



Mašinelne spóznawanje češćiny. Předskok za Serbow?

Anja Nedoluzhko



Plan přednoška

- ELITR: što my činimy
- Česká republika a češćina: kak daloko smy z mašinelnym spóznawanjom
- Što móže wužitne za Serbow być

Projekt ELITR

- European Live Translator
- Džěl Horizon 2020
- Mašinelne podtitule žiwych zetkanjow, přeložk, mašinelne protokole zetkanjow
- Hłowny podawk: EUROSAT Kongres w Praze, 2020



CHARLES
UNIVERSITY



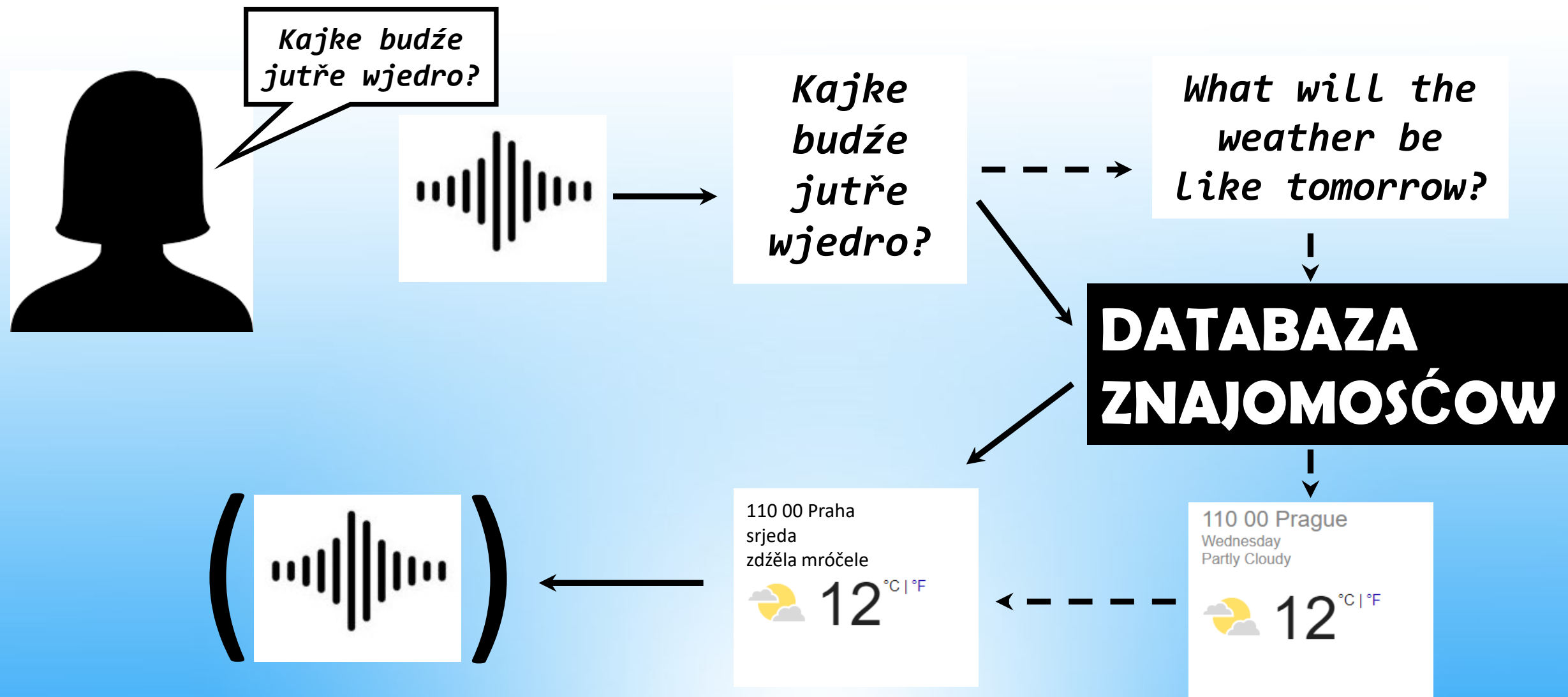
THE UNIVERSITY
of EDINBURGH

PerVoice

alfaview®

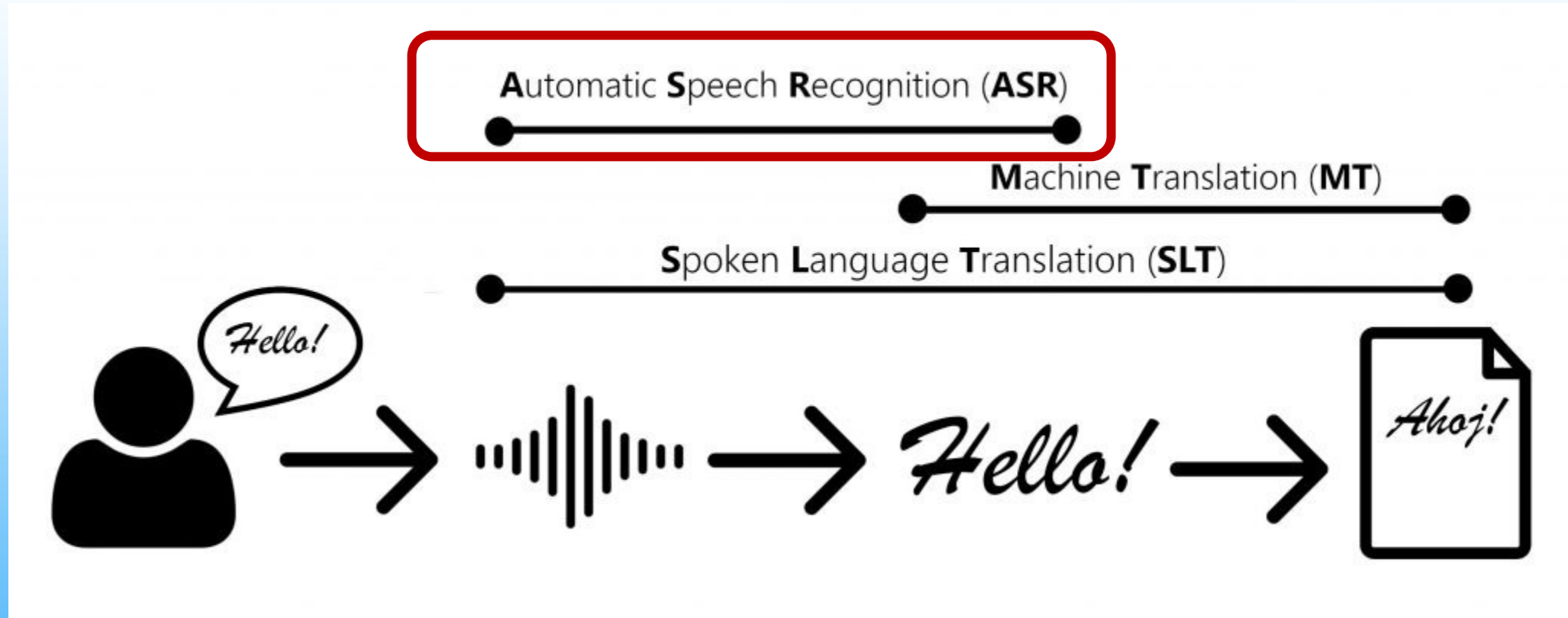


Rěčane wupytowanje (Voice search)





