Česká fonetická transkripce pomocí neuronových sítí

Markéta Jůzová[1]

1 Úvod - Fonetická transkripce

Jedním ze základních bloků systému syntézy řeči z textu (angl. *text-to-speech*, TTS) je fonetická transkripce vstupního textu do výslovnostní podoby. A přestože čeština je na první pohled poměrně jednoduchý jazyk z hlediska čtení a výslovnosti, není pravdou, že se vše čte tak, jak je napsáno; např.:

- **ě** – čte se `[je]` nebo `[ňe]`
- **i**, **í** – způsobují změkčení některých souhlásek
- skupiny **ou**, **au**, **eu** – čtou se někdy jako dvojhláska, někdy jako dvě samohlásky
- asimilace místa artikulace – nazály **m**, **n** se vyslovují více vzadu v určitých kontextech
- asimilace znělosti – skupina souhlásek se celá čte buď zněle nebo nezněle
- slova cizího původu – neplatí pro ně standardní pravidla české výslovnosti

V současném TTS systému *ARTIC* vyvíjeném na naší katedře se pro fonetickou transkripci používá sada ručně navržených pravidel spolu se slovníkem výjimek (který zajišťuje správný přepis častých slov cizího původu v českých textech). Používaná fonetická pravidla je možné zapsat jako produkční pravidla ve tvaru (převzato z Psutka et al. (2006)):

JESTLIŽE	řetězci znaků A bezprostředně předchází řetězec znaků C a je bezprostředně následován řetězcem znaků D,
PAK	se A přepíše na řetězec znaků B

2 Použití neuronové sítě pro natrénování modelu fonetické transkripce

Hlavním cílem popisovaného experimentu bylo otestovat možnost natrénování neuronové sítě pro účely fonetické transkripce na základě velkého množství vstupních dat. K dispozici bylo cca 770 tisíc českých frází, doplněných o fonetický přepis, které obsahovali celkem 40 grafémů a 48 fonémů. Navržená neuronová síť pro úlohu fonetické transkripce má následující strukturu:

- *embedding* vrstva – transformuje všechny grafémy do vektorové podoby
- *biLSTM* vrstva – obousměrná LSTM vrstva, která pracuje nad posloupnostmi grafémů v rámci slov a měla by se naučit fonetické závislosti uvnitř jednotlivých slov
- *LSTM encoder* – vytvoří vektorové reprezentace slov
- vrstva pracující nad slovy – měla by se naučit mezislovní závislosti; 2 typy:
  - **biLSTM** – obousměrná LSTM vrstva
  - **conv** – konvoluční vrstva
- *LSTM decoder+softmax* – generuje posloupnost fonémů

[1] studentka doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, specializace Počítačová syntéza řeči, e-mail: juzova@kky.zcu.cz

Trénování sítě spočívá v postupném předkládání posloupností slov, kdy každé slovo se skládá ze stejného počtu grafémů (v případě kratších slov doplněno symbolem „-“). Výstupem je pak posloupnost slov tvořených fonémy.

3 Výsledky a závěr

Úspěšnost fonetické transkripce byla vyhodnocena pomocí míry *accuracy* na 20.000 testovacích frázích, a to na úrovni fonémů (*phoneme Acc*) a celých slov (*word Acc*). Výsledky jsou uvedeny v tab. 1 a příklad výstupu na obr. 1.

	phoneme Acc	*word Acc*
pouze pravidla	95,08 %	96,75 %
pravidla + slovník	98,72 %	**99,16 %**
model 1 (biLSTM)	**99,69 %**	98,92 %
model 2 (conv)	99,30 %	97,65 %

Tabulka 1: Porovnání úspěšnosti fonetické transkripce na 20.000 testovacích frázích.

```
word          correct                 predicted               is_ok?
až--------    ? a S - - - - - - -     ? a S - - - - - - -     True
nás-------    n a: z - - - - - - -    n a: z - - - - - - -    True
znudí-----    z n u J\ i: - - - - -   z n u J\ i: - - - - -   True
i---------    ? i - - - - - - - -     ? i - - - - - - - -     True
městský---    m J e s t_s k i: - - -  m J e s t_s k i: - - -  True
ruch------    r u x - - - - - - -     r u x - - - - - - -     True
```

Obrázek 1: Ukázka fonetické transkripce.

Testované modely pro fonetickou transkripci využívající neuronové sítě jsou úspěšnější (na úrovni slov i fonémů) v porovnání s přístupem, kdy by se pro transkripci použila pouze pravidla. Kombinovaný přístup využívající i slovník výjimek je však stále o něco lepší na slovní úrovni. To je dáno tím, že testované modely někdy selhávají u dlouhých, i když českých slov.

Z ukázky výstupu výše je ale zřejmé, že se testovaný model využívající neuronové sítě dokáže naučit základní pravidla pro fonetický přepis (např. změkčení kvůli `í`) i spodobu znělosti, dokonce i přes hranici slov. Podrobnější analýza ukázala, že je schopný se naučit i výslovnost slov cizího původu – i ve tvarech, které se v trénovacích datech nevyskytovaly. Závěrem mi dovolte poznamenat, že tento článek popisuje pouze prvotní experiment trénování neuronových sítí pro účely automatické fonetické transkripce. Probíhá testování dalších struktur neuronových sítí a jejich nastavení i aplikace na jiné jazyky (viz Jůzová, Vít (2019) a Jůzová et al. (2019)).

Poděkování

Příspěvek byl podpořen grantovým projektem číslo SGS-2019-027.

Literatura

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

Jůzová, M., and Vít, J., 2019. *Using Auto-Encoder BiLSTM Neural Network for Czech Grapheme-to-Phoneme Conversion.* Přijato na TSD 2019.

Jůzová, M., Tihelka, D., and Vít, J., 2019. *Unified Language-Independent DNN-Based G2P Converter.* Odesláno na Interspeech 2019.