Ústav formální a aplikované lingvistiky

Bachelor's Thesis Topics Proposals

at UFAL

Wednesday 27 February 2019,
15:40, S3

ÚFAL overview

- choose your mix of:
  - Natural language processing (including Big Data)
  - Deep learning
  - Computational linguistics
  - Dialogue systems
- amazing concentration of experts, shared task winners (CoNLL, WMT,...)
- many international projects and industrial cooperation
- internship support (... universities)
- cluster: 100 GPUs (>1TB RAM), >2000 CPUs (>32TB RAM)
- try our web services at lindat.cz

Jak porozumět 75 jazykům?

- Lidský jazyk: spousta výjimek a nejednoznačností
- Jeho fungování lze „natrénovat“ z velkých dat

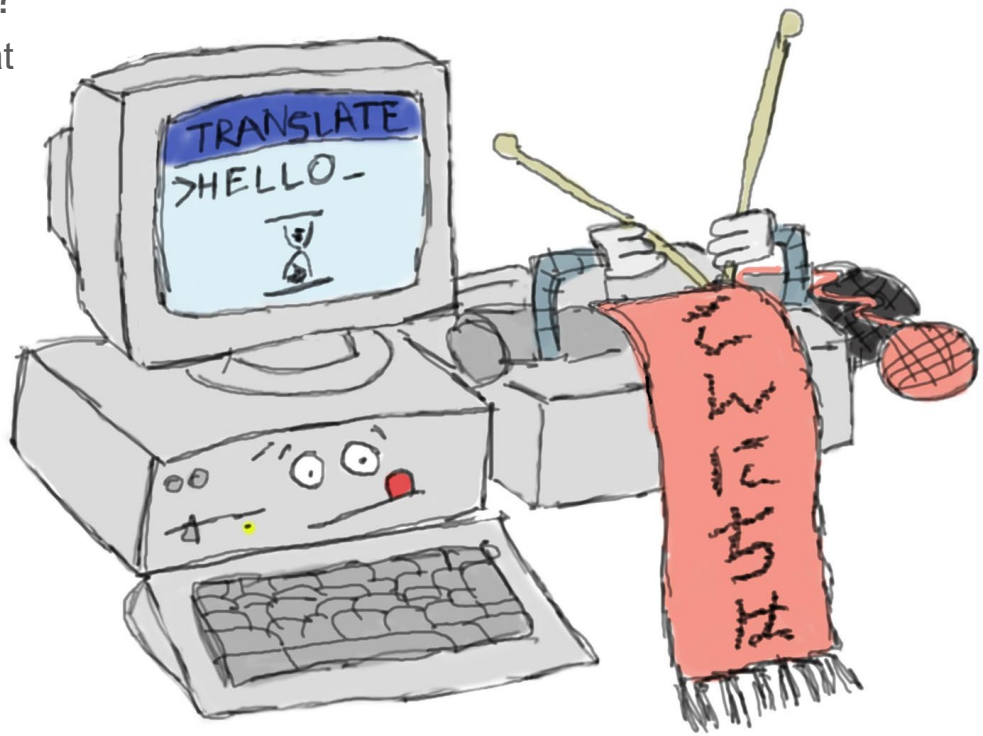
- Máme jazyková data pro 75 jazyků, od staré řečtiny až po vietnamštinu
- Práce s mnohojazyčnými daty = celá řada témat na bakalářky i diplomky!

- Věděli jste např., že muže být jeden minulý čas pro včerejšek a jiný pro vzdálenější minulost?

Prospective supervisor: Daniel Zeman

Strojový překlad (Neural Machine Translation, NMT)

- V anglicko-českém překladu jsme nejlepší na světě (ověřte si sami), ale jak pokrýt **mnoho** jazyků?
  Kolik ztratíme překladem přes Aj?
  - natrénovat přímé překlady a porovnat s překladem přes pivotní jazyk
- Překlad webových stránek
  - zachovat markup, přeložit obsah
- NMT pro reformátování
  - automaticky upravit např. bibliografické záznamy podle vzoru
- Dvojjazyčný Google Doc
  - textový editor pro dvousloupcové dokumenty (smlouvy ap.) s NMT



- Využití NMT pro generování poezie nebo překlad do ostravštiny/hantecu

Prospective supervisor: Martin Popel, Ondřej Bojar

Hraní si s jazykem

- generovat kartičky do hry iKnow (otázka, 3 nápovědy, řešení)

- automaticky vytvářet nebo řešit jazykové testy (FCE) nebo matematické slovní úlohy
- generovat textový popis ze strukturovaných údajů (např. předpověď počasí, popis trasy)
- generovat šifry či hádanky (např. synonyma, dvojznačná slova…)
- generovat diskuzní příspěvky “na iDnes” (rasistický, pro/protizemanovský…)
- spojení receptů (sloučit 3 podobné recepty na knedlík do jednoho)
- generovat jednoduchou křížovku (hřebenovka) včetně legendy
- převod textu minulost/budoucnost, vykání/tykání…
- …a další na http://ufal.cz/rudolf-rosa/projekty

Prospective supervisor: Rudolf Rosa

Vyhledávání nových slov v textech

- V nejrůznějších médiích se denně objevují nová slova.
- Chceme je mít všechny ve slovníku? Většinu ano.
- Dosud se to (někde) dělá ručně!
- Možnost spolupráce na tvorbě opravdových slovníků.

