LANGUAGE MODELING OF CZECH USING NEURAL NETWORKS

Tomáš Mikolov
Master Degree Programme, FIT BUT
E-mail: xmikol04@stud.fit.vutbr.cz

Supervised by: Pavel Smrž
E-mail: smrz@fit.vutbr.cz

ABSTRACT

Language models are used in many systems involving natural language processing, like speech and handwriting recognition. The most widely used techniques are based on back-off n-grams. However, it is commonly believed that this approach is insufficient. One of the best improvements over back-off language models has been achieved by using neural networks that project words onto a continuous space.

This work concentrates on comparison of standard 4-gram language model with modified Kneser-Ney smoothing and neural network, both trained on spoken corpora with 1M words. Significant improvements in perplexity are reported.

1. ÚVOD

Statistické jazykové modely jsou v současnosti důležitou součástí systémů, které zpracovávají přirozený jazyk – příkladem může být rozpoznávač spojité řeči nebo strojový překlad. Cílem jazykového modelu je vyhodnotit pravděpodobnost promluvy a to tak, aby promluva patřící do jazyka měla přidělenu vysokou pravděpodobnost. Tento úkol se pokládá za AI-complete, protože zjevně vyžaduje pochopení ohodnocované promluvy.

Mezi nejjednodušší, ale také nejúspěšnější přístupy pro jazykové modelování se řadí klasické n-gramové techniky [1]. Jejich zjevnou výhodou je nízká výpočetní náročnost umožňující vytvoření jazykových modelů založených na korpusech se stovkami miliónů slov.

Nicméně je zřejmé, že jazykové modely postavené na n-gramových statistikách problém AI neřeší. V posledních letech byla navržena celá řada rozšíření, jak do jazykových modelů dostat více pochopení přirozeného jazyka, s mírnými úspěchy. Příkladem těch úspěšných mohou být jazykové modely založené na třídách (class based language models), nebo jazykové modely využívající neuronové sítě. Oba zmíněné přístupy se snaží vylepšit počítané statistiky pomocí sdílení informací mezi blízkými symboly (většinou slovy).

Tato práce se zaměřuje právě na jazykové modely, které jsou založené na neuronových sítích. Jejich hlavní výhodou je, že jsou schopny zachytit jinou informaci obsaženou v přirozeném jazyce než klasické n-gramové techniky. Zkombinováním obou přístupů pak lze dosáhnout výrazných zlepšení v perplexitě [2]. Další výhodou je relativní jednoduchost im-

plementace umožňující snadné experimentování. Existují i implementace prokazující užitečnost takovýchto modelů v rozpoznávačích [3].

Mezi hlavní nevýhody neuronových sítí patří jejich vysoká výpočetní náročnost a nesnadná paralelizace výpočtů. Nicméně pro případy, kdy je trénovacích dat relativně málo (například přepisy mluvené řeči v češtině), lze natrénovat síť v rozumném čase.

Cílem této práce je vytvoření jazykového modelu, který by byl použitelný pro rozpoznávání spontánní české spojité řeči (například automatický přepis přednášek). Pro trénování jsou k dispozici Pražský a Brněnský mluvený korpus (dále PMK a BMK, dohromady cca 1 170 000 slov). Pro takto malý objem trénovacích dat je použití neuronových sítí velmi vhodné.

2. IMPLEMENTACE

Implementace jazykového modelu vychází z dřívějších prací v této oblasti [3], [4]. Základem jsou dvě neuronové sítě učené pomocí algoritmu backpropagation.

2.1. ARCHITEKTURA

První síť se učí převádět slova do spojitého prostoru tím, že se učí bigramový jazykový model (využívá tedy jen historii jednoho předchozího slova pro predikci následujícího). Vstupní a výstupní vrstvy mají velikost slovníku (např. 50 000), skrytá vrstva mezi nimi má velikost typicky 20-50 neuronů. Na vstupní vrstvě je použito kódování one-of-N, tedy aktivní je pouze neuron reprezentující poslední slovo v historii. Na výstupu je použita funkce softmax, která normalizuje výstupy skryté vrstvy tak, aby dohromady daly jedničku (tedy na výstupní vrstvě se objevuje rozložení pravděpodobnosti pro následující slovo). Učící algoritmus je backpropagation.

