Zpracování řeči a přirozeného jazyka

(Jan Černocký, Josef Psutka, Jan Nouza, Jan Hajič)

Řeč a psaný projev jsou základními lidskými komunikačními prostředky a jsou zkoumány od starověku. Počátky českého moderního lingvistického výzkumu se datují již do 20. let 20. století (Pražský lingvistický kroužek), v novodobé historii se o rychlý rozvoj oboru v Česku v souvislosti s AI zasloužili dva Čechoameričané - Prof. Frederick Jelinek (1932-2010, zakladatel statistického zpracování řeči a jazyka) a Prof. Hynek Heřmanský (nar. 1946, průkopník antropických a neurálních metod v automatickém zpracování).

České laboratoře se věnují dolování dat z řeči (převod řeči na text, rozpoznávání mluvčího, detekce jazyka), syntéze řeči z textu, zpracování řečových signálů, v oblasti zpracování přirozeného jazyka pak automatickému překladu, syntaktické a sémantické analýze, automatické sumarizaci, vyhledávání multimediálních informací a konverzačním (dialogovým) systémům. Využívají přitom techniky strojového učení jako statistické modely, neuronové sítě a nízkodimenzionální reprezentace dat (embeddings) - ve vybraných oblastech jsou na světové špičce a udávají celosvětové směry vývoje. Oba směry pracují s velkými daty pro trénování a testování systémů, množství práce a prostředků je tedy věnováno sběru a zpracování textových a řečových počítačových korpusů.

V oblasti zpracování řeči a přirozeného jazyka je Česká republika v mezinárodní vědecké komunitě považována za velmoc, v roce 2021 bude mj. hostovat největší konferenci v oboru řečových technologií - Interspeech - a v minulosti pořádala několik hlavních akcí v oboru zpracování jazyka (celosvětovou konferenci Association for Computational Linguistics a COLING) i řeči (mezinárodní konference IEEE ICASSP, ISCA Odyssey a IEEE ASRU).

Pracoviště (od západu na východ)	**Témata**
FAV ZČU v Plzni, skupiny „Strojové rozpoznávání, vnímání a porozumění“ a “Zpracování textových dat” Prof. Josef Psutka, Doc. Pavel Ircing, Doc. Jindřich Matoušek, Prof. Luděk Müller, Doc. Vlasta Radová, Doc. Pavel Král, Doc. Josef Steinberger	Rozpoznávání řeči, syntéza řeči, rozpoznávání a syntéza znakové řeči, hlasové dialogové systémy, porozumění řeči, vyhledávání informací v audiovizuálních archivech, inteligentní zpracování a porozumění vícejazyčným textovým dokumentům
MFF UK v Praze - Ústav formální a aplikované lingvistiky Prof. Jan Hajič, Doc. Ondřej Bojar, Doc. Zdeněk Žabokrtský, Doc. Markéta Lopatková, Doc. Pavel Pecina, Dr. Milan Straka, Dr. Daniel Zeman	Strojové hluboké učení, zpracování přirozeného jazyka a řeči, automatický překlad, aplikace pro komunikaci člověk stroj, internetové technologie a veřejnou správu

ČVUT CIIRC, Praha, skupina Conversational AI Jan Šedivý	Dialog management, automatické generování dialogů, question answering, automatické generování QA z FAQ, generování přirozeného jazyka, extrakce informací, znalostní databáze, sumarizace textu, chatboty
Laboratoř zpracování řečového signálu, FEL ČVUT Doc. Petr Pollák	Robustní rozpoznávání řeči, zvýrazňování řeči, databáze mluvené řeči
ITE TUL v Liberci - SpeechLab a ASAP Group Prof. Jan Nouza, Doc. Zbyněk Koldovský	Automatické rozpoznávání řeči v reálném čase s rozsáhlými slovníky a v různých jazycích, cloudová prostředí, analýza nezávislých komponent ve zpracování signálu, slepá separace audiosignálů
FI MU v Brně, Centrum zpracování přirozeného jazyka prof. Karel Pala, doc. Aleš Horák, doc. Pavel Rychlý	Reprezentace a odvozování znalostí, dolování znalostí, dialogové systémy, stylometrie, analýza autorství, komunikace člověk-stroj, tvorba a analýza velmi velkých textových korpusů pro desítky jazyků, slovníkové systémy.
FIT VUT v Brně, skupina BUT Speech@FIT Doc. Jan Černocký, Doc. Lukáš Burget, Dr. Pavel Matějka, Dr. Martin Karafiát	fonémové rozpoznávání, neurální přístupy k extrakci parametrů z řeči, multi-lingvální trénování DNN pro rozpoznávání řeči, i-vektorové a end-to-end systémy pro rozpoznávání jazyka a řečníka, neurální techniky pro jazykové modelování.