  - **přepodstatňovatel**
  - **shelfie** (ne, to není překlep!)
  - **haldoviště**
  - **trojchřtání**
  - **travolťácky**
  - **biozklamání**

- Víte, že v každých novinách je v průměru asi 4% slov, která neznají současné slovníky?

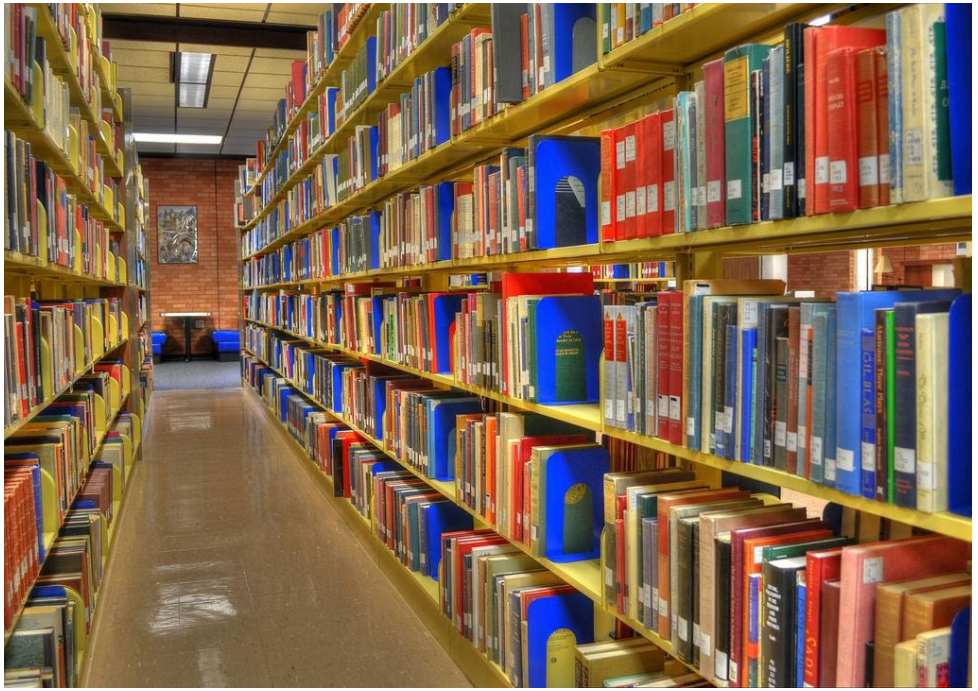
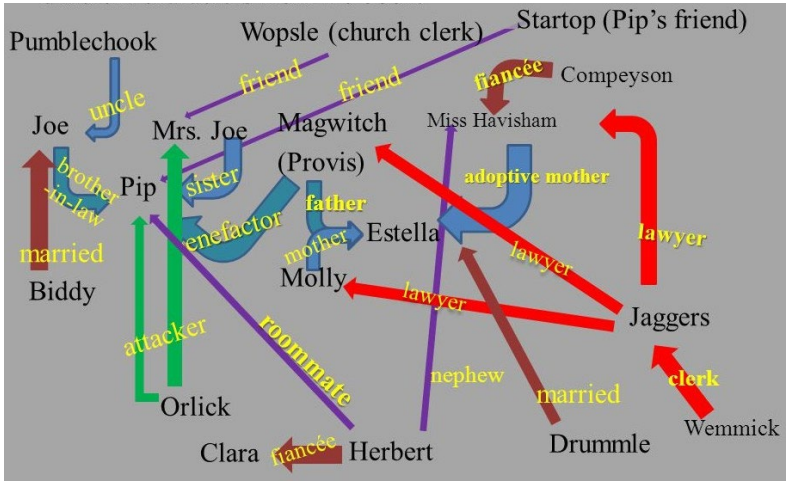


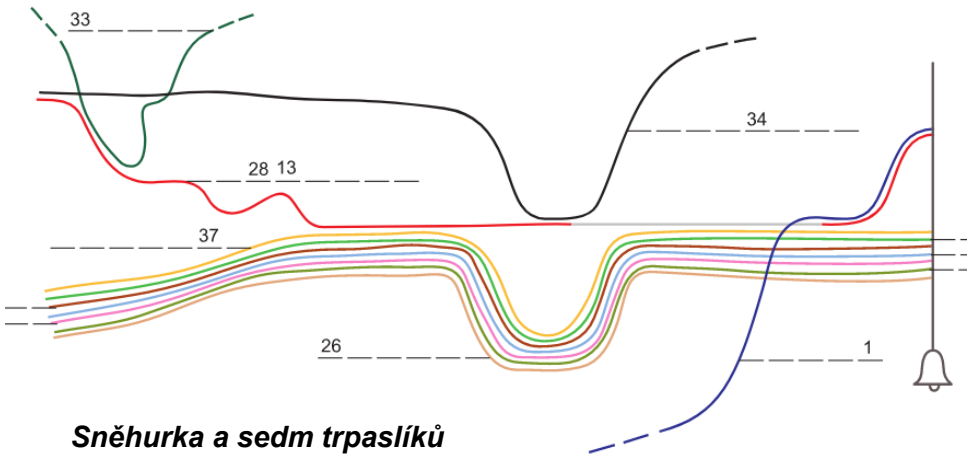
Photo: MSUB Library

Prospective supervisor: Jaroslava Hlaváčová

Extrakce informací z knihy

- vytvoření aplikace, která na základě vloženého textu/knihy vypíše základní informace o postavách a ději
- spousta možností od triviálních (nalezení hlavních postav) až po velmi složité a téměř nemožné (určení kdy a kde se která osoba nachází)
  - hlavní postavy
  - vztahy mezi postavami
  - časová osa: kde se o kom mluví, jak se jednotlivé postavy spolu setkávají
  - časová osa: kde se odehrává aktuální děj?