Što my w ELITRu za češćinu činimy

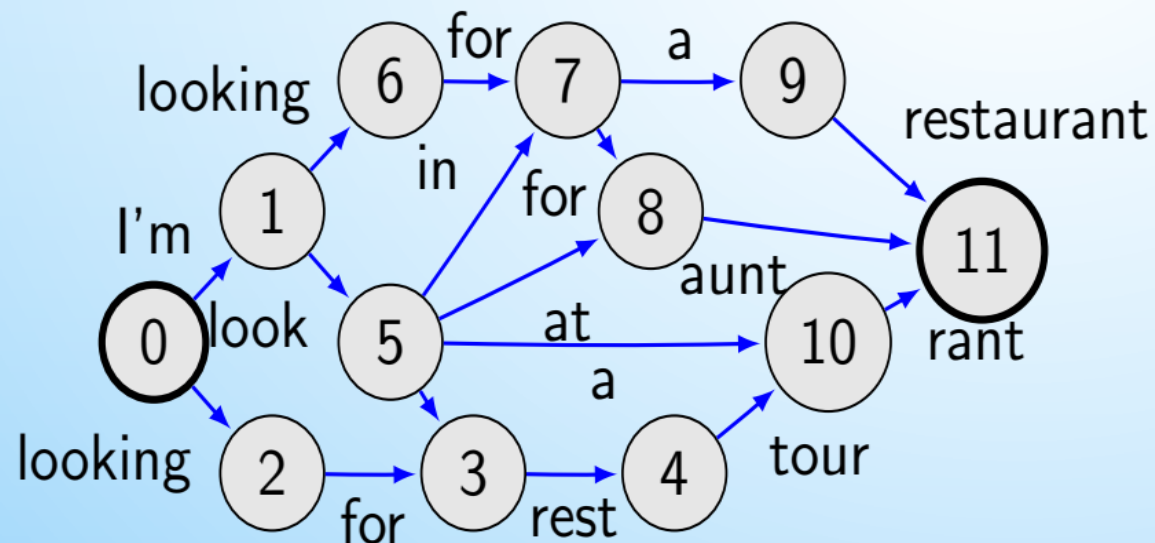


Mašinelne spóznawanje češćiny - teorija



$$\text{output} = \arg \max P(A|T)P(T)$$

- **Akustiski model:** $P(A|T)$
- **Rěčny model:** $P(T)$



0.8 I'm looking for a restaurant
0.4 I'm look in for aunt
0.2 looking for rest tour rant

senon → fonem → słowo → sada → rěčny model



Akustiski model, řečný model a slovník

- Džele zwuka: signale, 25 ms do the wotrězki
- Kóždy signal twori niskodimensionalny wektor, MFCC
- MFCC wektor dže přez akustiski model (wužiwamy neuralne syće), kotryž predikuje fonetiske jednotki
- Slovník z tych segmentow móže słowa tworić

several S EH V AX R AX L

- Slovník je jara wažna komponenta systema.
- Řečný model: n-grams

70%: European parliament in Brussels

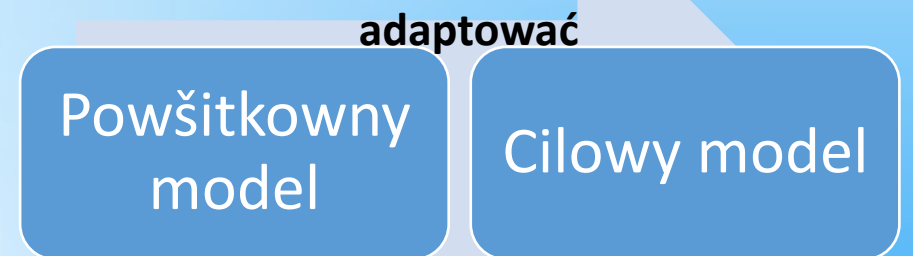
29%: European parliament in France

1%: European parliament in Zimbabwe



Wot teorije k realizaciji

- Trjebamy všelakore typy datow
 - Za akustiski model: zwukowe (awdijo) dataje z transkripciju
 - Za řečny model: ryzy pisomne teksty
- Rozdžěle w domenje!!!
 - Z nowinow: *Zaměr je sej nazhonjenja wuměnjeć, lěpšiny a njelěpšiny rozdžělnych modelow nastupajo organizowanje a financowanje mjeńšin w jednotliwych krajach rozważować a na tym zakładže wo móžnym zhromadnym puću rozmyslować.*
 - Handy: *Njechaš pola mje džensa nimo přińć?*





