Porovnání různých přístupů k frázování textu pro TTS

Markéta Jůzová[1]

1 Úvod

Pojmem *frázování* rozumíme dělení věty na kratší úseky, ucelené ve smyslu prozodie i významu (Palková (1974)). Hlavním důvodem frázování lidské řeči je lepší srozumitelnost promluvy, nezanedbatelná je také potřeba člověka se nadechnout. Velmi dlouhé věty bez pauz (hranic frází) jsou často vnímány jako nepřirozené a také jsou náročnější na pozornost posluchačů – proto je třeba zabývat se frázováním i pro účely systémů převodu textu na řeč (TTS).

Na detekci hranic frází lze pohlížet jako na klasifikaci klíčových okamžiků (mezi každými dvěma slovy) do dvou tříd: *break* a *no-break*. Protože člověk vkládá do řeči pauzu v určitých pravidelných intervalech – fráze se obvykle skládají z 3-6 slov, viz. Taylor (2009), zdá se vhodnější používat některý ze sekvenčních modelů, spíše než klasický klasifikační přístup vyhodnocující každé místo ve větě zvlášť.

Vkládání hranic frází souvisí se syntaktickou strukturou věty a pozicí interpunkce (převážně čárek). Různé studie také ukázaly, že konkrétní frázování věty je závislé na rychlosti a stylu mluvení a také na řečníkovi – proto jsou experimenty vyhodnocovány v sekci 3 pro každého řečníka zvlášť.

2 Porovnání přístupů

V rámci experimentů byly porovnávány následující přístupy:

- *Comma* – vkládá hranici fráze jen a pouze za každou čárku (používaný v TTS *ARTIC*)
- *SVC* – Support Vectore Machines (lineární) – parametry klasifikátoru byly nastavené na základě experimentů popsaných v Jůzová (2017a) (vybrán jako zástupce „klasických“ klasifikačních přístupů)
- *CRF* – Conditional Random Fields – sekvenční modelování (celá věta se vyhodnocuje najednou); přístup byl prezentován v Jůzová (2017b)
- *MLP* – Multi-layer Perception s 30 neurony ve vstupní vrstvě a 100 neurony ve skryté vrstvě (všechny plně propojené); trénováno 100 epoch (použito v Jůzová (2018))
- *LSTM* – neuronová síť s dvěma LSTM vrstvami, každá obsahuje 200 jednotek, výstupní plně propojená vrstva; trénováno 100 epoch (použito v Jůzová (2018))

K získání trénovacích a testovacích dat byly použity 2 české řečové korpusy nahrané pro účely TTS *ARTIC* vyvíjeném na naší katedře – tato data byla nahrána profesionálními řečníky, u kterých předpokládáme, že vkládají do řeči pauzu v „rozumných“ místech. Jako informace o

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

hranici fráze proto byly využity všechny pauzy a nádechy v korpusech, spolu s pozicí čárek v nahrávaných větách.

Pro samotné trénování byla použita sada příznaků obvyklá pro úlohu detekce frází: předchozí a následující slovo a jejich morfologické tagy, délka věty, pozice slova ve větě, vzdálenosti od nejbližší předchozí/následující čárky, zda předcházející slovo má čárku apod.

3 Výsledky a závěr

Výsledky porovnání jednotlivých přístupů jsou prezentovány v Tab. 1 pomocí standardních měr *accuracy* (A), *precision* (P), *recall* (R), *F1 skóre* ($F1$) a počtů *true positives* (tp), *true negatives* (tn), *false positives* (fp) and *false negatives* (fn).

data	přístup	tp	fn	fp	tn	A	R	P	$F1$
korpus1	*Comma*	2407	781	0	75472	0.990	0.755	1.000	0.860
	SVC	2784	404	134	75338	0.993	0.873	0.954	0.912
	CRF	2857	331	170	75302	0.994	0.896	0.944	**0.919**
	MLP	2521	667	95	75377	0.990	0.791	0.964	0.869
	LSTM	2344	844	269	75203	0.986	0.735	0.897	0.808
korpus2	*Comma*	2319	109	0	62534	0.998	0.955	1.000	0.977
	SVC	2336	92	1	62533	0.999	0.962	1.000	0.980
	CRF	2343	85	3	62531	0.999	0.965	0.999	0.982
	MLP	2361	67	0	62534	0.999	0.972	1.000	**0.986**
	LSTM	2189	239	152	62382	0.994	0.902	0.935	0.918

Tabulka 1: Srovnání přístupů k frázování textu trénovaných na českých řečových korpusech

Experimenty ukázaly, že testované neuronové sítě většinou nedosahují lepších výsledků v porovnání s *Conditional Random Fields* – důvodem může být nedostatečný počet trénovacích dat (velká trénovací množina bývá podmínkou funkčnosti mnoha neuronových sítí). V budoucnu plánujeme otestovat i další struktury sítí (i rekurentní) a vyzkoušet také použití číselné reprezentace slov (tzv. *word embedding*).

Poděkování

Příspěvek byl podpořen grantovým projektem číslo SGS-2016-039.

Literatura

Palková, Z. (1974) *Rytmická výstavba prozaického textu*. Studia ČSAV, Academia, Praha, 1974.

Taylor, P. (2009) *Text-to-Speech Synthesis*. Cambridge University Press, New York, USA, 2009.

Jůzová, M. (2017) *Prosodic Phrase Boundary Classification Based on Czech Speech Corpora*. TSD 2017, Lecture Notes in Computer Science, vol. 10415, pp. 165-173, Springer, 2017.

Jůzová, M. (2017) *CRF-Based Phrase Boundary Detection Trained on Large-Scale TTS Speech Corpora*. SPECOM 2017, Lecture Notes in Computer Science, vol. 10415, pp. 317-325, Springer, 2017.

Jůzová, M. (2018) *On the Comparison of Different Phrase Boundary Detection Approaches Trained on Czech TTS Speech Corpora*. Odesláno na SPECOM 2018.