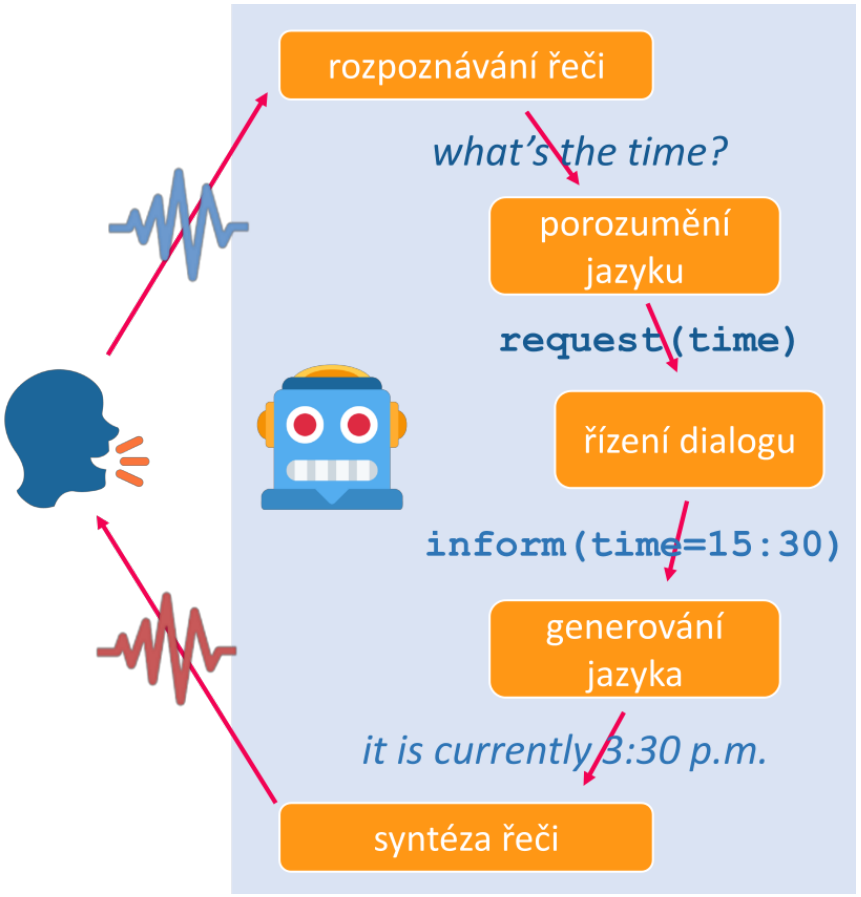


Sněhurka a sedm trpaslíků

Prospective supervisor: David Mareček <marecek@ufal.mff.cuni.cz>

Dialogové systémy – mluvte s počítačem

- Naučte počítač mluvit o konkrétním tématu:
  - počasí
  - městská doprava
  - televizní program
  - filmy
  - hudba
  - zprávy
  - …
- Naučte počítač zdvořilostní konverzaci
  - oblafněte Turingův test
- Nechte počítač, ať se to naučí sám
  - výběr vhodné odpovědi z velké databáze
  - strojové učení, neuronové sítě

