EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Zkušební okruhy pro státní doktorskou zkoušku

Modernizované zkušební okruhy pro státní doktorskou zkoušku doktorského studijního programu matematická lingvistika jsou v elektronické podobě k dispozici na webu

https://www.mff.cuni.cz/cs/studenti/doktorske-studium/studijni-plany/informatika/p4i3

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

Tato stránka vychází z podkladů pro tištěné studijní plány (tzv. Karolinku).

Studijní program P4I3 Matematická lingvistika

Oborová rada

Aktuální složení rady je na adrese http://mff.cuni.cz/phd/or/p4i3 .

Vypsaná témata

Jsou k nahlédnutí v SIS na adrese http://mff.cuni.cz/phd/temata/p4i3 .

Poskytovaná výuka

kód	Předmět	ZS	LS
NPFL004	Seminář z formální lingvistiky	0/2 Z	0/2 Z
NPFL006	Úvod do formální lingvistiky	2/0 Zk	—
NPFL015	Metody automatizovaného p řekladu	0/2 Z	—
NPFL024	Syntaktická analýza češtiny	—	0/2 Z
NPFL038	Základy rozpoznávání a generování mluvené řeči	2/2 Z+Zk	—
NPFL054	Úvod do strojového u ční v systému R	—	2/2 Z+Zk
NPFL063	Úvod do obecné lingvistiky	2/1 Z+Zk	—
NPFL067	Statistické metody zpracování p řirozených jazyk ů I	2/2 Z+Zk	—
NPFL068	Statistické metody zpracování p řirozených jazyk ů II	—	2/2 Z+Zk
NPFL070	Zdroje jazykových dat	1/2 KZ	—
NPFL073	Matematické metody v lingvistice	0/2 Z	—
NPFL075	Závislostní gramatiky a korpusy	—	2/2 Z+Zk
NPFL079	Algoritmy rozpoznávání mluvené řeči	—	2/2 Z+Zk
NPFL083	Lingvistické teorie a gramatické formalismy	—	2/2 Z+Zk
NPFL087	Statistický strojový p řeklad	—	2/2 Z+Zk
NPFL094	Morfologická a syntaktická analýza	2/0 KZ	—
NPFL095	Moderní metody v po čítačové lingvistice	0/2 Z	—
NPFL097	Neřízené strojové u ční v NLP	1/1 Z	—
NPFL099	Statistické dialogové systémy	2/1 Z+Zk	—
NPFL100	Variabilita jazyk ů v čase a prostoru	1/1 Z	—
NPFL103	Vyhledávání informací	2/2 Z+Zk	—
NPFL106	Obecná lingvistika	—	1/1 KZ
NPFL109	Číslicové zpracování zvukových signál ů	—	2/2 Z+Zk
NPFL114	Hluboké u ční	—	3/2 Z+Zk
NPFL116	Kompendium neuronového strojového p řekladu	—	0/2 Z
NPFL118	Zpracování p řirozeného jazyka na výpo četním clusteru	0/2 Z	—
NPFL120	Mnohojazy čné po čítačové zpracování jazyka	—	1/1 KZ
NPFL122	Hluboké zp ětnovazební u ční	2/2 Z+Zk	—
NPFL123	Dialogové systémy	—	2/2 Z+Zk
NPFL124	Zpracování p řirozeného jazyka	—	2/1 Z+Zk
NPFL125	Základy jazykových technologií	0/2 KZ	—

NPFL129	Úvod do strojového u ení v Pythonu	2/2 Z+Zk	—

Seznam po žadavků ke státní doktorské zkou šce

Zkouška se skládá ze dvou částí. V první části uchazeč představí výsledky své dosavadní rešeršní a výzkumné práce v návaznosti na téma zadané doktorské disertační práce. Ve druhé části jsou kladeny otázky ze tří okruhů. Okruh 1 je povinný. Ze zbývajících osmi okruhů uchazeč volí dva (všechny kombinace jsou možné).

Okruh 1 - Společný základ
Typy úloh a aplikací v počítačovém zpracování přirozeného jazyka. Základy pravděpodobnosti, jazykové modelování. Základní pojmy řízeného a neřízeného strojového učení. Výpočetní model neuronových sítí. Základy teorie grafů. Automaty a gramatiky, Chomského hierarchie. Popis jazykového systému jako souboru rovin. Základní pojmy lexikologie. Základní pojmy jazykové typologie. Korpusy, klasifikační kritéria. Principy lingvistické anotace. Morfologická anotace. Syntaktická anotace (složkové a závislostní korpusy). Paralelní korpusy. Specializované korpusy. Lexikální zdroje (slovníky, ontologie ad.). Zásady využívání dat při vyhodnocování experimentů, užití základních evaluačních měr, měření mezianotátorské shody. Vyhledávání v jazykových datech.