Wot teorije k realizaciji - šumy

- Specialny model za spóznawanje šumow
 - nadychowanje
 - smjeće
 - *ehmm, hmm..* njewěstota
 - pozadkowe zwuki
 - mjelčenje
- Je wažne to w datach spóznać a tež do anotacijow zadžětać
- Model dyrbi nawuknyć, zo někajki džěl zwukoweje žołmy je jenož mjelčenje a njetrjeba transkriběrowany być



Wot teorije k realizaciji – nowe słowa

- Słowa, kotrež w słowniku njejsu
- Mjena, pomjenowanja firmow, produktow, skróćenki, fachowa terminologija
- Hdyž model tute słowa we słowniku njenamaka, spyta najbliše akustiske słowa namakać a druhy je to jara žortne a njekorektne.

Referenca: kn. Brösanowa
Predikcija: kn. brěza nowa

Towarstwo Mohorjeva
Towarstwo mocha rjewjenja

Mamy małoh' syna. (małkeho syna)
Mamy mało syna. (mało sušeneje trawy)



Wot teorije k realizaciji – technika

- Technisce chětro naročne
- Trenowanje našich modelow sej žada dosahace wulku kompjuterowu syć





Eksperiment z firmami

- Překwapjenje z wuslědkom – realny vs. podawany
- State-of-the-art modely, kotrež wozjewja 60-procentowsku wuspěšnosť, druhdy scyła ničo pokazali njejsu

Reference	Google	UEDIN	KIT
why do you wear those high heels , if you would wear some sneakers ? I know one really good store , that deals with the sale of freetime footwear .	why do \emptyset where does high heels if you would wear some sneakers I no one really good star that deals with the sale of freedown food to our	're ready where tells us if you would \emptyset sneaker us I know won the really good store that deals with the sale of freedown food	why do are those highs heels if you would where some sneakers i no one really good story that deals with the sale of freetime food to our



