

Úvod do počítačové lingvistiky

Aleš Horák

E-mail: hales@fi.muni.cz
http://nlp.fi.muni.cz/poc_lingv/

Obsah:

- ▶ Organizace předmětu IB030
- ▶ Počítačová lingvistika
- ▶ Situace na FI MU

Organizace předmětu IB030

Hodnocení předmětu:

- ▶ **závěrečná písemka** (max 80 bodů)
 - jeden řádný a dva opravné termíny
- ▶ průběžný **úkol** (max 20 bodů)
- ▶ navíc možnost **1 bodu** za netriviální vylepšení slajdů
- ▶ **hodnocení** – součet bodů za písemku i úkol (max 100 bodů)
- ▶ rozdíly **zk, k, z** – různé limity

např.:

| | |
|---|----------|
| A | 80 – 100 |
| B | 73 – 79 |
| C | 65 – 72 |
| D | 58 – 64 |
| E | 50 – 57 |
| F | 0 – 49 |

| | |
|---|----------|
| K | 45 – 100 |
| Z | 40 – 100 |

Základní informace

- ▶ **cvičení** – občas doporučené malé úkoly
- ▶ jeden **hodnocený úkol** (viz další slajdy)
- ▶ **web** předmětu – http://nlp.fi.muni.cz/poc_lingv/
- ▶ **slajdy** – průběžně doplňovány na webu předmětu
- ▶ kontakt na přednášejícího – Aleš Horák <hales@fi.muni.cz>
(**Subject: IB030 ...**)

Samostatný hodnocený úkol – programátorský

- ▶ dva **typy** – *programátorský* × *lingvistický*
- ▶ **programátorský úkol** – **upravit** některou z dostupných jazykových knihoven pro **češtinu**:
 - **NLTK – Natural Language Toolkit** www.nltk.org
 - **FreeLing** nlp.lsi.upc.edu/freeling/
 - **Stanford University Natural Language Software** nlp.stanford.edu/software/
 - **Grammatical Framework** www.grammaticalframework.org
 - **AIML** www.aiml.foundation,
 - **ChatScript** github.com/ChatScript/ChatScript,
 - **ParlAI** github.com/facebookresearch/ParlAI
 - **Pepper** nlp.fi.muni.cz/projects/pepper
 - **Facebook research – fastText** github.com/facebookresearch/fastText
- ▶ k **odevzdání** je zapotřebí:
 - naprogramovaný odsouhlasený vybraný algoritmus na češtině (**návrh tématu** je součástí úkolu)
 - dokumentace programu s ukázkami a návodem na instalaci/spuštění na serveru aurora.fi.muni.cz a **vyhodnocením úspěšnosti** algoritmu na ne zcela triviálních českých datech
 - vše uložit v komprimovaném archivu do **odevzdávacího** do **termínu na webu předmětu**
- ▶ **hodnocení** bude od 0 do 20 bodů podle:
 - složitosti vybraného algoritmus
 - kvality zpracování algoritmu i dokumentace

Samostatný hodnocený úkol – lingvistický

- ▶ **lingvistický úkol** – tvorba **specializovaných jazykových dat** pro evaluaci automatických nástrojů

příklad z roku 2019: SQAAD – Simple Question Answering Database:

- čeština, 300 otázek a odpovědí podle textů z Wikipedie

Jak se nazývá strom, jehož zrna jsou využívána k výrobě čokolády?
Theobroma cacao

Čokoláda se vyrábí z kvašených, pražených a mletých zrněk tropického kakaového stromu *Theobroma cacao*.