Volitelné okruhy pro komputa ční zam ěření

Okruh 2 - Základní statistické metody a strojové učení ve zpracování přirozeného jazyka
Pravděpodobnostní přístup ke zpracování jazyka. Jazykové modely, vyhlazování. Model šumového kanálu. Metody řízeného učení (lineární regrese, logistická regrese, rozhodovací stromy, perceptron, metoda podpůrných vektorů, K nejbližších sousedů ad.). Kernelové metody. Metody neřízeného učení (shluková analýza, EM ad.). Skryté Markovovy modely (algoritmy Baum–Welch, Forward–Backward, Viterbi). Algoritmy pro statistický tagging. Algoritmy pro složkový a závislostní statistický parsing. Statistický strojový překlad. Základy neuronových sítí pro využití v počítačovém zpracování jazyka. Testy signifikance.

Okruh 3 - Pokročilé strojové učení
Trénování neuronových sítí. Regularizace neuronových sítí. Konvoluční sítě. Rekurentní sítě. Distribuované reprezentace a embeddingy slov. Architektury pro zpracování přirozeného jazyka. Generativní modelování textu a obrázků. Zpětnovazební učení. Optimalizace diskrétních latentních proměnných. Bayesovská inference. Metody typu Markov Chain Monte Carlo.

Okruh 4 - Strojový překlad
Úloha strojového překladu (obtížnost MT, prostor správných a nesprávných překladů, víceznačnost a vágnost, mezivětné vztahy). Vyhodnocování překladu (ruční, automatické; proti referenci, bez reference). Data pro strojový překlad (zarovnání dokumentů, vět, slov a jiných jednotek, modely IBM). Klasický statistický strojový překlad (frázový překlad a další metody používající překladové jednotky). Heuristické přístupy (transfer-based MT, hybridní překlad). Neuronový strojový překlad (architektury, vztah diskrétní a spojité reprezentace výrazů přirozeného jazyka). Pokročilé metody (multi-task, multi-lingual MT, ad.). Formální popis jazyka pro překlad (morfologie a syntax v překladu). Počítačem podporovaný překlad (CAT, TM, inkrementální překlad).

Okruh 5 - Vyhledávání informací
Booleovský model. Invertovaný index, komprese indexu. Tolerantní vyhledávání. Oprava pravopisných chyb. Vektorový model. Evaluace a benchmarky. Metody zpětné vazby, rozšiřování dotazů. Pravděpodobnostní modely. Jazykové modely. Klasifikace dokumentů. Learning to rank. Shlukování dokumentů. Latentní sémantické indexování.

Okruh 6 - Zpracování mluvené řeči a dialogové systémy
Modelování akustiky fonému. Implementace Baum-Welch a Viterbi algoritmu pro rozpoznáváni řeči. Adaptační techniky. Metody syntézy řeči. Dialogové systémy. Základní komponenty dialogového systému. Stav dialogu, řízení dialogu. Porozumění mluvené řeči. Generování promluvy. Neuronové dialogové systémy. Hodnocení kvality dialogových systémů.

Volitelné okruhy pro lingvistické zam ěření

Okruh 7 - Formální popis jazykového systému
Základy fonetiky a fonologie. Morfologická stavba jazyka. Základní slovotvorné postupy. Syntaktická stavba jazyka: reprezentace větné stavby, povrchová a hloubková stavba věty, role valence, aktuální členění věty. Jazykový význam, asymetrie formy a významu; jazykový význam a kognitivní obsah. Výstavba textu: mezivětné významové vztahy (discourse relations), koreference a asociační anafora. Základní pojmy stylistiky; přehled stylů a žánrů, využití jazykových prostředků pro ztvárnění stylu textu. Základní pojmy sémantiky a pragmatiky.

Okruh 8 - Lingvistické formalismy (základní charakteristiky)
Lingvistické formalismy a jejich účel. Funkční generativní popis. Generativní gramatika, Government & Binding, minimalismus. Vztah gramatiky a slovníku (gramatické vs. lexikální jevy vs. jevy pomezní). Srovnání přístupů orientovaných lexikalisticky vs. gramaticky. Porovnání lingvistických formalismů z hlediska reprezentace syntaktické struktury. Zachycení slovesa jako syntaktického centra věty v různých formalismech. Sémantická reprezentace v různých jazykových modelech.

