PA153 Počítačové zpracování přirozeného jazyka

11 – Znalosti, parafráze, odvozování

Karel Pala, Zuzana Nevěřilová

Centrum ZPJ, FI MU, Brno

4. prosince 2014

Znalosti a odvozování

- znalosti o jazyce (lexikon, gramatické kategorie, syntax)
- znalosti o světě

Znalostní báze (knowledge base, KB): obsahuje fakta, která jsou premisami v deduktivním odvozování
lidmi čitelné KB: how-to, FAQ, recepty, návody, diagramy
strojově čitelné KB: ontologie (SUMO-MILO), sémantické sítě (WordNet), dbPedia, ConceptNet

Reprezentace znalostí (knowledge representation): znalostní báze + odvozovací pravidla

Deklarativní vs. procedurální znalost

Deklarativní (formálně verifikovatelná, obecně platná) vs. procedurální (implicitní, méně obecná)
Příklad: robot, který se umí pohybovat po budově
procedurální znalost: ,,dojdi do místnosti“
deklarativní znalost: mapa objektu + základní kroky

Deduktivní odvozování: monotónní a nemonotónní odvozování [Allen, 1995]

KB: Ptáci létají. Vrabec je pták. Pštros je pták. Pštros nelétá.

Vrabec létá. Pštros létá. ~~Pštros létá.~~

Znalosti o světě

- encyklopedické (Jaké je hlavní město ČR?)
- common-sense (Jak je vhodné obléci se 4. prosince 2014?)

neostrá hranice
počítačově zpracovatelné zdroje encyklopedických znalostí:

- encyklopedie
- znalostní hry
- dbPedia: strojově zpracovaná Wikipedie

Common sense a odvozování

common sense: sdílená znalost, ne vždy v souladu s (vědeckými) fakty (V noci nesvítí slunce.)

Cheap apartments are rare.
Rare things are expensive.

Cheap apartments are expensive.

Deduktivní odvozování není možné použít vždy (ve skutečnosti skoro nikdy).

Common sense: nejznámější projekty

- CyC: vývoj od r. 1985(!), reprezentace pomocí vlastního jazyka CyCL, mikroteorie
- ConceptNet: syntaktická analýza OpenMind, propojení s Wiktionary
- Never-ending Language Learning (NELL): prochází web a odvozuje, občas nutný lidský zásah ("I deleted my Internet cookies", "I deleted my files" $\Rightarrow$ soubor je stejná kategorie jako pečivo)

Parafráze

Parafráze: promluva x je parafrází promluvy y, pokud x a y mají stejný nebo podobný význam.

Tento most postavila Nejlepší firma s.r.o.
Nejlepší firma s.r.o. postavila tento most.
Stavitelem tohoto mostu je Nejlepší firma s.r.o.

Přesnější definice

Textové vyplývání $\neq$ logické vyplývání

Z text t textově vyplývá hypotéza h ($t \Rightarrow h$), pokud lidé, kteří přečtou t, odvodí, že h je nejspíš pravda. [Dagan et al., 2007]
parafráze $= h \Rightarrow t \wedge t \Rightarrow h$

Rozpoznávání textových vyplývání/parafrází

hledání podobností:

- na řetězcích (např. Levenshteinova vzdálenost)
- na slovech
- na slovech s použitím znalostní báze (např. slovník synonym)
- na syntaktických stromech
- kombinace předchozích

Rozpoznávání textových vyplývání/parafrází

využití:

- odpovídání na otázky
- chatbots
- detekce plagiátů
- výuka
- automatická sumarizace textu
- doplnění implicitní znalosti
  - logická analýza textu
  - znalostní modely v umělé inteligenci
- ...

Korpusy parafrází

- Microsoft Research Paraphrase Corpus[1]
- The Boeing-Princeton-ISI (BPI) Textual Entailment Test Suite[2]
- Multiple Translation Chinese Corpus[3]
- The SEMILAR Corpus: The SEMantic SimILARity Corpus[4]
- Paraphrase Discovery[5]

[1] http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/

[2] http://www.cs.utexas.edu/users/pclark/bpi-test-suite/

[3] https://catalog.ldc.upenn.edu/LDC2002T01

[4] http://deeptutor2.memphis.edu/Semilar-Web/public/semilar-api.html

[5] http://nlp.cs.nyu.edu/paraphrase/

Paraphrase Discovery

vztahy mezi pojmenovanými entitami v korpusových datech:

```
[lemma="Hannibal"] []* [lemma="Hopkins"] within <s/>
```

ztvárnit	jako
hrát	odmítnout
s	na roli
si	hrající
/	se objevil
v podání	představoval
alias	působí v roli
se svým přítelem	
(	
po boku	

Generování parafrází

Základní způsoby parafrázování:

- aktivní–pasivní větná konstrukce: Tento most byl postaven Nejlepší firmou s.r.o.
- synonyma: Tuto lávku postavila Nejlepší firma s.r.o.
- hyperonyma: Tuto stavbu postavila Nejlepší firma s.r.o.
- substantivizace, deverbalizace: Stavitelem tohoto mostu je Nejlepší firma s.r.o.
- kombinace: Tento most byl vytvořen Nejlepší firmou s.r.o.

