MGR. JAN HAJIČ, JR. ***CURRICULUM VITAE***

CONTACT INFORMATION

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

Malostranské nám. 25
118 00 Prague, Czech Republic

E-mail: hajicj@ufal.mff.cuni.cz
WWW: http://ufal.mff.cuni.cz/jan-hajic-jr
https://github.com/hajicj/

EDUCATION

Charles University in Prague, Czech Republic
Faculty of Mathematics and Physics, School of Computer Science

Ph.D.: Computational Linguistics, 2014 – (ongoing)
- Thesis: Neural Network Models for Multimodal Data Interpretation
- Advisor: RNDr. Pavel Pecina, Ph.D.

M.Sc.: Computational Linguistics, 2011 – 2014 (September)
- Thesis: Matching Images to Texts
- Advisor: RNDr. Pavel Pecina, Ph.D.

Bc.: General Computer Science, September 2008 – 2011 (September)
- Thesis: Automatically measuring popularity of persons
- Advisor: RNDr. Ondřej Bojar, Ph.D.

INDUSTRY EXPERIENCE

Apple, Inc. Siri Speech team internship, 29. 6. — 16. 10. 2015. Cupertino, CA

RESEARCH INTERESTS

Optical music recognition

Machine Learning: Neural networks, Bayesian methods, Unsupervised learning

Natural Language Processing: Analysis of multimodal documents, Sentiment analysis, Topic modeling, Language modeling.

(former) Bioinformatics: RNA secondary structure prediction

PUBLICATIONS

Hajič jr., J. & Pecina, P. *The MUSCIMA++ Dataset for Handwritten Optical Music Recognition.* Proceedings of the International Conference on Document Analysis and Recognition 2017, Kyoto, 2017, pp. 39-46

Dorfer, M., Hajič jr., J., Widmer, G.: *On the Potential of Fully Convolutional Networks for Musical Symbol Detection.* In: Graphics Recognition Workshop 2017, Kyoto, 2017, pp. 53-54

Hajič jr., J. & Pecina, P. *Exploiting Music Notation Syntax for OMR.* In: Graphics Recognition Workshop 2017, Kyoto, 2017, pp. 55-56

Hajič jr., J. & Pecina, P. *Groundtruthing (not only) Music Notation with MUSCIMarker: a Practical Overview.* In: Graphics Recognition Workshop 2017, Kyoto, 2017, pp. 47-48

Hajič jr., J.; Notovný, J., Pecina, P., Pokorný, J. *Further Steps Towards a Standard Testbed for Optical Music Recognition.* Proceedings of the International Society for Music Information Retrieval Conference 2016, New York, 2016, pp.157-163

Hajič jr., J. & Pecina, P.: Žabokrtský, Z. (Ed.) *Matching Illustrative Images to "Soft News" Articles* .UFAL WDS 2015 (Conference of PhD Students in Mathematical Linguistics), Institute of Formal and Applied Linguistics, Charles University in Prague, 2015, pp.49-56

Straka, M.; Hajič, J.; Straková, J. & Hajič jr., J. *Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle.* 14th International Workshop on Treebanks and Linguistic Theories (TLT 2015), IPIPAN, 2015, pp.208-220

Josef Pánek, Jan Hajič jr., David Hoksza (2014): *Template-based Prediction of Ribosomal RNA Secondary Structure.* In: Proceedings of the IEEE conference for Bioinformatics and Biomedicine, Belfast, UK, November 2014.

Šindlerová, J.; Veselovská, K. & Hajič jr., J.: Abel, A.; Vettori, C. & Ralli, N. (Eds.) *Tracing Sentiments: Syntactic and Semantic Features in a Subjectivity Lexicon.* Proceedings of the XVI EURALEX International Congress: The User in Focus, EURAC research, 2014, pp. 405-414

Veselovská, K.; Hajič jr., J. & Šindlerová, J. *Subjectivity Lexicon for Czech: Implementation and Improvements.* Journal for Language Technology and Computational Linguistics, German Society for Computational Linguistics and Language Technology, 2014, 29, pp.47-61

Jan Hajič, jr., Kateřina Veselovská (2013): *Developing a Sentiment Annotator in UIMA – The Unstructured Information Management Architecture for Data Mining Applications.* In: Proceedings of the Workshop on Data Mining and Preference Learning on the Web, ITAT 2013, Donovaly, Slovakia.