Po naučení první sítě je k dispozici možnost přepsání každého slova ze slovníku do spojitého prostoru o dimenzi dané počtem neuronů ve skryté vrstvě první sítě. Takto lze vytvořit delší kontexty, typicky 4-8 slov. Tato data jsou pak vstupem druhé sítě - použijeme-li tedy kontext pěti předchozích slov, pak velikost vstupu první sítě je 5*N, kde N je velikost skryté vrstvy první sítě. Velikost skryté vrstvy druhé sítě je typicky 30-100, výstupní vrstva je stejná jako v případě první sítě.

2.2. TRÉNOVÁNÍ

Na začátku trénování se nastaví rychlost trénování – používaná hodnota je 0.1. Trénování pak probíhá v etapách do té doby, dokud síť vykazuje znatelné zlepšení na validační sadě. Po té dochází ke snižování rychlosti trénování (algoritmus "new-Bob"); pokud nenastane zlepšení, je trénování ukončeno.

Pro zrychlení trénování jsou všechna slova vyskytující se v trénovací sadě méně než 3x spojena do jednoho virtuálního slova (při vyhodnocování pravděpodobnosti je pak těmto málo vyskytujícím se slovům přiřazena pravděpodobnost virtuálního slova dělená počtem slov takto spojených dohromady, tzn. uvažuje se zde uniformní rozložení pravděpodobnosti). Toto vede v případě PMK+BMK korpusu na redukci velikosti slovníku z 68 500 na 20 300 slov.

3. VÝSLEDKY

Vyhodnocení bylo provedeno porovnáním výsledků neuronové sítě a výsledků dosažených pomocí SRI LM toolkitu. Jako baseline byl použit 4-gramový jazykový model s modifikovaným Kneser-Ney smoothing jako vyhlazovací metodou. Pro natrénování jazykového modelu pomocí SRI LM toolkitu i neuronové sítě byla použita část Pražského a Brněnského mluveného korpusu (prvních cca 1 155 000 slov). Oba jazykové modely tedy byly trénovány na úplně stejných datech. Jako validační data pro řízení rychlosti trénování neuronové sítě byla použita následující část PMK+BMK korpusu (5 500 slov), jako testovací data zbytek (10 000 slov). Velikosti skrytých vrstev byly 30 neuronů pro první síť, 50 pro druhou, délka kontextu byla zvolena 5 slov. Doba trénování byla zhruba dva dny na počítači s procesorem AMD Opteron 2,8GHz.

	Validační sada	Testovací sada	Přepis přednášek
KN 4-gram	284.37	299.96	613.39
Neuronová síť	242.77	274.70	533.17
Interpolace	221.34	250.30	486.68

Tabulka 1: Výsledná perplexita

V řádku interpolace je uvedena perplexita po interpolaci výsledků z obou modelů. Je použito jednoduché lineární interpolování s mixovacím koeficientem odhadnutým na validačních datech. Výsledné jazykové modely byly také použity pro ohodnocení jedné přepsané přednášky (5 275 slov); perplexita n-gramového modelu byla v tomto případě o 26% vyšší než u modelu vzniklého interpolací.

4. ZÁVĚR

Dosažené výsledky jsou na stejné úrovni jako u předchozích prací v této oblasti [4], přičemž celková implementace je o něco jednodušší (dvě třívrstvé neuronové sítě namísto jedné čtyřvrstvé). V průběhu řešení tohoto projektu bylo vyzkoušeno více variant trénování a je zřejmé, že výsledky lze nadále zlepšovat v mnoha oblastech, především co se týče rychlosti trénování. Další výrazné zlepšení výsledků by bylo možno očekávat s lepší interpolační technikou, případně při trénování sítě na slovech i jejich částech (morfémech) dohromady.

LITERATURA

[1] Stanley F. Chen, Joshua T. Goodman. 1999. An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 13(4):359-394.

[2] Joshua T. Goodman. 2001. A bit of progress in language modeling. Microsoft Technical Report MSR-TR-2001-72.

[3] Holger Schwenk, Jean-Luc Gauvain. 2004. Neural network language models for conversational speech recognition. In ICSLP, pages 1215-1218, 2004.

[4] Yoshua Bengio, Rejean Ducharme, Pascal Vincent a Christian Jauvin. 2003. A neural probabilistic language model. Journal of Machine Learning Research, 3(2):1137–1155.