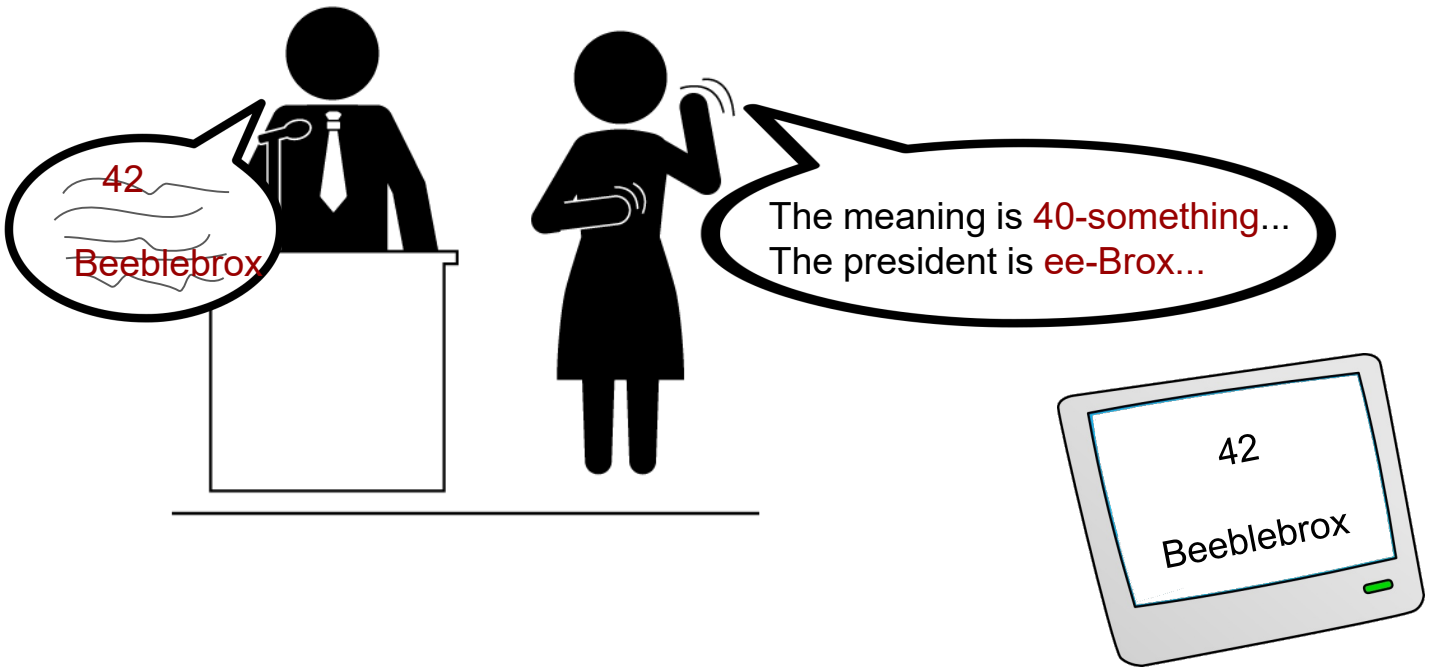


Více o dialogových systémech na předmětu http://ufal.cz/npfl123 !

Prospective supervisor: Ondřej Dušek <odusek@ufal.mff.cuni.cz>

Našeptávač pro tlumočníky

- Tlumočníci mají potíže, když někdo řekne mnoho čísel, vlastních jmen, ap.

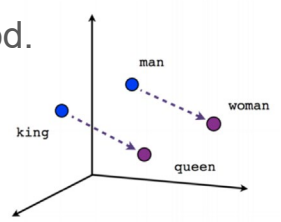


- Cíl: Vytvořit specializované rozpoznávání řeči na pojmenované entity, tlumočníkovi ukazovat nápovědu: konkrétní věci, které právě zazněly.
- Forma: Jednodušší: Server-based. Užitečnější: Off-line mobile app.
  - Může využít i podklady (slidy, materiály) dodané předem.

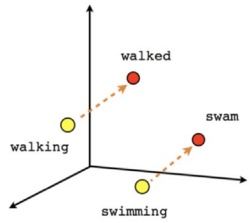
Prospective supervisor: Ondřej Bojar

Pestrost jazyka, modelování prostoru vět

- Jednu větu lze přeformulovat **tisíci způsoby** a zachovat význam.
  - Včera přišel Novák pozdě. František Novák dorazil jsem včera se zpožděním. F. Novák má další pozdní příchod. Franta zas přišel pozdě.
- Nepatrnou úpravou lze význam věty **zásadně změnit**.
  - Včera ***ne***přišel Novák pozdě. ***E***. Novák má další pozdní příchod.
- Neuronové sítě umějí větu převést do mnoharozměrného spojitého prostoru.
  - Pro slova je slavný word2vec ---->



Male-Female



Verb tense

- Dají se v tomto prostoru najít směry nebo transformace, které **odpovídají "normálním lingvistickým úpravám"**?
  - Nováka kritizovali. -(drsně)-> Nováka seřvali.
  - Nováka kritizovali. -(otočit význam)-> Nováka chválili. Nováka nekritizovali.
  - Nováka kritizovali. -(neurčitě)-> Někomu tam něco vytýkali.

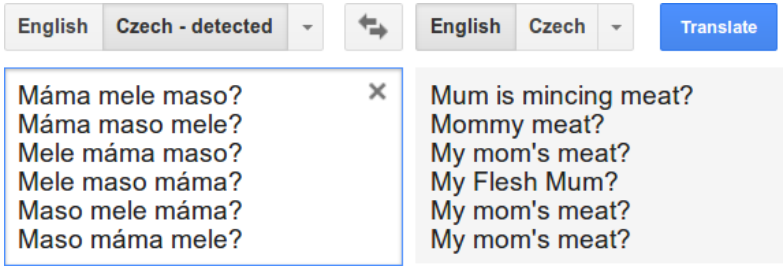
- První cíl: Za pomoci crowdsourcingu sestavit velký korpus úprav vět. (Bc)
- Druhý cíl: Promítnout sebrané věty do spojitého prostoru. (Mgr)
- Třetí cíl: Namapovat na sebe prostor úprav a spojitý prostor vět. (PhD ;-)

Prospective supervisor: Ondřej Bojar

Strojový překlad s porozuměním obsahu

- Lidé strojovému překladu věří někdy až moc:

- Překlad se opravdu podstatně zlepšil, ale o *porozumění* nelze mluvit:

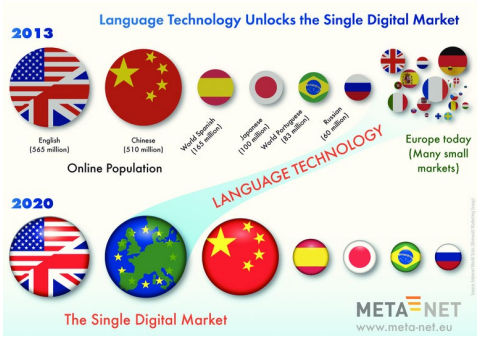


Cíl: Trénovat NMT technikou **multi-tasku**, aby zohlednilo **něco z významu**:

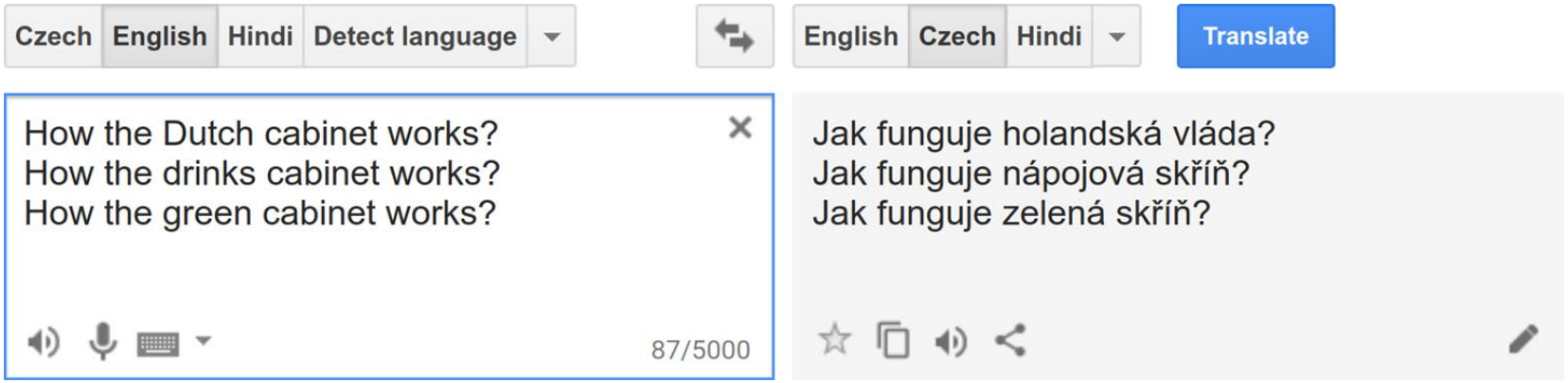
- Zachování sémantických rolí (kdo dělal komu co), pojmenované entity, …
- Odpovídání na otázky o obsahu věty...

Prospective supervisor: Ondřej Bojar

Překlad z více zdrojů; Multi-Source MT



- Cíl EU: Vše dostupné ve všech jazycích.
- Strojový překlad stále potřebuje ruční kontrolu.



- **Zjednoznačnění** provedené při revizi překladu do češtiny může pomoct strojovému překladu do dalších jazyků.

Cíl: Navrhnout architekturu NMT, která umí zpracovat **libovolnou podmnožinu jazykových mutací** dané věty.

Prospective supervisor: Ondřej Bojar and/or Jindřich Helcl

Master Thesis Topics Proposals

at UFAL

Wednesday 27 February 2019, 15:40, S3

Spoken Language Translation, Summarization

- **Neural machine translation** has made a big leap in translation quality.
  - This was thanks to end-to-end training, removal of independence assumptions.
- NN methods in **automatic speech recognition** (ASR).

Topic 1: Train ASR+MT end-to-end.
Challenge: Make use of (independent) MT and speech data. Network architecture.

- Online meetings (skype, hangouts, gotomeeting) are very popular.
- **Minutes** are needed to keep track of what has been agreed upon.
- Writing minutes makes people busy during the call and is tedious afterwards.

Topic 2: Design (NN-based) **speech summarization** to create compact minutes.
Automatically populate pre-defined meeting agenda.

Prospective supervisor: Ondřej Bojar

Visualization of deep learning components

A good visualization of DL-based models can lead to intuitions which in turn help us understanding what is going on under the hood of the “magic” black box.

Ideally, the student will:

1. develop a framework for visualizing different network components, and
2. use the framework to analyze a few models (e.g. NMT) on various inputs, trying to find patterns characterizing some (e.g. linguistic) features.