Veselovská, K. & Hajič jr., J. *Why Words Alone Are Not Enough: Error Analysis of Lexicon-based Polarity Classifier for Czech.* Proceedings of the 6th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, 2013, pp. 1-5

Kateřina Veselovská, Jan Hajič jr., Jana Šindlerová (2012): *Creating Annotated Resources for Polarity Classification in Czech.* In: Proceedings of the 11th Conference on Natural Language Processing, Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI), Vienna, Austria, 2012. ISBN 3-85027-005-X

GRANTS

Grant Agency of the Charles University **(2017 - 2019)**

- Grant ID: GAUK 1444217
- Topic: Multimodal Optical Music Recognition using Deep Learning
- Role: Principal Investigator

Grant Agency of the Charles University **(2017 - 2018)**

- Grant ID: GAUK 170217
- Topic: Optical Music Recognition with Convolutional Neural Networks
- Role: Team Member

Grant Agency of the Czech Republic **(2012 - 2018)**

- Grant ID: P103/12/G084
- Topic: Center for Large-Scale Multimodal Data Interpretation
- Role: Team member, 2015-2017 (?)

Grant Agency of the Charles University **(2014 - 2015)**

- Grant ID: GAUK 550214
- Topic: rRNA Secondary Structure Prediction
- Role: Principal Investigator

Grant Agency of the Charles University **(2011 - 2013)**

- Grant ID: GAUK 353711
- Topic: Sentence-Level Polarity Detection in a Computer Corpus
- Role: Team Member

PROFESSIONAL EXPERIENCE

Member of the International Society for Music Information Retrieval.

MUSCIMarker: a GUI tool for annotating musical symbols (2016)

SAFIRE: a library for mostly Deep Learning experimentation (2014-2015)

rPredictor: an infrastructure for rRNA secondary structure prediction (2013-), team member, bioinformatics research leader, documentation coordinator

GeParse: generative parsing with Pitman-Yor processes based on the method of Wallach, 2008 (independent tool, 2013)

Consulting services on Sentiment Analysis for Yeseter Now (2013)

Implementing a Sentiment Analysis component for IBM Content Analytics (2012 – 2013)

GSEG: unsupervised morphological segmentation using sparse Dirichlet priors, based on the method of Lee, Haghigi and Barzilay, 2011 (independent tool, 2012)

Evaluating Machine Translation results for the EuroMatrix project (2009) Data preparation for the PCEDT corpus – translations (2008 – 2010)

SOFTWARE SKILLS

Programming languages: **Python,** bash, Perl (Past: C, C++; Elementary: Java, R)

Toolkits: PyTorch, Theano+derivatives, Scikit-Learn, gensim (contributed), biopython, Kivy (Past: UIMA, Treex; Elementary: Django)

Other: Git, LaTeX, SVN, Sphinx, ...

LANGUAGE SKILLS

Czech: native speaker

English: CPE Cambridge certificate (qualification at level C2 of CEFR)

German: Deutsches Sprachdiplom certificate (qualification at level B2-C1 of CEFR)

Latin: elementary

EXTRACURRICULAR INTERESTS

Music: piano (incl. admission to Prague Conservatory), composition (studied at the Janáček Academy of Performing Arts, Brno, Czech Republic for two years, 2011-2013), organ/ harpsichord, baroque music, improvisation. Leading an amateur early music ensemble.

Hiking, mountains, photography.

Past: Pralinka linguistics competition and Czech Linguistic Olympiad organizing team member

Class B driver's licence.