Rising computing power gives us opportunities to comunicate with computer in a way of human speech. Human speech synthesis and recognition are developed but more attention is given to recognition. Speech recognition is based on acoustic and language modeling. Language models are build on the principle of probability distribution over word sequences. According to this probability the most suitable word is selected. Newly developed language models are tested with perplexity at first and finally with speech recognition itself. The performance is analysed from achieved recognition accuracy.

1 ÚVOD

Touha po tom, aby jsme mohli s počítačem komunikovat prostřednictvím řeči, nejlépe v mateřském jazyce je zde již velmi dlouho. Ale prostředky a především metody jak toho dosáhnout jsou teprve ve vývoji. Hledají se různé cesty jak tohoto cíle dosáhnout. V dnešní době již není syntéza umělé řeči problém, vždyť se s ní můžeme potkávat na každém rohu. Čekáme-li na vlakovém nádraží, při jízdě tramvají a v mnoha dalších komerčních využití. Ale obráceně je to zatím problém. Úkolem rozpoznávaní řeči je vyvinout takové metody, které nám umožní řeč ve formě analogového signálu převést na písmena, slabiky, slova a následně celé věty. Tento proces využívá více prostředků k dosažení svého cíle. Jedním z nich jsou jazykové modely.

2 JAZYKOVÝ MODEL

V dnešní době neexistují žádné univerzální jazykové modely. Výzkumníci zabývající se zpracováním řeči zpravidla používájí jazykové modely od jiných skupin nebo si vytvářejí vlastní jazykové modely. Pro tvorbu těchto modelů se používají trénovací data, která jsou ve formě textu. K vytvoření jazykového modelu je zapotřebí velké množství trénovacích dat, které se obtížně shání a většinou nejsou ve vhodném formátu. Další problém je tematický okruh trénovacích dat. Trénovací data získaná z běžné hovorové mluvy se nehodí pro vytváření jazykových modelů na rozpoznávaní odborné řečí a naopak.

Jazykový model slouží k modelování posloupnosti slov. Skládá se z N-tic slov *–N-gramu–* přičemž každé N-tici je přiřazená pravděpodobnost výskytu slova podmíněná výskytem $N-1$

předchozích slov. Nejčastěji používané n-gramy jsou mono-gramy, bi-gramy a tri-gramy. Jazykový model můžeme pak definovat jako pravděpodobnostní rozdělení sekvence slov. K vytváření jazykových modelů byl používán program *ngram-count* od SRI International s implicitním nastavením. Jazykovým modelem můžeme odhadnout, která slova mohou následovat za sekvencí slov na základě pravděpodobností n-gramů. Tyto odhady slov pak dále využíváme.

3 ZPRACOVÁNÍ DAT

Pro kvalitní jazykový model je zapotřebí mít vhodné trénovací data. Data, která nám poskytla firma seznam.cz, jsou tvořena textovým obsahem českých webových stránek. Tento text v původní formě se nehodí pro trénování jazykových modelů. Obsahují velké množství nevhodného textu jako nabídková menu, navigační lišty, odkazy, tabulky atd. Vzhledem k velkému množství dat není v lidských silách zkontrolovat takto získané texty. Celkové množství dat poskytnuté firmou seznam.cz bylo okolo 50GB textových souborů.

Proto byly vytvářený skripty, pomocí nichž se má vydolovat vhodný text pro trénování. Skripty jsou napsány v programovacím jazyku Python. Jejích úkol je odstranit všechen nevhodný text a získat celé věty, které se v textu vyskytují. Skritpy na základě vytvořených pravidel nejdříve odstraní z textu všechen odpad. Pravidla byla odvozená na základě formátu dat od seznamu. Následovně se ve zbylém textu vyhledávají celé věty. Nakonec se každá věta opět zkontroluje a je-li korektní převede se do požadovaného výstupního formátu a uloží.

Bohužel není možné stoprocentně vydolovat pouze vhodný text. Na druhou stranu je množství získaných trénovacích dat velké a množství nevhodného textu je velmi malé, a proto můžeme tuto skutečnost zanedbat. Velikost získaných trénovacích dat je 7GB, tedy okolo 14%.