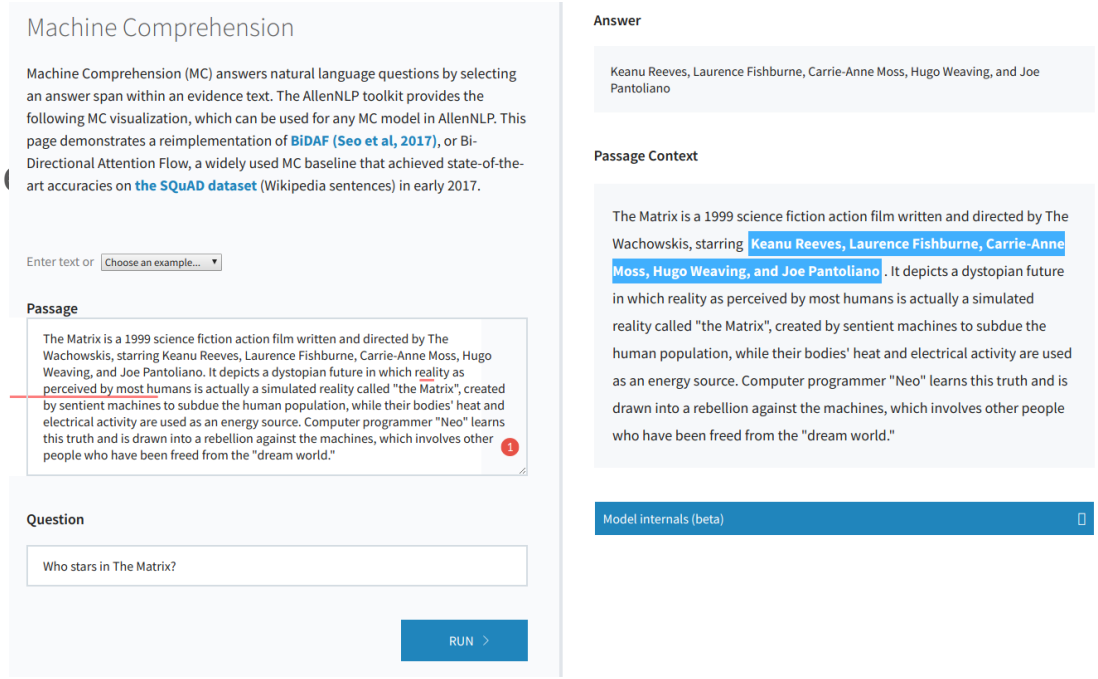
Prospective supervisor: Jindřich Helcl

Neural information search

Attention models work greatly for the searching for answers for natural language questions in coherent text. However, they require a lot of training data which is available in English only. The question is what should we do if we want the model to work in another language, e.g., Czech?

The thesis would include:

- Implementing and train a mod
  for answer span selection
- Exploring technique for
  transfer learning, multi-task
  learning



Prospective supervisor: Jindřich Libovický

Generating source code from description

- Task of generating source code in Python based on its description in natural language - questions mined from StackOverflow
- International challenge announced by CMU in Pittsburgh this summer, released a large dataset

https://conala-corpus.github.io

```
{
  "intent": "How do I check if all elements in a list are the same?",
  "snippet": "len(set(mylist)) == 1",
  "question_id": 22240602
}
```

Prospective supervisor: Jindřich Libovický

2 topics related to **Modelling Language Evolution**

- **1. Word migration across languages**
  - Words frequently borrowed from one language to another
  - Evolution of natural language vocabulary can hardly be captured by a human, because
    - zillions of morphemes in thousands of languages
    - all kinds of sound changes along the way, different alphabets…
  - No, we don't want just to reproduce the work of etymologists for individual languages. We want to find a ML model that fits the evolution of the natural language as a world-wide phenomenon.
- **2. Competition in Natural Languages**
  - Different languages are just different solutions to basically identical communication needs.
  - Like in living nature, possible solutions of particular "tasks" compete with each other, which results in a dynamic equilibrium. We want to model the balancing mechanisms mathematically.
  - Examples:
    - Typological view: fixed word order vs. richer morphology
    - Word-formation view: different suffixes competing with each other
- In both cases:
  - Timely topics: massively multilingual data available only in the last few years
  - Well-located too: rich experience with multi-lingual data at UFAL MFF UK
  - Inspiration could be taken from mathematical models in biology (predator-prey models, dynamic equilibria in microbio ecosystems, horizontal gene transfer...)
  - The last missing piece is YOU - a mathematically gifted machine learning practitioner
  - If interested, let's meet at **11:00 Wednesday May 23, room 409**

Prospective supervisor: Zdeněk Žabokrtský

Deep Understanding by Deep Learning

Despite applications like Machine Translation, true language understanding is still elusive. The focus of the thesis will be to analyze plain text in Czech and English to a knowledge representation graph by using supervised training with DNNs. Basic tools are available (up to natural language syntax), but semantic and knowledge extraction part is unsolved and will be the main problem to tackle. Datasets are available for at least two meaning/knowledge graph types (trees/DAGs).

Prospective supervisor: Jan Hajič

Wikification

Wikification is the process of annotating the mentions of concepts in a document with the URL of the Wikipedia page about that concept. Such a process can be used to improve annotation of existing Wikipedia articles (even cross-lingual) or for annotating any user text. Training material is large - the entire Wikipedia!

Foucaultovo kyvadlo

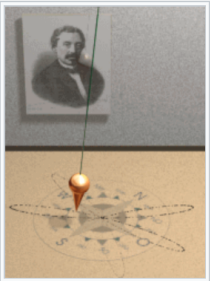
Tento článek je o fyzikálním experimentu. O knize Umberta Eca pojednává článek Foucaultovo kyvadlo (kniha).

Foucaultovo kyvadlo, pojmenované po francouzském fyzikovi J. B. Léonu Foucaultovi, představuje důležitý experiment potvrzující otáčení planety Země kolem své osy.

