Počítačové zpracování přirozeného jazyka

Daniel Zeman
http://ufal.mff.cuni.cz/course/popj1/
zeman@ufal.mff.cuni.cz

Předpoklady

- Žádné (téměř)
- Pouze:
  - Základní znalost programování v některém programovacím jazyku
  - Středoškolské znalosti mluvnice (ne nutně české)
  - Pasivní znalost angličtiny výhodou

Zápočet

- Zápočet za domácí úkoly (Perl, Python, Java…; práce s cizími nástroji pro NLP)
- Dva možné modely:
  1. Jeden větší úkol pro jednotlivce na semestr. Krátká prezentace před ostatními v prosinci.
  2. Několik drobnějších úkolů v průběhu semestru. Součást „společného projektu“?

Přehled aplikací a problémů počítačové lingvistiky 1

- Vyhledat slovo v textu ve všech tvarech (morfologická analýza).

Vyhledání slova ve všech tvarech

- Lze nechat vyhledat jen část slova, ale:

- Internetové vyhledávače mají v indexu slova, ne jejich části.
  - Ovšem mnohé z nich už morfologii zvládají (Google).
- Povolíme-li hledání částí, dostaneme i texty, které nás nezajímají: hledáme *hodit*, zadáme *hod*, dostaneme nejen *hodit, hodím, hodil*, nejen *hod, hody, hodování*, nejen *přehodit, náhoda, přehodnotit*, ale dokonce i *chodit* nebo *schody*.
- Navíc nedostaneme *hoď, hoďme, hoďte* — to bychom museli hledat jenom *ho*, a to by bylo vůbec katastrofální. Zadat dotazy pro každý tvar zvlášť nejde, jsou jich desítky až stovky.
- Některá slova mění i kmen (stůl – stolu; brát – beru – bral).

Přehled aplikací a problémů počítačové lingvistiky 1

- Vyhledat slovo v textu ve všech tvarech (morfologická analýza).
- Kontrola pravopisu (spell checking). Inteligentní pro češtinu: např. kontrola tvrdého a měkkého *i* v příčestí minulém.
- Kontrola gramatiky (grammar checking) a stylu.

Kontrola pravopisu

- Prohledávání slovníku: jednoduché, téměř nelingvistické
- Problém je rozumně vybrat podobná slova (náhrada)
  - Jak měřit podobnost?
  - Odhadnout, co uživatel myslel?
  - Gramatika může na daném místě požadovat sloveso, význam ostatních slov zase může napovědět, které sloveso je nejpravděpodobnější
- V češtině málo účinné, mnoho koncovek OK, ale ne v daném kontextu, chyby v tvrdém a měkkém *i*. Viz též kontrolu gramatiky.
- Chybí přepínač stylu
  - Rozhodnu-li se pro koncovky typu *–ej*, je to jiná varianta jazyka a mělo by být vyžadováno držení jedné linie
- Zákaz dvou stejných slov za sebou: obecně ano, ale jsou výjimky, které by mohla rozpoznat kontrola gramatiky: *Nesnese se se sestrou, snědl jí tu tu buchtu a teď jí jí její koláč*.

Kontrola gramatiky a stylu

- Ideální stav: kompletní syntaktická analýza vztahů ve větě. To je ale těžké.
- Současná kontrola gramatiky v angličtině ve Wordu v sobě má i kontrolu stylu (bouří se proti dlouhým větám apod.)
- Univerzální úkoly: velké písmeno na začátku věty, zakázat dvě mezery za sebou apod.
- Úkoly v češtině: shoda podmětu s přísudkem (tvrdé a měkké *i*), shoda přívlastků s rozvíjenými podstatnými jmény, čárky kolem vnořených klauzí, interpunkce v uvozovkách.

Přehled aplikací a problémů počítačové lingvistiky 1

- Vyhledat slovo v textu ve všech tvarech (morfologická analýza).
- Kontrola pravopisu (spell checking). Inteligentní pro češtinu: např. kontrola tvrdého a měkkého *i* v příčestí minulém.
- Kontrola gramatiky (grammar checking) a stylu.
- Háčkování: vrátit diakritiku do textu, ze kterého byla odstraněna. Obdobný úkol: doplnit samohlásky do arabského nebo hebrejského textu.