Podrobněji v [Bhagat and Hovy, 2013].

Přirozená logika [Lakoff, 1970]

nástrojem této logiky je přirozený jazyk

- monotonicita (monotonicity): víc než tisíc je hodně
  Mám víc než tisíc knih. Mám hodně knih.
  Nemám víc než tisíc knih. Nemám hodně knih.
- obsažení/omezení (containment): červené auto je auto
  Po ulici jelo červené auto. Po ulici jelo auto.
  Po ulici nejelo červené auto. Po ulici nejelo auto.
- exkluze (exclusion): pes není kočka
  Na dvorku seděl pes. Na dvorku seděla kočka.
  Na dvorku neseděl pes. Na dvorku neseděla kočka.

odvození vs. presupozice:
Mark David Chapman zastřelil Johna Lennona. $\Rightarrow$ John Lennon nežije.
Mark David Chapman nezastřelil Johna Lennona. $\not\Rightarrow$ John Lennon nežije.
Brazílie vyhrála mistrovství světa.$\Rightarrow$ Brazílie hrála na mistrovství světa.
Brazílie nevyhrála mistrovství světa.$\Rightarrow$ Brazílie hrála na mistrovství světa.

Přirozená logika [Lakoff, 1970]

nástrojem této logiky je přirozený jazyk

- monotonicita (monotonicity): víc než tisíc je hodně
  Mám víc než tisíc knih. Mám hodně knih.
  Nemám víc než tisíc knih. Nemám hodně knih.
- obsažení/omezení (containment): červené auto je auto
  Po ulici jelo červené auto. Po ulici jelo auto.
  Po ulici nejelo červené auto. Po ulici nejelo auto.
- exkluze (exclusion): pes není kočka
  Na dvorku seděl pes. Na dvorku seděla kočka.
  Na dvorku neseděl pes. Na dvorku neseděla kočka.

odvození vs. presupozice:
Mark David Chapman zastřelil Johna Lennona. ⇒ John Lennon nežije.
Mark David Chapman nezastřelil Johna Lennona. ⇏ John Lennon nežije.
Brazílie vyhrála mistrovství světa.⇒ Brazílie hrála na mistrovství světa.
Brazílie nevyhrála mistrovství světa.⇒ Brazílie hrála na mistrovství světa.

V přednášce jsem se spletla, šipky na snímcích jsou dobře.

Analýza textu „bez analýzy“

Z textu můžeme získat dost informací bez analýzy obsahu textu (kódování nebo jazyk, délka textu, počet odstavců, počet slov . . .).

Můžeme získat informace o obsahu bez analýzy obsahu?

Ano, ale . . .

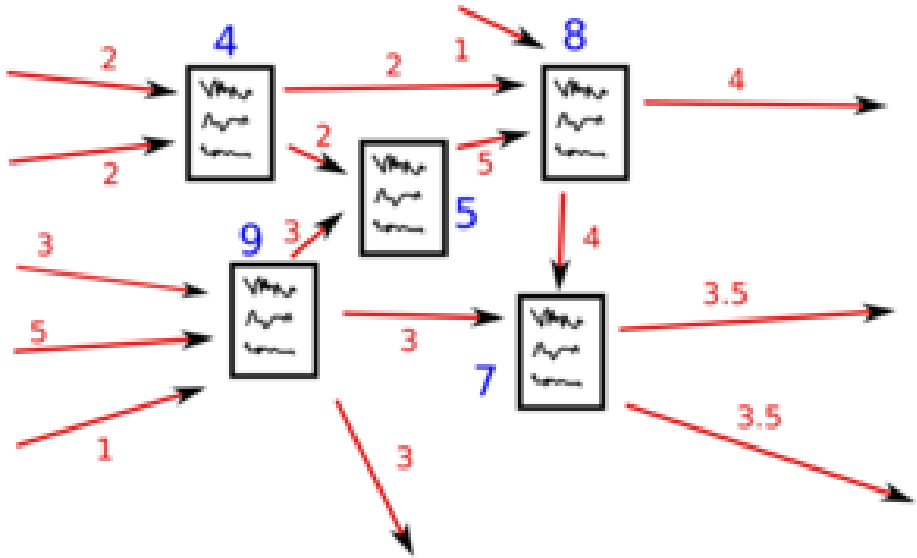
Analýza textu ,,bez analýzy“: jak?

jazykově nezávislé metody jsou založeny na faktu, že

- některé části textu jsou důležitější než jiné
- pokud ty důležitější identifikujeme, můžeme dále pracovat jen s nimi

Odbočka k PageRank: důležité jsou odkazy [6]

$$R(a) = \sum_{u \in B_a} \frac{R(u)}{N_u}$$



[6] http://cs.wikipedia.org/wiki/Soubor:Pagerank1.png

Rozpoznávání témat (topic recognition)

