PLIN021 Sémantická analýza v praxi

OP VK Mezi bohemistikou a informatikou
www.projekt-inova.cz

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

15. března 2012

Pokračování z minule

Myslím si zvíře

Common sense

Minule . . .

. . . jsme se dívali na techniky strojového učení v praxi.

20q.net – 20 questions, Myslím si zvíře . . .

Myslím si zvíře

- objekty (řešení hádanky)
- otázky (max. 20 na hru, celkem ale mnohem víc)
- pořadí otázek (jak zvolit?)

Myslím si zvíře

	velký	savec	býložravec	umí skákat?	má rád vodu?
slon	99 %	92 %	87 %	43 %	63 %
kůň	78 %	93 %	91 %	93 %	45 %
velryba	99 %	74 %	65 %	71 %	100 %
motýl	12 %	2 %	24 %	4 %	9 %

Myslím si zvíře

Co je výsledkem?

Znalostní báze o zvířatech.

Nejde ale o vědecká fakta, jde o *common sense*.

Common sense

common sense

= sdílená znalost

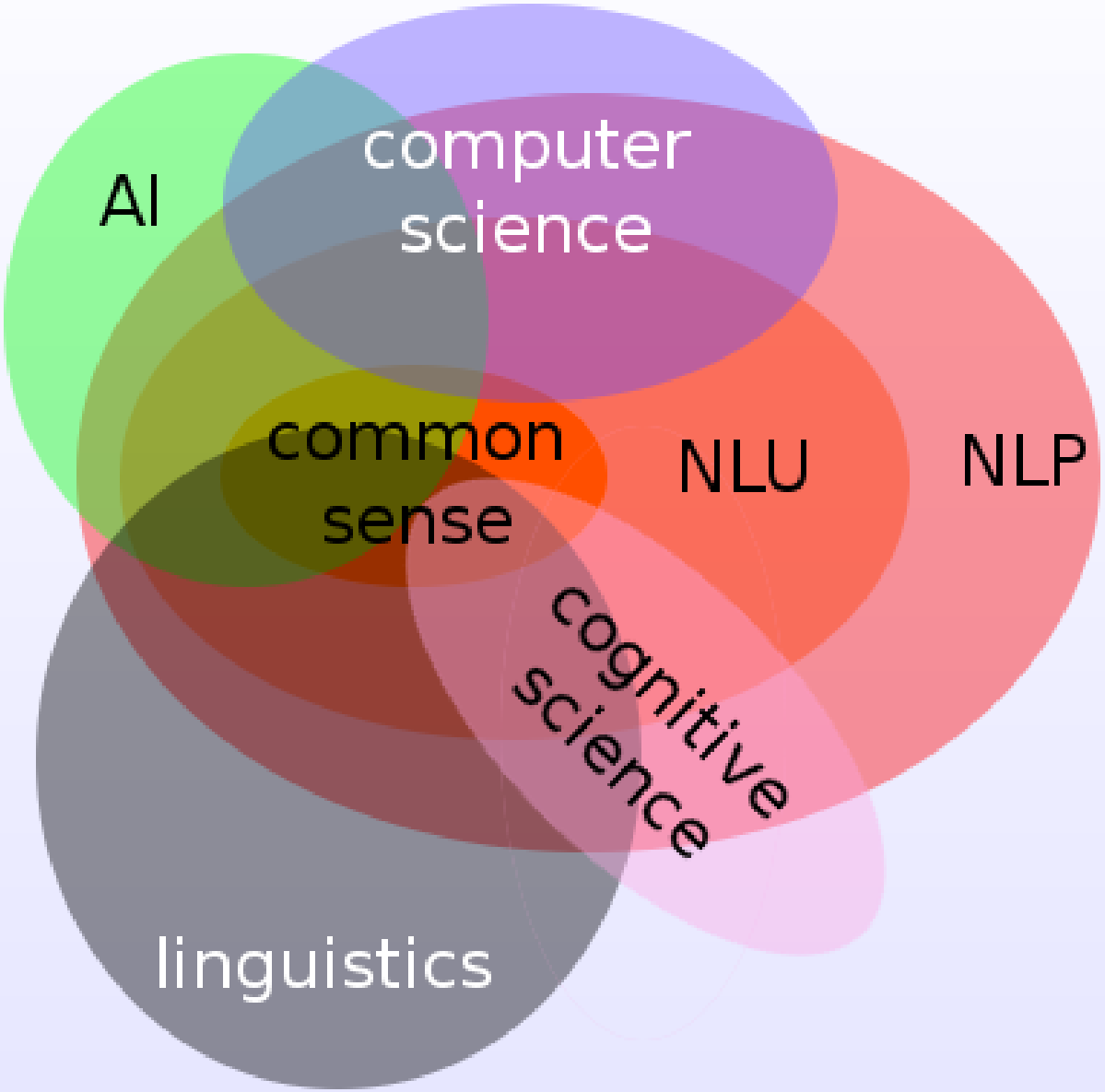
> *"Common sense includes commonsense knowledge – the kinds of facts and concepts that most of us know – but also the commonsense reasoning skills which people use for applying their knowledge. We each use terms like commonsense for the things that we expect other people to know and regard as obvious"*