<http://cs.wikipedia.org/wiki/%C4%8Cokol%C3%A1da>

aktuální zadání je popsáno na webu předmětu

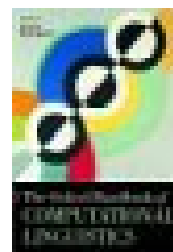
- ▶ k **odevzdání** je zapotřebí:
 - oznámit včas výběr úkolu (viz [www stránka předmětu](#))
 - odeslat výsledek v termínu dle instrukcí na webu
- ▶ **hodnocení** bude od 0 do 20 bodů podle:
 - výsledků kombinovaného hodnocení navržených sad

Literatura



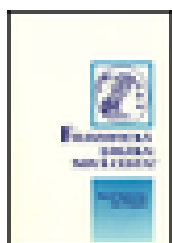
Jurafsky & Martin: **Speech and Language Processing**, 3rd edition draft, 2020. 615 s.

The Oxford handbook of computational linguistics, ed. by Ruslan Mitkov. Oxford University Press, 2003, 784 s.



Pala, Karel: **Počítačové zpracování přirozeného jazyka**, Brno FI MU, 2000. 190 s.

Chomsky, Noam: **Syntaktické struktury**, Praha : Academia, 1966. 209 s.



Materna, Pavel - Štěpán, Jan: **Filozofická logika: nová cesta?**, Olomouc (Univerzita Palackého), 2000. 127 s.

slajdy na webu předmětu

Náplň předmětu

- ▶ počítačové **zpracování přirozeného jazyka** (*Natural Language Processing, NLP*)
- ▶ roviny **analýzy jazyka**
- ▶ reprezentace morfologických a syntaktických **struktur**
- ▶ **analýza a syntéza**: morfologická, syntaktická, sémantická
- ▶ formy reprezentace **znalostí** o lexikálních jednotkách
- ▶ porozumění jazyku: **reprezentace významu** věty, inference a reprezentace znalostí

Co je “počítačová lingvistika”

Lingvistika:

- ▶ **jazykověda** (*lingua* = lat. *jazyk*)
- ▶ věda o **jazycích**, jejich třídění, stavbě, zvukové i psané podobě
- ▶ zkoumá **strukturu jazyka** – slovtvorba, kombinace slov do vět, význam věty, ...

Počítačová lingvistika:

- ▶ od 60. let, *Computational linguistics*, často **NLP** (*Natural Language Processing*)
- ▶ spojení **umělé inteligence** (informatiky) a **lingvistiky** – jako jedna z **kognitivních věd**
- ▶ zkoumá problémy **analýzy** či **generování** textů nebo mluveného slova, které vyžadují určitou (ne absolutní) míru porozumění přirozenému jazyku strojem.
- ▶ tvoří **jazykové modely** – pojmy **algoritmus**, **datová struktura**, **(formální) gramatika**, ...

Turingův test

- ▶ z roku 1950, založen na tzv. **imitační hře**
- ▶ úkol – **program komunikující jako člověk**
- ▶ zahrnuje:
 - zpracování přirozeného jazyka (NLP)
 - reprezentaci znalostí (KRepresentation)
 - vyvozování znalostí (KReasoning)
 - strojové učení
 - (počítačové vidění)
 - (robotiku)
- ▶ od 1991 – **Loebnerova cena** (*Loebner Prize*) → každý rok \$4.000 za “nejlidštější” program, nabízí \$100.000 a zlatá medaile za složení celého Turingova testu



Eliza, Parry, ...

- ▶ **Eliza** – Joseph Weizenbaum, MIT, 1964–66
- ▶ program napodobující chování psychoterapeuta:

young woman: *Men are all alike.*

eliza: *In what way?*

young woman: *They're always bugging us about something.*

eliza: *Can you think of a specific example?*

young woman: *Well, my boyfriend made me come here.*

eliza: *Your boyfriend made you come here?*

young woman: *He says I'm depressed much of the time.*

eliza: *I am sorry to hear you are depressed.*

young woman: *It's true. I'm unhappy.*

- ▶ **Parry** (Kenneth Colby, 70. léta), protějšek Elizy – počítačová simulace pacienta postiženého paranoií
- ▶ oba využívají spíš “**manipulace s řetězci**” než analýzu
- ▶ praktický význam – tzv. **expertní systémy, chatboti**