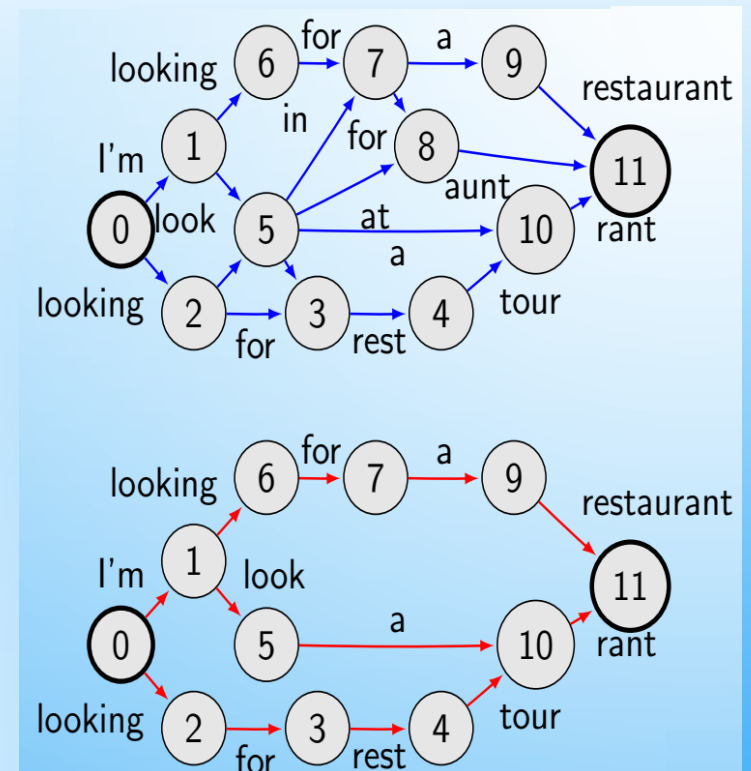
Offline vs. Online ASR

- Spěšnosť – kvalita outputa
 - Wupytowanie najlěpších wariantow w akustiskim modelu traje někajki čas
 - dyrbimy wobmjezować ličbu wariantow
 - to rěka, zo najlěpša warianta njetrjeba nutřka być
- Wulkosć a spěšnosť
 - Přewulki wobsah
 - Dyrbi spěšnje wobdžělać a z rezultatom přińć
 - output móže być pomałki
- Koncept parcielneje hypotezy

I'm

I'm look

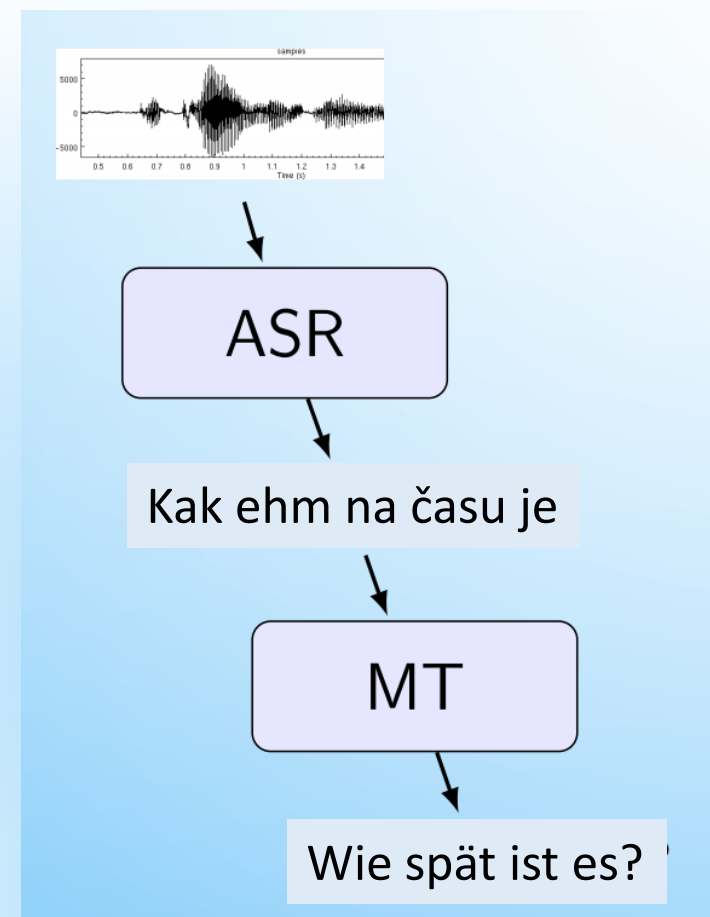
I'm looking for



Přeložk outputa mašinelneho spóznawanja



- Spomalenje
- Mašinelny přeložk so dyrbi na cilowu domenu adaptować
- Kontekstowe wokno: němsce so sada móže hakle na kóncu formować
 - Chcemy wužiwarjam cyły kontext poskić, tohodla přeco dawamy přechodne hypotezy
- Zawěsćić wěstotu (robustness) přeložka:
 - Hewak tón najmjenší zmylk w spóznawanju móže wulke změny w přeložku implikować
 - Eksperimentujemy z někotrymi ASR hypotezami a systemami MT
- Intergacija systemow ASR a MT





Wužiwanje

- Kak to prezentować wužiwarjam
- Wulke pismiki a interpunkcija

Input: HI IT'S JOHN CAN YOU TALK

Output: Hi, it's John. Can you talk?