Okruh 9 - Variabilita jazyků a základy jazykové typologie
Variabilita jazyků a možnosti jejich klasifikace (genetická, areálová, strukturně-lingvistická). Genetická klasifikace jazyků, jazykové rodiny. Areálová klasifikace jazyků, jazykové svazy. Inventáře hlásek, distinktivní rysy a suprasegmentální jevy z kontrastivního pohledu; mezinárodní fonetická abeceda; tvoření slabik. Mluvená vs. psaná forma jazyka; typy písma. Morfologická stavba jazyků (jazykový typ flektivní, aglutinační, izolační a polysyntetický); typologie gramatických významů (pád, číslo, čas, aspekt, modalita aj.). Slovní druhy a jejich porovnatelnost přes hranice jazyků. Slovosled v kontrastivním pohledu; volný a pevný slovosled; dominantní slovosled; korelace slovosledných vzorců. Slovotvorné procesy napříč jazyky. Harmonizace anotačních schémat.

Doporu čená literatura

Okruh 1 - Spole čný základ

–Manning C. D., Schuetze, H.: Foundations of Statistical Natural Language Processing. *MIT Press, Cambridge, 1999.*
–Lüdeling, A., Kytö, M.: Corpus Linguistics: an International Handbook *, Volume 1. W. de Gruyter, 2008*
–Ide, N., Pustejovsky, J. (eds.): Handbook of Linguistic Annotation. *Springer, 2017.*

Okruh 2 - Základní statistické metody a strojové u čení ve zpracování p řirozeného jazyka

–Manning C. D., Schuetze, H.: Foundations of Statistical Natural Language Processing. *MIT Press, Cambridge, 1999.*
–Jurafsky, D. and J. H. Martin: Speech and Language Processing. *Prentice-Hall, 2nd edition. 2009.*
–Bishop, C.: Pattern Recognition and Machine Learning. *Springer, 2007.*

Okruh 3 - Pokro čilé strojové u čení

–Ian Goodfellow and Yoshua Bengio and Aaron Courville: Deep Learning. *MIT Press, 2016.*
–Richard, S. Sutton and Andrew G. Barto: Reinforcement Learning: An Introduction (Second Edition). *MIT Press, Cambridge, MA, 2018.*
–Murphy, K.: Machine Learning: a Probabilistic Perspective. *MIT Press, 2012.*

Okruh 4 - Strojový p řeklad

–Philipp Koehn: Statistical Machine Translation. *Cambridge University Press New York, 2010*
–Philip Williams, Rico Sennrich, Matt Post, Philipp Koehn: Syntax-based Statistical Machine Translation *. Morgan & Claypool Publishers, 2016.*
–Goldberg, Y.: Neural Network Methods for Natural Language Processing. *Morgan & Claypool Publishers, 2017*

Okruh 5 - Vyhledávání informací

–Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: Introduction to Information Retrieval *. Cambridge University Press, 2008.*
–Charu Aggarwal, Chengxiang Zhai: Mining Text Data *. Springer, 2012*
–David A. Grossman, Ophir Frieder: Information Retrieval, Algorithms and Heuristics *. Springer, 2004.*

Okruh 6 - Zpracování mluvené řeči a dialogové systémy

–Jurafsky, D. and J. H. Martin: Speech and Language Processing. *Prentice-Hall, 2nd edition. 2009.*
–Yu, D., Deng, L.: Automatic Speech Recognition: A Deep Learning Approach. *Signals and Communication Technology, Springer London, 2014.*
–Gao, J., Galley, M., Li, L.: Neural Approaches to Conversational AI. *Foundations and Trends in Information Retrieval, Vol. 13, No. 2-3, pp 127-298. 2019.*

Okruh 7 - Formální popis jazykového systému

–Booij, G.: Morphology. An International Handbook on Inflection and Word-Formation. *Volume 1, de Gruyter, 2000.*
–Ágel, V. et al. (eds.): Dependency and Valency. An international Handbook of Contemporary Research. *Volume 1. de Gruyter, 2003.*
–Cruse, D. A.: Meaning in language: an introduction to semantics and pragmatics. *Oxford: Oxford University Press, 2011.*

Okruh 8 - Lingvistické formalismy (základní charakteristiky)

– Allan, K. (ed.): The Oxford Handbook of the History of Linguistics. *Oxford Univesity Press. 2013*
– Ágel, V. et al. (eds.): Dependency and Valency. An international Handbook of Contemporary Research. *Volume 1. de Gruyter, 2003.*

Okruh 9 - Variabilita jazyk ů a základy jazykové typologie

–Haspelmath, M. et al. (eds.): Language typology and language universals. *De Gruyter, 2001.*
–Comrie, B.: Language universals and linguistic typology. *University of Chicago press, 1989.*