Chatbot – dialogový robot

- ▶ mnoho **proprietárních řešení** pro návrh dialogových robotů
- ▶ **praktické dialogy** i pouze **udržení zájmu**
- ▶ přístupy:
 - **pravidla** založená na **vzorech**
 - **Artificial Intelligence Markup Language, AIML**
 - robot **ALICE, Mitsuku** – vítězí v Loebnerově ceně
 - praktičtější použití, hodně závisí na podkladových pravidlech
 - **učení** z předchozích dialogů
 - potřebuje velké **množství dialogů** na učení
 - lépe se přizpůsobí **novým tématům** (což je plus i minus, viz *Microsoft Tay*)
 - robot **Cleverbot, Xiaoice**

Sociální roboti

- ▶ **rozšiřují** princip dialogových robotů o
 - **počítačové vidění**
 - **rozpoznávání emocí** ve tváři, v hlasu
 - **gesta**
 - **pohyb**
- ▶ důležitý je i **vzhled**
- ▶ umožňuje přidat do komunikace (programovou) **empatii**
- ▶ příklady robotů:
 - **hitchBOT** – sociální robot, který stopoval po USA v roce 2014,
 - **Matylda** – český stopující robot cestující v roce 2018 (OpenTechLab Jablonec nad Nisou Česká republika)
 - **Kismet** – robotí hlava (MIT AI Lab), která rozumí lidským emocím
 - **Bandit** – sociální robot určený autistickým dětem (USA, <https://robot.cfp.co.ir/en/newsdetail/368>)
 - robot **Pepper** od Softbank Robotics

Turingův test – jiné varianty

Winograd Schema Challenge:

- ▶ vyhlášený organizacemi [Commonsense Reasoning](#) a [Nuance](#) od 2015
- ▶ “strukturovanější” test – založený na [rozpoznávání anafor](#)
- ▶ podrobněji v přednášce o sémantice

Turing tests in Creative Arts:

- ▶ [DigiLit](#), [DigiKidLit](#) – generování povídek
- ▶ [PoetiX](#), [LimeriX](#), [LyriX](#) – generování sonetů, limeriků nebo básní
- ▶ Human-Computer Music Interaction – [AccompaniX](#), [AlgoRhythm](#) – generování doprovodné hudby pro duet s člověkem

IBM Watson – DeepQA

- ▶ stroj označovaný jako [Watson](#) – [DeepQA](#) vyvinutý za účelem porazit lidské šampiony ve hře [Jeopardy](#) ([Riskuj](#))
navazuje tím na stroj [DeepBlue](#), který v roce 1997 porazil Kasparova v šachu
- ▶ po 5 letech vývoje se to Watsonovi podařilo 16. února 2011
- ▶ princip:
 - vytvoření [databáze tvrzení](#) z internetových dat
 - analýza částí otázky, členění otázek podle [typu](#)
 - vysoce [paralelní hledání](#) odpovědi s určením [míry jistoty](#)
 - vyladěný algoritmus pro [kombinaci](#) stovek výsledků do výsledného rozhodovacího skóre
 - viz [Jak a proč Watson vyhrál Jeopardy!](#)
- ▶ [nejedná se o umělou inteligenci](#) podle Turingova testu
- ▶ praktický význam – [inteligentní](#) zpracování obrovského množství textů pro [hledání odpovědi](#)

Historie počítačové lingvistiky

- ▶ 1957 – rusko-anglický překlad
- ▶ Chomsky (60. léta) – generativní gramatika, vrozenost jazyka, ...
- ▶ strojový překlad není ani dnes dokonalý – potřebuje porozumět obsahu textu (Paretův zákon – pravidlo 80/20)
- ▶ problémy – víceznačnost, množství významů slov, různé způsoby užití slov k vyjádření významu, “Commonsense” a lidské uvažování
- ▶ Robert Wilensky: NLP je “AI-complete”
- ▶ 80. a 90. léta – rozvoj formalismů pro syntaktickou analýzu PJ (LFG, LTAG, HPSG)
- ▶ současně – zkoumání kvality statistických metod s rozsáhlými daty → srovnatelné výsledky!
- ▶ 90. léta až 200x – tvorba zdrojů vyšší úrovně (syntakticko-sémantické lexikony, wordnety, ...)
- ▶ 2013 až nyní – slovní vektory (word embeddings) a velmi velké neuronové jazykové modely
- ▶ stále není na obzoru splnění Turingova testu