Čistý zisk energetické společnosti ČEZ za tři čtvrtletí letošního roku meziročně klesl o 4,7 procenta na 31,7 miliardy korun. Tržby se meziročně snížily o 0,3 procenta na 161,9 miliardy korun. Hlavním důvodem poklesu byly odpisy aktiv kvůli regulacím evropského energetického sektoru a související snižování velkoobchodních cen elektřiny , sdělila firma . Výsledek je tak výrazně pod očekáváním. Analytici totiž předpokládali, že čistý zisk ČEZ stoupne o víc než čtyři procenta na 34,8 miliardy korun. Společnost také oznámila, že kvůli snížení velkoobchodních cen elektřiny a regulatorním zásahům do evropského energetického sektoru snížila celoroční výhled čistého zisku na 35 miliard korun . Původně počítala s výsledkem o 2,5 miliardy vyšším. " Očekávané celoroční výsledky hospodaření ČEZ odrážejí současný stav energetiky v Evropě . Fakt, že na naše výsledky tato krize doléhá později a výrazně méně než na naše evropské konkurenty, reflektuje zejména naši úspěšnou strategii předprodejů elektřiny na roky dopředu a důraz na vnitřní úspory," uvedl k výsledkům předseda představenstva a generální ředitel Daniel Beneš .

Rozpoznávání témat (topic recognition)

- extrakce klíčových frází (key phrases)
- klasifikace textu do kategorií (sport, fotbal, finance, půjčky, ekonomie, energetika. . .)

Extrakce klíčových frází (key phrases) obecně

- podobný úkol jako extrakce klíčových slov
- klíčové n-gramy (slovo = unigram)
- zkoumaný korpus a referenční korpus
- potřebujeme (předpočítané) frekvence n-gramů
- frekvence n-gramu není srovnatelná s frekvencí m-gramu pro $n \neq m$

Extrakce klíčových frází (key phrases), projekt To|P|icks

- zkoumaný korpus je (krátký) text
- referenční korpus je (velký) korpus
- text rozdělíme na možné fráze (pomocí regulární gramatiky)
- každá fráze získá skóre: frekvence n-gramů v textu / frekvence n-gramů v korpusu
- vyhledáváme základní tvary n-gramů (např. energetický společnost ČEZ)
- skóre fráze posiluje, pokud má podfráze také nějaké skóre
- skóre fráze posiluje, pokud fráze obsahuje pojmenovanou entitu
- skóre fráze oslabuje, pokud je fráze krátká nebo pokud je číslo

Projekt To|P|icks: analýza „bez analýzy“

- pracujeme s tokeny (použili jsme tokenizaci)
- pracujeme s n-gramy lemmat (použili jsme lemmatizaci)
- počítáme poměr frekvencí (používáme korpus konkrétního jazyka)
- extrahujeme kandidáty pomocí regulární gramatiky (používáme parciální syntaktickou analýzu)
- rozpoznáváme pojmenované entity

- neprobíhá úplná analýza
- nepracujeme s lexikálním významem

Projekt To|P|icks: hodnocení

Čistý zisk energetické společnosti ČEZ za tři čtvrtletí letošního roku meziročně klesl o 4,7 procenta na 31,7 miliardy korun. Tržby se meziročně snížily o 0,3 procenta na 161,9 miliardy korun. Hlavním důvodem poklesu byly odpisy aktiv kvůli regulacím evropského energetického sektoru a související snižování velkoobchodních cen elektřiny , sdělila firma . Výsledek je tak výrazně pod očekáváním. Analytici totiž předpokládali, že čistý zisk ČEZ stoupne o víc než čtyři procenta na 34,8 miliardy korun. Společnost také oznámila, že kvůli snížení velkoobchodních cen elektřiny a regulatorním zásahům do evropského energetického sektoru snížila celoroční výhled čistého zisku na 35 miliard korun . Původně počítala s výsledkem o 2,5 miliardy vyšším. " Očekávané celoroční výsledky hospodaření ČEZ odrážejí současný stav energetiky v Evropě . Fakt, že na naše výsledky tato krize doléhá později a výrazně méně než na naše evropské konkurenty, reflektuje zejména naši úspěšnou strategii předprodejů elektřiny na roky dopředu a důraz na vnitřní úspory," uvedl k výsledkům předseda představenstva a generální ředitel Daniel Beneš .

Extrahuje program „ty správné klíčové fráze“?

⇒ obecnější otázka: dává program správný výstup?

- je třeba stanovit přesně cíl
- je třeba stanovit vzdálenost (nejlépe metriku) mezi výstupem a cílem

Rozpoznávání témat

. . . je zatím velmi vágně definovaný problém, tudíž má jen omezeně dobrá řešení.

Odkazy I


Allen, J. (1995).
Natural Language Understanding (2nd ed.).
Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.


Bhagat, R. and Hovy, E. (2013).
What is a paraphrase?
Computational Linguistics, 39(3):463–472.


Dagan, I., Roth, D., and Zanzotto, F. M. (2007).
Tutorial notes.
In *45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic. The Association of Computational Linguistics.


Lakoff, G. (1970).
Linguistics and natural logic.
Synthese, 22(1-2):151–271.