Model pro rozpoznání diktovaných čísel pro společnost poskytující Interactive Voice Response (IVR)

Ing. Martin Nykodem, Mgr. Alexander Kovalenko, Ph.D.

Fakulta informačních technologií
České vysoké učení technické v Praze

Motivace

Mezilidská interakce se nejvíce zprostředkovává verbální komunikací, jež se začíná prosazovat i jako jedna z možností interakce mezi člověkem a počítačem. Nicméně ještě stále nedosahuje lidských kvalit ve všech oblastech a u jazyků s menší popularitou je rozdíl ještě hlubší. Děje se tak převážně z důvodu, že momentálně v oblasti strojového rozpoznání řeči dominuje pár technologických firem, pro které menší trh není tak zajímavý. Navíc nabízená řešení jsou obecná a neumožňují jednoduchou aplikaci i pro případy rozpoznání z omezené domény.

V rámci diplomové práce řešíme praktický problém rozpoznání diktovaných čísel českého jazyka na úroveň lepší, než na jaké jsou momentálně běžně dostupná řešení. Zároveň vybudovaný framework je aplikovatelný na jakékoliv rozpoznání řeči z omezené domény.

Praktické využití

- Ve voicebot řešeních je doména rozpoznání řeči předem známá.
- Číselný vstup je součástí naprosté většiny voicebot řešení komunikací (klientské číslo, číslo účtu/zásilky, RČ, PSČ atd.).
- Takový údaj bývá často nezbytný k úspěšnému odbavení.
- Zvýšení kvality rozpoznání tak může zlepšit zkušenosti (nejen) s voiceboty nás všech.

- *IVR:* ... prosím diktujte.
- *Klient:* Dva, sedm, sedm, osm, tři devět, čtyři, jedna, nula, nula.
- *IVR:* Promiňte, ale slyšel jsem devět cifer správně by však číslo mělo mít deset cifer. Zkuste prosím diktovat znovu.
- *Klient:* Dva sedm
- *Přerušení od IVR...*
- *IVR:* Promiňte, ale slyšel jsem jedna cifru.
- *Klient:* Ježiš tak ale poslouchej ne?!!?
- *IVR:* Správně by však číslo mělo mít deset cifer, zkuste prosím diktovat znovu.
- *Klient:* Dva, sedm, sedm, osm, tři devět, čtyři, jedna, nula, nula.
- *IVR:* Číslo se mi nepodařilo rozpoznat vyčkejte prosím na lince přepojím vás na živého kolegu.

Cíle práce

- Obstarání vhodných datasetů pro rozpoznání řeči
- Rešerše a analýza současných přístupů pro rozpoznání řeči
- Zvolení vyhovující architektury pro specifické účely rozpoznávání řeči z telefonních hovorů
- Předzpracování a očištění datasetů
- Implementace a natrénování systému

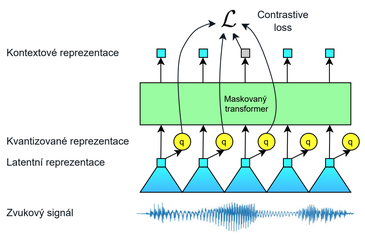
Řešení stavit tak, aby bylo možné jej případně nasadit do produkce a zároveň jednoduše modifikovat na další domény jako je rozpoznání diktovaných adres, jmen či dalších běžných údajů při komunikaci firmy se zákazníkem. Navíc musí být zohledněny licence dat, programů a algoritmů, aby bylo možné komerční využití.

Přístup k řešení

1 Volba architektury

Na základě analýzy jsme zvolili jako základ pro náš model architekturu Wav2vec2.

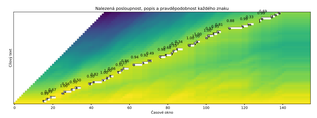
- Model této architektury je možné předtrénovat sebesupervizovaným učením pouze na nahrávkách bez přepisu a následně dotrénovat na menším datasetu obsahující i přepisy nahrávek.
- Důležité pro český jazyk, jelikož pro něj není dostupné velké množství ASR datasetů.