- Zadžěwać nježadnemu wobsahej
 - wulgarne wurazy wufiltrować
- Hardware a problemy z WiFi
- Je dobre měć backup plan
- Je dobre so na tajkich akcijach najhusćišo wobdžělić a system wuspytać, dokelž kumštnje to njeńdže derje modelować





Što mamy w ELITRu – zjimanje

- Prěnje wersije modelow (ASR, MT)
 - output njeje idealny, ale je to započatk
- Tři dźěle ASR modela, kotrež móžemy kontrolować
- Modele su jara čućiwe na cilowu domenu a na rěčacu wosobu
- Online setup so jara rozeznawa wot offline setupa, a hišće bóle so změni, hdyž činimy tež přełožk
- Designowe rozrisanje – kak to wšo wužiwarjej předstajić



Štó, hdže a kak na češćinje ASR čini?

- VUT Brno fakulta IT
 - džětaja z KALDI, sobuzatožerjo (Karel Veselý)
 - Phonexia
- ZČU, Plzeň
 - prof. Josef Psutka, dr. Pavel Král
 - džětaja z KALDI
- TU, Liberec
 - Petr Červa, Jiří Málek
 - maja swójski system
 - Newton technologies,a.s. Jindrich Slovak Praha, Na Pankraci
- IBM Prague Watson AI
 - Jan Cuřin, Jan Macek
 - maja swójski system
 - džětaja na mašinelnym spóznawanju wot 1962
- ČVUT, FEL, Praha
 - wužiwaja KALDI
- FIMU, Masarykova Universita, Brno
 - Pavel Rychly,
- LINGEA
- Omilia
- SpeechMatics
- Speech Tech – Plzen
- Google

Dvě technologické možnosti řečného spoznávání



Chcemy ASR činić

Mamy wjele datow (wjac hać 1000 hodź.)

Wužiwamy end-to-end systemy, neuralne syće na rzyzy daty, wuwija so w poslednich někotrych lětach

např. w Google za jendželšćinu a němćinu

Mamy mało datow (měnje hać 1000 hodź.)

Wužiwamy hybridne modele
1. Markoff předprocessing + GMM (gauzowe měšeńcy)
2. Neuralna syć

Wužiwamy za čěšćinu we wjetšinje systemow



Open-source tehnologija KALDI

- <https://kaldi-asr.org/>
- Powšitkowna informacija
 - Markoff předprocessing + GMM (gauzowe měšeńcy, gauzowky): 25 m/s partaje → MFCC wector
 - Derje za serbščinu, dokelž wužiwa tehnologije, kotrež trjebaja mało datow, min. 15 hodź. awdija
 - Aktiwny tool, podpěra so, da so jich woprašec (wotmołwjeja)
 - Wužiwa so we wjele systemach
- Móžne problemy
 - W KALDI kodach je druhdy ćežko, so orientować (njejsu xml, wjele blokow)
 - Njefunguje na dołhich rěčowych wotrězkach (pjeća, po hinašich žórłach funguje)
 - KALDI systemy su skerje za wědomostne dźěło tworjene hač za komercielne produkty