Výsledky / aplikace

- technologie rekurentních neuronových sítí pro jazykové modelování a extrakci nízkodimenzionálních vektorových popisů dat (embeddings), které jsou nyní používány všemi významnými světovými hráči (Google, Facebook, ...)
- řada open-source softwarových nástrojů, které získaly světový věhlas: MOSES, KALDI, NeuralMonkey, UDPipe.
- korektor pravopisu a gramatiky v Microsoft Word
- automatické titulkování pořadů České televize v reálném čase a automatický přepis a monitoring televizního a rozhlasového vysílání ve 20 evropských jazycích
- automatické diktování a přepis rozsudků na českých i zahraničních soudech
- automatický přepis a indexace zvukového archivu archívu Českého rozhlasu
- automatický překlad pro lokalizaci, používaný velkými IT firmami
- systémy pro analýzu řeči pro policii a zpravodajské služby
- syntéza řeči pro nákupní centra a pražské metro.
- rozpoznávání řeči a indexace pro digitalizované archivy orální historie (svědectví přeživších holocaustu MALACH, archiv Ústavu pro studium totalitních režimů).
- Dialogový systém Alquist (2 místo v soutěži Alexa Prize 2017 a 2018 v konkurenci více než 100 akademických týmů z celého světa

- řada korpusů a databází využívaných světovou výzkumnou komunitou, např. SpeechDat, Prague Dependency Treebank a nástrojů pro práci s nimi: korpusový manažer Manatee, Sketch Engine

Významné projekty (výběr z několika desítek evropských, amerických, národních a průmyslových projektů)

- NICOP – Adaptive Algorithms for Independent Component/Vector Extraction (Office of Naval Research Global) - adaptivní metody zpracování řečových signálů
- RATS - Robust Automatic Transcription of Speech (DARPA) - algoritmy dolování informací z řeči pro vojenské využití (přenosové kanály, kodeky, prostředí bojiště).
- LINDAT/CLARIAH-CZ - Výzkumná infrastruktura pro jazykové technologie (MŠMT)
- Automatická konverze a rekonstrukce hlasu se zaměřením na pacienty po totální laryngektomii (TAČR)
- CEMI - Centrum pro multi-modální interpretaci dat velkého rozsahu (Centrum excelence GAČR),
- NEUREM3 - projekt GAČR EXPRO pro modelování řeči a jazyka hlubokými sítěmi
- ELITR - European Live Translator, (EC H2020) - multijazykové automatické simultánní tlumočení a shrnutí mluvených projevů.
- ELG, European Language Grid (EC H2020) - vytvoření Language Technology Marketplace – mnohojazyčné, komerční a průmyslově řízené platformy pro jazykové technologie v EU.
- BISON - BIg Speech data analytics for cONtact centers (EC H2020) - analytika velkých řečových dat pro kontaktní centra.
- DeepSpot - Multilingual technology for spotting and instant alerting (TAČR)
- Automatické určení sémantické podobnosti kurzů na univerzitách (smluvní výzkum pro americkou firmu Owen software)

Vize rozvoje a příspěvku ekonomice

Zpracování řeči a přirozeného jazyka má mezi obory AI v České republice výsadní postavení z hlediska množství a kvality publikací, množství a prestiže mezinárodních projektů (včetně amerických agentur NSF, ONR, DARPA a IARPA) a průmyslových spoluprací či firem založených nebo úzce spolupracujících s výzkumnými skupinami (Phonexia, Lexical Computing, Lingea, SpeechTech, MemSource, Newton Technologies, Newton Media, Replaywell a další). Řada z nich je aktivní v zahraničí a realizuje většinu svého obratu mimo ČR. Obor již nyní ekonomice přispívá několika stovkami pracovních míst (samotné laboratoře a navázané firmy).

Budoucnost rozvoje v technologické oblasti leží v propojování modalit (řeč, text, video), vývoji algoritmů na nepřesně popsaných či zcela nepopsaných datech dostupných ve velkém množství na internetu, zlepšování robustnosti (např. při zpracování dat z nového typu telefonu či v novém dialektu) a ve zrychlení vývojového cyklu pomocí tzv. end-to-end trénování. V oblasti aplikací budeme cílit tradiční oblasti jako obranu/bezpečnost, média a státní správu, ale budeme cílit i nové domény jako sociální sítě, chytré domy či propojování řeči a textu s podporou business procesů. V oblasti projektové a organizační bude usilovat o udržení a zvýšení excelence v evropských, amerických a národních projektech, budeme se nadále věnovat práci s mezinárodní vědeckou komunitou (včetně pokračování již započaté internacionalizace našich týmů) a budeme pracovat na pevnějším propojení české “speech/NLP” komunity pomocí organizace spolupracující úzce s AICzechia. Budeme udržovat

a rozšiřovat aktivity v mezinárodních organizacích v oboru (předsednictví META-NET, výbory CLARIN ERIC, členství v LT Innovate, BDVA, ISCA, ACL, IEEE, ELRA a LDC)