Obsah [zobrazit]

Historie experimentu [editovat | editovat zdroj]

Původní pokus byl proveden v roce 1851 v pařížském Pantheónu, kde bylo v kupoli zavěšeno závaží o hmotnosti 28 kilogramů na 68 metrů dlouhém laně. Doba kmitu kyvadla byla 16 sekund. Na závaží kyvadla byl hrot, kterým se do písku na podlaze zakresloval pohyb kyvadla. Pozorovatelé tak mohli vidět, jak se postupně mění rovina kyvu.



Animace pohybu Foucaultova kyvadla.

V roce 1851 bylo všeobecně známo, že se Země otáčí. Bylo také pozorováno zploštění Země na pólech. Foucaultovo kyvadlo však bylo prvním jasně viditelným důkazem a způsobilo tak velkou senzaci jak v odborných kruzích, tak u široké veřejnosti.

Na obou pólech Země se rovina kyvu nemění vzhledem k okolním hvězdám, zatímco Země se jednou za den zcela otočí. Vzhledem k Zemi se tedy rovina kyvu na severním nebo jižním pólu jednou za den zcela otočí po směru nebo proti směru hodinových ručiček. Pokud bychom Foucaultovo kyvadlo zavěsili na rovníku, rovina kyvu zůstane vzhledem k Zemi nezměněna. V ostatních zeměpisných šířkách se rovina kyvu vzhledem k Zemi sice otáčí, ale pomaleji, než na pólech.

Foucaultovo kyvadlo

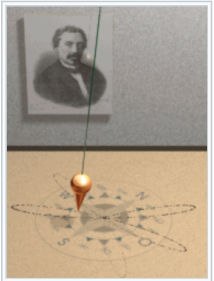
Tento článek je o fyzikálním experimentu. O knize Umberta Eca pojednává článek Foucaultovo kyvadlo (kniha).

Foucaultovo kyvadlo, pojmenované po francouzském fyzikovi J. B. Léonu Foucaultovi, představuje důležitý experiment potvrzující otáčení planety Země kolem své osy.

Obsah [zobrazit]

Historie experimentu [editovat | editovat zdroj]

Původní pokus byl proveden v roce 1851 v pařížském Pantheónu, kde bylo v kupoli zavěšeno závaží o hmotnosti 28 kilogramů na 68 metrů dlouhém laně. Doba kmitu kyvadla byla 16 sekund. Na závaží kyvadla byl hrot, kterým se do písku na podlaze zakresloval pohyb kyvadla. Pozorovatelé tak mohli vidět, jak se postupně mění rovina kyvu.



Animace pohybu Foucaultova kyvadla.

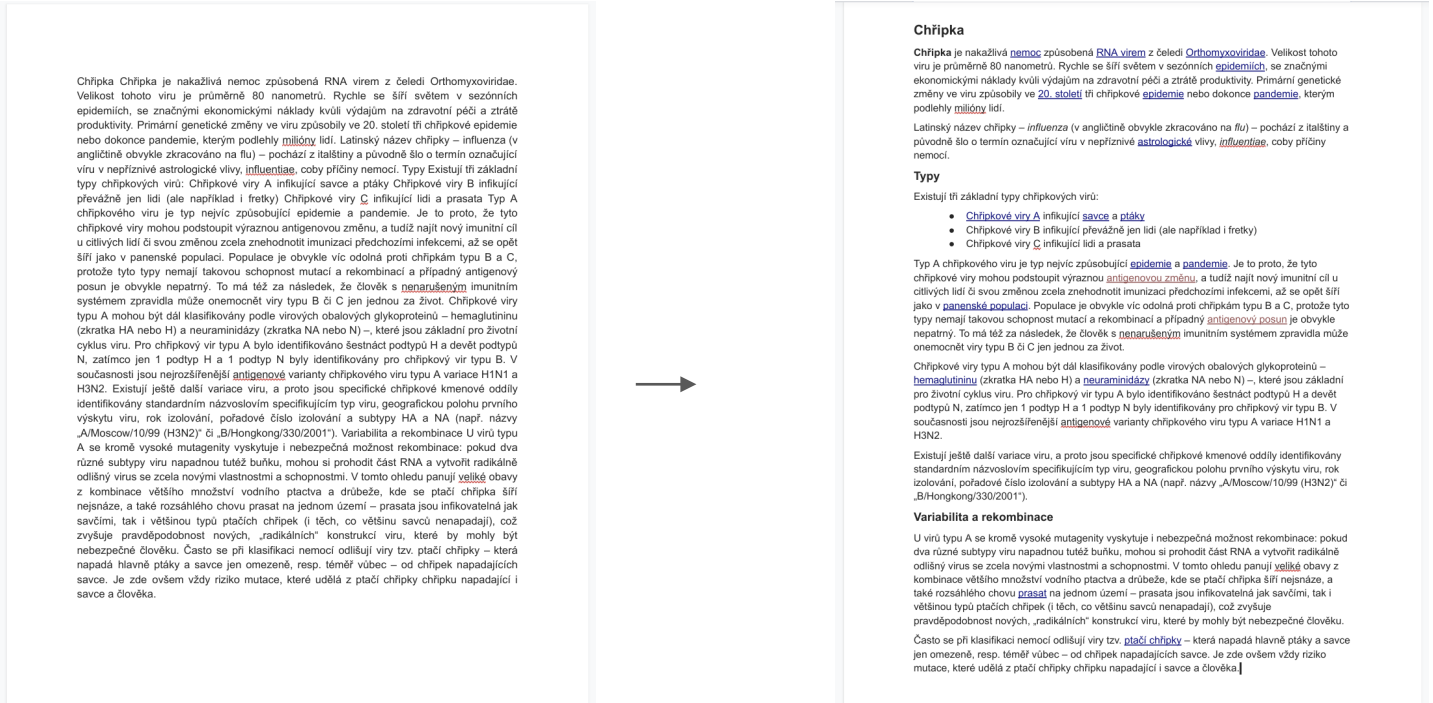
V roce 1851 bylo všeobecně známo, že se Země otáčí. Bylo také pozorováno zploštění Země na pólech. Foucaultovo kyvadlo však bylo prvním jasně viditelným důkazem a způsobilo tak velkou senzaci jak v odborných kruzích, tak u široké veřejnosti.