Háčkování

- Háčkování a doplňování samohlásek, doplňování hranic slov (Asie)
- Pouhé hledání ve slovníku nestačí, jak ukazují následující příklady:

  horka cokolada (horká / hořká), mala (malá / mála), udelana (udělána / udělaná), uspi (uspi / uspí), mami (mami / mámí), zadejte uhel (zadejte úhel / žádejte uhel), cesky (česky / český / Češky / céšky), rad (rad / raď / rád / řad / řaď / řád)
- Stejné problémy mají v řadě dalších jazyků, někde dokonce občas znaménka vynechávají, aniž by je k tomu tlačila neschopnost techniky (francouzština, rumunština).

Problémy podobné háčkování

- Rozlišování malých a velkých písmen
  - Ve statistickém strojovém překladu bývá výhodné převést vstupní text na malá písmena.
  - Pak je ale nutné na výstupu odhadnout, kam patří velké písmeno.
- T9 v mobilech
  - Např. „852536“ může znamenat „tlakem“, „vlakem“, „vláken“, „vlčkem“.

Doplňování krátkých samohlásek v arabštině a hebrejštině

- Krátké samohlásky v semitských jazycích jsou podobný problém jako naše háčkování.
- V dialektech se navíc samohlásky (ale i souhlásky) liší, přestože zápis arabským písmem je jen ten jeden!

- جيب لي ثلاثة قحوة
- HWHhQ HTh'LTh YL BYJ
- jyb ly vlAvp qHwp (Buckwalterův přepis)
- Jīb lī thlīthah qaẖwah. (spisovná arabština)
- žib lí tléta dil qahwa (mar)
- žib lí thlétha qahwa (tun)
- gib lí taláta ahwa (egy)
- žib lí tléte ahwe (syr)
- džib lí theláthe gahwe (irq)
- přineste mi tři kávy (čsk)

Arabské samohlásky: nejednoznačnosti

- Nejednoznačnosti: nejen v dialektech, ale i v mluvnici.
- كتاب = kitáb (kniha)
- كتب = katab (psát), kutub (knihy)
- كتبت = katabt (napsal jsem, napsala jsem, napsal jsi), katabti (napsala jsi), katabit (napsala)
- Arabština umí zapisovat samohlásky pomocí diakritiky, ale Arabové to nedělají, vyskytuje se snad pouze v Koránu.
- Zapisují se dlouhé samohlásky (Á = ’, Í = Y, Ú = W), proto cizí slova mívají všechny samohlásky dlouhé, aby Arab poznal, jak se vyslovují (Československo = Tšíkúslúfákijá).
- Totéž platí pro hebrejštinu: v tóře bývají samohlásky zapsány, aby židé roztroušení po světě a hovořící jinými jazyky nezapomněli výslovnost, ale po Izraeli nic takového nenajdeme.

Segmentace

- Doplňování hranic slov (asijské jazyky, zejména čínština). I to je trochu analogie k doplňování samohlásek či diakritiky.
- Číňané nemají pojem slova, ale pro počítačové zpracování jazyka se tento pojem hodí. Znak (slabika) není vždy ideální ekvivalent.

– 这个多少钱?

- zhè ge duō shǎo qián ? (če ke tuo šao čchien?)
- tenhle kus mnoho málo peníze ?
- Zhège duōshǎo qián?
- Tohle kolik peněz? ... Kolik to stojí?

Přehled aplikací a problémů počítačové lingvistiky 1

- Vyhledat slovo v textu ve všech tvarech (morfologická analýza).
- Kontrola pravopisu (spell checking). Inteligentní pro češtinu: např. kontrola tvrdého a měkkého *i* v příčestí minulém.
- Kontrola gramatiky (grammar checking) a stylu.
- Háčkování: vrátit diakritiku do textu, ze kterého byla odstraněna. Obdobný úkol: doplnit samohlásky do arabského nebo hebrejského textu.
- Rozpoznávání naskenovaného písma (optical character recognition, OCR).
- Rozpoznávání řeči (speech recognition). Diktát se vrací ve smartphonech. Vyhledávání v nahrávkách (speech Google?)
- Strojový překlad z jednoho (přirozeného) jazyka do druhého.
- Generování textů (např. manuálů) v různých jazycích.