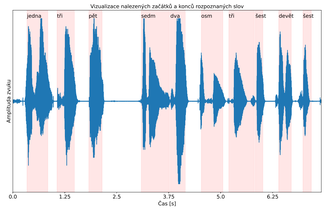


2 Implementace

Poskytnuté přepisy obsahu nahrávek byly ve většině poskytnuty jako posloupnost čísel. Nelze z nich tedy poznat v jakém formátu bylo slovo dva/dvě vysloveno. Při použití čísel vyšších než jednociferných řádů není jednoznačné určení obsahu nahrávky. Dále obsahovaly systematické chyby i obecně chybné přepisy.

- Pro vyřešení problému jsme přišli s frameworkem pro iterativní vývoj na nevěrohodných datech: postupným trénováním modelu na nekvalitních datech a jejich následném opravení pomocí natrénovaného a nezávislého modelu jsme dostali kvalitní trénovací množinu.





Pro praktické použití je vhodné predikovat slova ze zvoleného korpusu, ale povolit i predikce mimo doménu.

- Aplikovali jsme speciální způsob dekódování, unikátní pro tuto doménu. Ten má pokročilé možnosti parametrizace a umožnil nám tak vypořádat se s rozpoznáním slov nahrávek, ve kterých dochází k ztrátám signálu i nahrávek, kde se vyskytuje poměrně silný zvuk na pozadí.
- Současně jsme přidali pro náš dekodér přesné vrácení času začátků a konců rozpoznaných slov, což umožňuje rychlejší analýzy dat a přípravu dat pro další trénování. Dále přesně najít zamýšlené číslo při diktování čísel vyšších cifer při známé očekávané délce číselné posloupnosti.

Výsledky

Naše řešení jsme porovnali s aktuálně využívaným řešením a dalšími komerčními řešeními pro rozpoznání řeči českého jazyka. Testovací dataset tvořily nahrávky deseticiferné posloupnosti čísel po telefonu od společnosti poskytující voicebot řešení, které byly lidsky verifikovány. Rozpoznaný text z ostatních řešení je následně ještě postprocesován a při víceciferných číslech uznány i varianty, které šly převést na zamýšlenou posloupnost. Pro voicebot řešení je zapotřebí správně rozpoznat celou posloupnost. Zde jsme dosáhli desetinásobného zlepšení v procentu neúspěšně rozpoznaných nahrávek oproti stávajícímu nejlepšímu řešení postaveném na speciálně přetrénovaném řešení rozpoznání řeči na úlohu rozpoznání čísel od společnosti Microsoft.

Model	Špatně/Celkem	Nerozpoznáno %	NER
Anonymizováno 1 (CZ)	36/201	17.91%	0.0400
Anonymizováno 2 (CZ)	38/201	18.91%	0.0503
Microsoft$^+$	27/201	13.43%	0.0197
Google	160/201	79.60%	0.1333
Naše (dekodér)	3/201	**1.49%**	**0.0015**

Number Error Rate (NER), měří chybovost na úrovni cifer a naše řešení tedy zde nerozpoznalo pouze tři cifry ve třech nahrávkách.

Kromě jsme dosáhli nejnižšího skóre Word Error Rate (WER) z dostupných nekomerčních modelů pro doménově neomezenou datovou sadu pro český jazyk Common Voice 8.

Závěr a přínos

Ukázali jsme, že ačkoliv mají velké technologické společnosti už kvalitní rozpoznání mluvené řeči, je minimálně na omezené doméně možné přijít s řešením, jež dosahuje výrazně lepších výsledků.

Toho jsme dosáhli díky aplikaci speciálního způsobu dekódování pro omezenou doménu. Ten lze použít k rozpoznání různých domén, nejen čísel. Současně v rámci toho byl vybudován i framework pro transformaci nepřesných/nevěrohodných dat v kvalitní trénovací množinu.

Na základě výsledků byla licence k diplomové práci od školy odkoupena a založen start-up, jež na vybudovaných základech má v plánu přetransformovat v škálovatelné produkční řešení.