Na obou pólech Země se rovina kyvu nemění vzhledem k okolním hvězdám, zatímco Země se jednou za den zcela otočí. Vzhledem k Zemi se tedy rovina kyvu na severním nebo jižním pólu jednou za den zcela otočí po směru nebo proti směru hodinových ručiček. Pokud bychom Foucaultovo kyvadlo zavěsili na rovníku, rovina kyvu zůstane vzhledem k Zemi nezměněna. V ostatních zeměpisných šířkách se rovina kyvu vzhledem k Zemi sice otáčí, ale pomaleji, než na pólech.

Prospective supervisor: Pavel Pecina <pecina@ufal.mff.cuni.cz>

Text segmentation

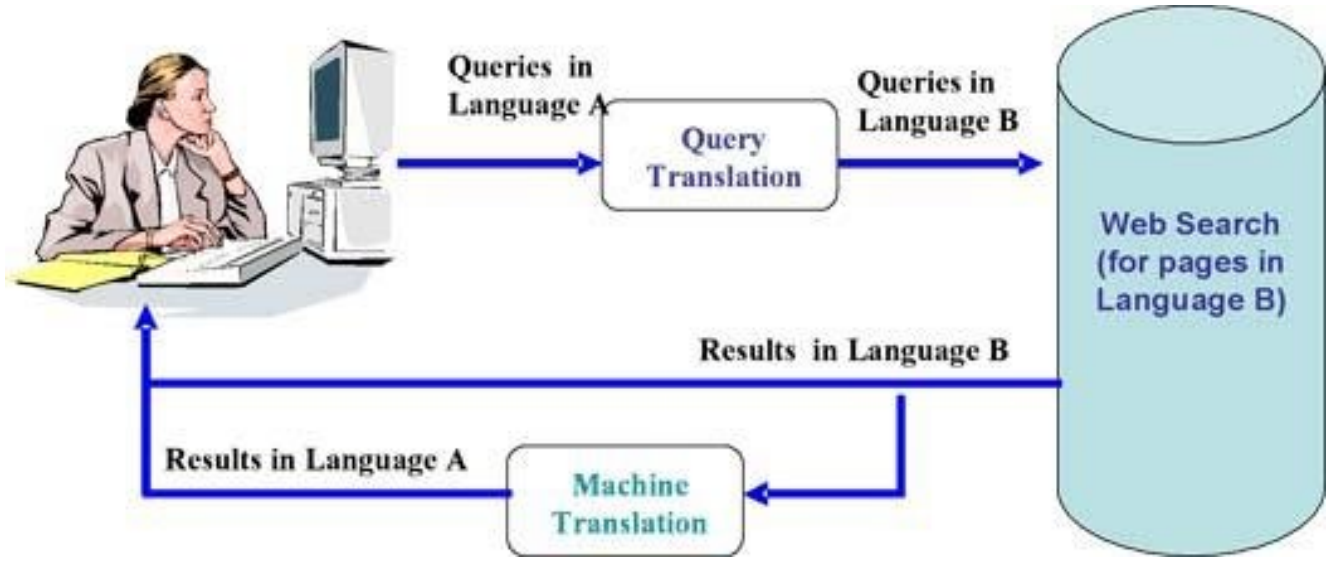
Text segmentation is the process of dividing written text into **meaningful units**, such as **words**, **sentences**, or **topics**. The term applies both to mental processes used by humans when reading text, and to artificial processes implemented in computers, which are the subject of natural language processing.



Prospective supervisor: Pavel Pecina <pecina@ufal.mff.cuni.cz>

Multilingual embeddings for cross-language search

Cross-language information retrieval (CLIR) is a subfield of information retrieval dealing with retrieving information written in a language (B) different from the language of the user's query (A).



The traditional approach to CLIR is translation of queries (in A) into the language of documents (B). We would like you to solve the same task by using multilingual (word embeddings).

Prospective supervisor: Pavel Pecina <pecina@ufal.mff.cuni.cz>

Neural information retrieval

Deep Learning has not affected the field of information retrieval (i.e. web search) much yet but there is a plenty of potential applications of neural technologies in information retrieval:

- Learning query/document representations (embeddings)
- Matching query/document representations
- Ranking search results
- Summary/snippet construction

Based on The NIST Text Retrieval Conference (TREC) Logo

Prospective supervisor: Pavel Pecina <pecina@ufal.mff.cuni.cz>