Přehled aplikací a problémů počítačové lingvistiky 2

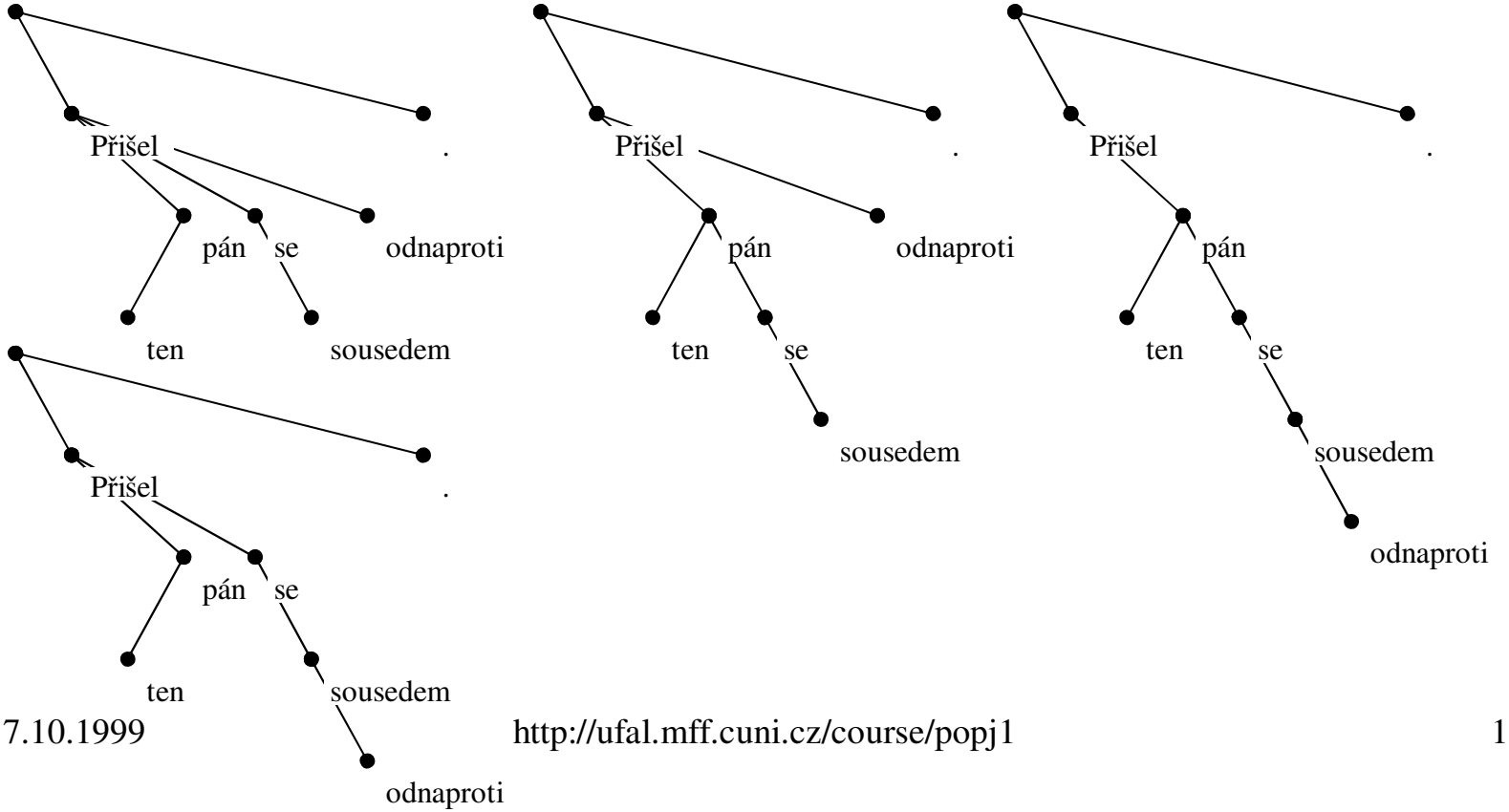
- Rozpoznat (převažující) jazyk, ve kterém je určitý dokument.
- Nalézt relevantní dokumenty v textové databázi (Google!). Nebo třeba: rozpoznat spam v mailu.
- Vytáhnout informace ze zpráv nebo článků s jistým tématem (např. všechny obchody s akciemi v daný den)
- Sumarizace textů (např. vytvořit třístránkové shrnutí tisícistránkového dokumentu).
- Dotazy na databázi v přirozeném jazyce (např. rezervace letenek).
- Automatická komunikace se zákazníkem po telefonu.
- Hlasové řízení stroje. Generování řeči strojem.
- Výukové systémy interagující se studentem, obecné systémy pro řešení konkrétních problémů.

