Vedoucí týmu: Jan Hajič

Označení týmu: NLP - Ústav formální a aplikované lingvistiky MFF UK

1. **Obsah výzkumu** – state-of-the art

Ústav formální a aplikované lingvistiky na informatické sekci Matematicko-fyzikální fakulty Univerzity Karlovy (ÚFAL) se zabývá teoretickými i aplikovanými aspekty počítačové lingvistiky a zpracování přirozeného jazyka a řeči (dialogové systémy). ÚFAL je zároveň koordinátorem výzkumné infrastruktury pro jazykovědu a digitální humanitní vědy a umění, zahrnující 11 vedoucích institucí z ČR. Ústav dosahuje špičkových výsledků v oblasti uplatnění strojového učení ve zpracování přirozeného jazyka a v oblasti teoretické počítačové lingvistiky; v aplikační oblasti je na špičce zejména v oblasti strojového překladu, vyhledávání multimediálních informací a konverzačních (dialogových) systémů.

2. **Klíčoví výzkumníci**
   Prof. RNDr. Jan Hajič, Dr., Doc. RNDr. Ondřej Bojar, Ph.D., Doc. Ing. Zdeněk Žabokrtský, Ph.D., Doc. RNDr. Pavel Pecina, Ph.D., Doc. RNDr. Markéta Lopatková, Ph.D., Prof. PhDr. Eva Hajičová, DrSc., Mgr. Ondřej Dušek, Ph.D., RNDr. Barbora Vidová Hladká, Ph.D., RNDr. Milan Straka, Ph.D., RNDr. Daniel Zeman, Ph.D., Ing. Tomáš Mikolov, Ph.D. (externí spolupracovník)

3. **Klíčové metody a technologie**
   - Metody:
     - Strojové učení
     - Zpracování řeči a přirozeného jazyka
   - Technologie:
     - Komunikace člověk stroj
     - Internetové technologie
     - Veřejná správa

4. **Top 3 výsledky**

- Teoretický: teorie Funkčního generativního popisu a „Pražský závislostní korpus“ implementující syntakticko-sémantický popis jazyka, a to i ve formě vhodné pro strojové učení (kniha, Hajičová – Panevová – Sgall, Meaning of the Sentence in its Semantic and Pragmatic Aspects, 1. Vyd. 1986, vydáno Kluwer); k tomu

databáze anotovaných vět pro strojové učení podle FGP: Hajič et al. (2006), Prague Dependency Treebank, 2006, Lingusitic Data Consortium, Cat. No. LDC2006T01, Philadelphia, PA, USA).

- Aplikovaný: systém strojového učení pro strojový překlad, demo viz https://lindat.mff.cuni.cz/services/translation/ (patent pending, článek Popel et al. submitted 2019)
- Aplikovaný: nástroj pro analýzu textu (základní sadu, od textu po syntaktickou analýzu) v 70+ jazycích „UDPipe“, http://lindat.mff.cuni.cz/services/udpipe/, článek s popisem: Straka Milan, Hajič Jan, Straková Jana: UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: Proceedings of the 10th LREC 2016, ELRA, Paris, France, ISBN 978-2-9517408-9-1, pp. 4290-4297, 2016

5. Top 5 projektů

- ELITR, EC projekt H2020, 2019-2021, multijazykové automatické simultánní tlumočení a shrnutí mluvených projevů. Koordinuje ÚFAL (MFF UK), O. Bojar
- ELG, European Language Grid, EC projekt H2020, 2019-2021, vytvoření Language Technology Marketplace – mnohojazyčné, komerční a průmyslově řízené platformy pro jazykové technologie v EU. Koordinuje DFKI Berlin, v ČR je partner ÚFAL MFF UK (J. Hajič)
- QT21, EC projekt H2020, Quality Translation 21, 2015-2018, výrazné zvýšení kvality překladu přechodem na technologii Neuronových sítí a hlubokého učení ve strojovém překladu, koordinovalo DFKI Saarbruecken, v ČR byl partner MFF UK ÚFAL (J. Hajič, O. Bojar)
- LINDAT/CLARIN a navazující LINDAT/CLARIAH-CZ, velká výzkumná infrastruktura pro jazykovědu, humanitní vědy a umění, program VI MŠMT, 2010-2022
- Malach, NSF (USA) projekt pro vyhledávání v archivu přeživších Holokaustu Visual History Foundation (zakl. S. Spielberg), 2001-2007, ve spolupráci s USC Los Angeles, JHU Whiting School of Engineering, IBM T. J. Watson Research Center a University of Maryland.