4 TRÉNOVÁNÍ A TESTOVÁNÍ JAZYKOVÝCH MODELŮ

K vyladění pravidel dolování a správnosti skriptu se používál malý vzorek dat, který se nejdříve zpracoval, pak se z něj vytvořil jazykový model. Vzniklý jazykový model se pak testoval programem *ngram* od SRI International. Výstupní hodnoty programu jsou perplexita a *OOV* neboli out of vocabulary.

Perplexita je pomocný ukazatel správnosti jazykového modelu, ale nemůžeme jí brát jako odpovídající. Až samotné testování při rozpoznávaní řeči nám ukáže jak „dobrý“ jazykový model je. Ale pro testování jazykového modelu je to rychlý a odpovídající ukazatel. Perplexita se vypočítává z pravděpodobností n-gramu v jazykovém modelu. Matematicky je to geometrický průměr inverzní hodnoty podmíněné pravděpodobnosti po sobě jdoucích slov a je dána vztahem

$$PPL = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i|w_{1..i-1})}} \tag{1}$$

kde n je celkový počet slov, w_i představuje i-té slovo.

Nižší hodnota perplexity znamená menší větvení a tedy ukazuje, že jazykový model se podobá testovanému textu. Ale ne vždy nižší znamená lepší. Protože druhým ukazatel je *OOV*, který představuje počet slov, který se v jazykovém modelu nevyskytl. Těmto slovům se přiřazuje pravděpodobnost 1OO% a snižuje perplexitu. Proto i jazykové modely s vyšší perplexitou, ale s nižší *OOV* mohou být lepší. Nelze tedy přesně určit, který model je lepší.

Po vyladění skriptů se veškerá data zpracovala a vytvořil se jeden velký jazykový model *seznam*. Programem *ngram* se porovnával nový jazykový model s všeobecným jazykovým modelem *general*, který má k dispozici Speech@FIT na obecné texty.

testovací data	ISS			IRP		
jazykový model	ppl	ppl1	OOVs	ppl	ppl1	OOVs
seznam	1066	3021	367	1081	1561	134
general	1233	3572	369	1183	1717	144
sloučení seznamu a general	1111	3156	347	1093	1578	108

Tabulka 1: Závislost perplexity a OOV na jazykovém modelu

Jazykové modely byly testovány na přepisech přednášek z Fakulty informačních technologií. Jedna se o ISS - Signály a systémy a IRP - Řízení projektů informačních systémů. Výstupní údaje jsou ppl - perplexita, ppl1 - normovaná perplexita a *OOV* - out of vocabulary. Nakonec se jazykové modely porovnávaly při rozpoznávaní řeči prostředky, které má k dyspozici Speech@FIT.

jazykový model	správnost [%]
seznam	65,42
general	64,25
sloučení seznamu a general	64,95

Tabulka 2: Úšpěšnost jazykových modelů při rozpoznávaní řeči

5 ZÁVĚR

Účelem této práce bylo z poskytnutých textových dat od firmy seznam.cz, vytěžit vhodná data pro vytvoření jazykového modelu, který by měl rozšířit a vylepšit současné modely. Pro testování byly pouze dostupné technické řeči. Proto úspěšnost modelů je nižší. Ale i tak nově vytvořený model od seznamu má úspěšnost lepší přes jedno procento při rozpoznávaní řeči. Zvláštní je, že při kombinací obou modelů byla úspěšnost mezi úspěšnosti obou modelů, i když se očekávala vyšší. Co se týče perplexity i *OOV* je jazykový model od seznamu lepší. Při sloučení obou slovníku je perplexita někde mezi oběmi modely, ale má lepší *OOV*.

Zajímavé bude sledovat úspěšnost při rozpoznávaní obecné řeči, kde se očekává lepší uplatnění. Zde se až ukáže přínos nově získaných dat. Ale zatím všechno naznačuje, že nové data budou mít přínos. Dosažené výsledky úspěšnosti nejsou konečné.

REFERENCE

[1] The State of The Art in Language Modeling [online] [cit. 2008-02-29; čas 16:05] Dostupný z http://research.microsoft.com/ joshuago/lm-tutorial-public.ppt

[2] Young, S.: The HTK Book, Cambridge University Engineering Department, 2006, s. 368