Dílčí problémy

- Tokenizace (hranice slov a vět)
- Morfologická analýza (slovník + ohýbání)
- Zjednoznačňování, značkování (tagging), lemmatizace
- Rozlišení významu slov (word sense disambiguation)
- Zařazení slov do tříd podle použití (clustering)
- Synchronizace různojazyčných verzí téhož, párování vět a odst.
- Syntaktická analýza (parsing).
- Hledání základních jmenných frází (base noun phrase chunking)
- Zavěšování předložkových frází (prepositional phrase (PP) attachment)
- Slovesné rámce
- Hloubková analýza
- Základ a ohnisko (topic and focus), hloubkový slovosled.
- Doplnění chybějících členů
- Koreference
- Analýza diskurzu, anafora.

Předložkové skupiny

- *„Přišel ten pán se sousedem odnaproti.“*



Předložkové skupiny

- Anglický příklad:
  - *I saw the man with a telescope.*
    1. *Viděl jsem ho dalekohledem.*
    2. *Viděl jsem ho s dalekohledem.*

Předložkové skupiny

- *V období, kdy prudce* ***poklesl zájem*** *na domácím trhu, dokázala továrna část výroby exportovat.*

Předložkové skupiny a syntaktické nejednoznačnosti

- *V **letech** 1991 – 1993 jsem **absolvovala kurzy** řízení a marketingu na Collège Bart v kanadském Québecu.*

  - *absolvovala na Collège Bart*
  - *kurzy na Collège Bart*
  - *řízení a marketingu na Collège Bart*
  - *marketingu na Collège Bart*
  - *Collège Bart v Québecu*
  - *marketingu v Québecu...*

Předložkové skupiny a syntaktické nejednoznačnosti

- *„říjnové jednání OSN o klimatických změnách v Kodani“ (Události ČT, 27.2.2009)*

- Otázka: Došlo ke klimatickým změnám v Kodani?

Webové služby: Majka

- http://nlp.fi.muni.cz/projekty/wwwajka/WwwAjkaSkripty/morph.cgi?jazyk=0

- Český morfologický analyzátor Masarykovy univerzity v Brně

Hajičova morfologie

- https://lindat.mff.cuni.cz/services/morph/index.html
- Český morfologický analyzátor + tagger na MFF UK v Praze

Morfo: generátor tvarů

- http://quest.ms.mff.cuni.cz/cgi-bin/zeman/morfo/index.pl

- Opačné rozhraní k morfologickému slovníku: zadejte lemma (základní tvar slova) a nechte si vygenerovat všechny tvary.

Literatura

- James Allen: *Natural Language Understanding*. Benjamin/Cummings 1994, ISBN 0-8053-0334-0
- Adolf Erhart: *Základy jazykovědy*. Státní pedagogické nakladatelství; Praha, 1990
- Christopher D. Manning, Hinrich Schütze: *Foundations of Statistical Natural Language Processing*. The MIT Press 1999, ISBN 0-26213-360-1
- Mé prezentace a další odkazy na webu.