Cíle počítačové lingvistiky

Významné úkoly v NLP:

- ▶ analýza přirozeného jazyka – morfologická, syntaktická, sémantická
- ▶ generování přirozeného jazyka
- ▶ syntéza a rozpoznávání řeči
- ▶ strojový překlad (*Machine translation*)
- ▶ odpovídání na otázky (*Question answering*)
- ▶ získávání informací (*Information retrieval*)
- ▶ korektura textu (*Spell-checking, Grammar checking*)
- ▶ extrakce informací (*Information extraction, Text Mining*)
- ▶ výtah z textu (*Text summarization*)
- ▶ určení typu dokumentu (*Text Classification/Clustering*)
- ▶ určení stylu dokumentu/autora (*Stylometry, Authorship Attribution*)

Přednášky se vztahem k NLP na FI MU

- ▶ program **Umělá inteligence**, specializace **zpracování přirozeného jazyka**
- ▶ vybrané Bc přednášky:

| | | |
|-------|---|------------------|
| IB030 | Úvod do počítačové lingvistiky | Horák |
| IB047 | Úvod do korpusové lingvistiky a počítačové lexikografie | Rychlý, Pala |
| IV029 | Logická analýza přirozeného jazyka | Materna, Duží |
| PB016 | Úvod do umělé inteligence | Horák |
| PB095 | Úvod do počítačového zpracování řeči | Bártek |
| PV277 | Programování sociálních robotů | Horák, Rambousek |
| PV056 | Strojové učení a dobývání znalostí | Popelínský |
| IA161 | Pokročilé techniky ZPJ | Horák et al. |
| PV173 | Seminář zpracování přirozeného jazyka | Horák, Rychlý |

NLP Centre – Centrum ZPJ na FI MU

- ▶ sdružení lidí (studentů Bc., Mgr. a PGS i zaměstnanců) z **oblasti NLP**
- ▶ webový server **nlp.fi.muni.cz**
- ▶ fyzicky – 2 “skleníky” ve 2. patře budovy B, místnosti **laboratoře zpracování přirozeného jazyka**
- ▶ vlastní laboratorní servery a stanice s OS Linux
- ▶ řeší několik velkých **grantových projektů**, pořádá **mezinárodní konference** (TSD, GWC, Lexicom, ...)
- ▶ práce studentů:
 - “malé projekty,” které se využijí v rámci “velkých projektů”
 - bakalářské, diplomové i disertační práce
 - někdy i zaměstnanecký poměr
- ▶ **PV173 Seminář Laboratoře zpracování přirozeného jazyka** – pravidelná společná výměna informací

NLP projekty a SW na FI MU

Vybrané projekty:

- ▶ **ajka**, **majka**, **desamb** – morfologický analyzátor, tagger
- ▶ **synt**, **set**, **zuzana** – syntaktické (a logický) analyzátoři
- ▶ **X.plain** – hra na hádání slov, člověk × počítač
- ▶ **Watsonson** – hra na hledání parafrází
- ▶ **DEB** – platforma pro XML databáze/slovníky
- ▶ **(DEB)VisDic** – editor wordnetů
- ▶ **VerbaLex** – slovník slovesných valencí
- ▶ **bonito**, **manatee**, **Word Sketches** – korpusový manažer
- ▶ **Visual Browser** – grafické znázornění (sémantických) sítí
- ▶ **GDW** (Grammar Development Workbench) – GUI pro vývoj gramatiky
- ▶ **demosthenes**, **text2phone (mbrola)** – syntetizátory řeči
- ▶ korpusy, slovníky, encyklopedie, ...