Što trjebamy za serbščinu

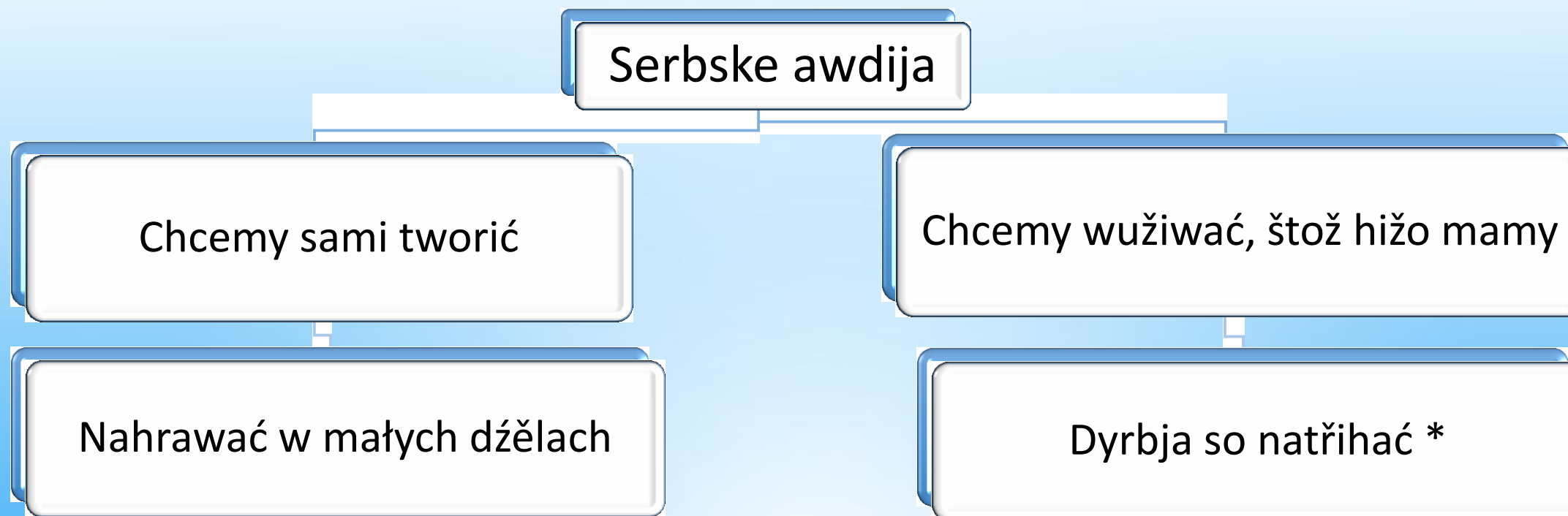
- Kajke daty su trěbne **Daty su najvažniše!**
 - awdija
 - transkripty
 - słownik
 - pisomne teksty
- Trenowanje syćow
 - wužiwanje eksistentnych modelow
 - wužiwanje kodow KALDI
 - wutwarić nowe?
- Administratiwne wěcy
 - software, engine
 - móžna kooperacija
- Diseminacija
 - z wulkimi koncernami sobu džětać
 - swójsku aplikaciju wutworić



Što trjebamy za serbščinu – awdijowe daty



Minimalnje 10-15 hodź. awdija. To je woprawdže minimum. Lěpje znajmjeńša 20 hodź., najlěpje hišće wjace. Ale z 10 hodź. móžemy započec modele trenować.



*Móže wužiwać program za čěšćinu, kotryž je hižo hotowy

Što trjebamy za serbščinu – awdijowe daty



Minimalnje 10-15 hodź. awdija. To je woprawdže minimum. Lěpje znajmjeńša 20 hodź., najlěpje hišće wjace. Ale z 10 hodź. móžemy započec modele trenować.



*Móže wužiwać program za čěšćinu, kotryž je hižo hotowy



Tworzenie awdijow

- Awdija dyrbja krótke być (5-15 sekundow)
- Awdija dyrbja wšelakore być
 - Wjele rěčacych wosobow, nic wjac hač 100 sadow wot jedneje wosoby
 - Fonetisce: wšitke fonemy dyrbja wjele króć reprezentowane być
 - Tematisce: nic jenož „Dobry dzeń“ wjele króć
- Jeli awdija cyle nowje nahrawaja, dobra strategija je na kwalitu prjedy myslić a snano samo sady wumyslić, kotrež so narěča
- Awdija dyrbja wot samsneje domeny pochadźeć kaž je planowana finalna aplikacija
 - Jeli chcemy přihotować rozpóznawanje za handyjowe prašenja, je dobre runje tajke dataje zběrać, abo znajmėńša chětro podobne
 - Nic jenož Serbski rozhlós, nic nahrawanje z telewizije



Tworzenie transkryptów

- Wszystkie słowa, które są nagrane, będziemy też transkribowane mieć, przepisane na fonemy.
- Fonemy w transkrypcji: Wulke tema
 - IPA
 - Husto so primarnje gramatiske fonemy wużiwaja a pak so tón system zjednori
- Transkripty dyrbja nic jenoż přepis podawać, ale tež specialne słowa mieć za např.
 - mjelčenje, pauzy
 - nadychowanje
 - smjeće
 - njewěstotu
 - pozadkowe zwuki



Słownik a daty za řečný model

- Słownik
 - wobsahuje dvě špalce
 - ortografiska forma
 - fonetiska transkripcja (např. IPA)
- Daty za řečný model
 - Za tvorjenje n-gramow
 - Dosahaja pisomne teksty

Ortografiska forma	IPA transkripcja
zetharny	'zɛvɐrni
dwě	'dɥɛ
wjesoła	'ɥɛsoβa
lubić so	lub'itʃ so

Wšitke pisane korpusy a daty, kotrež so za serbščinu namakaja, je na to dobre wužiwać



Trenink neuralnych syćow

- Neuralne syće so najprjedy na všech řečach trenuja a potom hakle so na cilowej řeči tunuja (tuned)
 - Tohodla je dobre eksistentne modele wužiwać
- Je móžne tež modele za čěšćinu a póľšćinu wužiwać a z toho model za serbšćinu tworić
- Čěšćina, póľšćina a němčina móžeja we přenje fazy pomhać, potom so móže póľšćina a němčina wužiwać
 - We fazy, hdyž je serbskich datow hišće přemało, k serbskim so pólsce abo němske dodać móžeja



Administrativne věce

- Software, engine
 - Dyrbi so natwarić
 - Dołhe lěta podpěrować
 - Móže być problem do komercielneho projekta open-source kody zakomponować
 - Update systemow (Linux, Windows, Apple) móže wulke změny přinjesć a wjele předžěłanja trjebać





Administrativne věcy

Spóznawanje serbšćiny

Chcemy cyle sami programěrować

Je (snano) lepje KALDI wužiwać

Chcemy z nazhonjenimi kolegami
sobu džětać

Móžna kooperacija z
Liberecom

- > Ústav Informačních technologií a elektroniky (ITE, Ing. Petr Červa, Ph.D)
- > Region NISA → je možnost wulke zhromadne projekty zapodać a wulka šanca jón dostać
- > Maja dobre modele za češćinu, pólsćinu, němćinu. Wjac hač 10 lět na tym džětaja



A hdyž to mamy, što z tym dale?

- Zakomponować do wulkich koncernow (Google Home, Apple Siri, Amazon Alexa, Microsoft Cortana)
 - Pjeća so za to zajmuja, jeli so jim při tym daty poskiča, samo hdyž je tych datow mało
 - Dyrbja na to swoje systemy pušćić a njemóžemy sej wěste być, zo to derje funguje
- Awdija spóznać a dale tekstowje pytać
- Samostatna aplikacija
 - Nazhonjenja z čěšćinu (minimalne):
 - **ALEX** - dialogue system framework (<https://ufal.mff.cuni.cz/alex>)
 - telefonuje a na prašenja wotmołwi
 - čěžko so podpěruje: update Linuxu → njefunguje
 - **Antelli** - Česká hlasová asistentka (<https://antelli.io/>)
 - instaluje so na handy
 - info wo wjedrje, telewizijnym programje, recepty atd.



Zakónčenje

- Pokrok cyle spěšnje doprědka běži
 - Za poslednich mało lět je so wjele změniło
 - Jeli dale w samsnym tempu džemy, móže so stać, zo bórže nimamy žane rěčne bariery
- Móžemy hinaše rěče wužiwać
- Daty su jara wažne!

Džakuju so!