[Minsky, 1999].

Charakteristiky common sense

- nemá pevnou hranici
- má velký rozsah
- není nutně vědecká znalost (někdy jde i proti ní)
- tvrzení *common sense* jsou příliš obyčejná, než aby je někdo někam psal
- bez *common sense* není možné úspěšně modelovat porozumění

Common sense v kontextu



Common sense is too common

Kde najdeme *common sense*?

- výkladové slovníky
- encyklopedie
- korpus
- sémantické sítě
- specializované kolekce

Kde najdeme *common sense*?

- výkladové slovníky
- encyklopedie
- korpus
- sémantické sítě
- specializované kolekce

Pozornost zasluhuje výkladový slovník WordSmyth (http://www.wordsmyth.net), který obsahuje výklady hesel na třech úrovních: začátečník (není rodilý mluvčí), dítě (nezná odborné termíny a cizí slova), pokročilý.

Ukázka Word Sketches (ve Sketch Engine). Korpusy všichni studenti znají, ale Word Sketches jsou vhodné pro zpracování velkých korpusů (které už dnes máme). WS dokážou sdružit slova, která se vyskytují ve stejných gramatických relacích (např. po předložce „na“ nebo adjektivum před slovem). Způsob sdružování je popsán ve Sketch Grammar, pomocí poměrně málo pravidel. Díky velikosti korpusu se zanedbají okrajové případy (je možné nastavit práh frekvence nebo skóre pro zobrazení).

Specializované kolekce *common sense*

- CyC (OpenCyC, ResearchCyC) http://www.cyc.com
- OpenMind http://openmind.media.mit.edu/
- ConceptNet http://conceptnet5.media.mit.edu/
- Games With a Purpose (GWAP) http://www.gwap.com/gwap/

- CyC (OpenCyC, ResearchCyC) http://www.cyc.com
- OpenMind http://openmind.media.mit.edu/
- ConceptNet http://conceptnet5.media.mit.edu/
- Games With a Purpose (GWAP) http://www.gwap.com/gwap/

Anotační hry jsou (kupodivu) předmětem zájmu. Jazykových dat máme totiž stále málo (divné, což?). Projekty jako Wikipedia ukázaly, že „neexperti“ jsou velmi užiteční, málokdy se dopouštějí vandalismu a jsou velmi levní. Vznikly i související portály pro *crowdsourcing*, např. Amazon Mechanical Turk.


Minsky, M. (1999).
The emotion machine: from pain to suffering.
In *C&C '99: Proceedings of the 3rd conference on Creativity & cognition*, page 7–13, New York, NY, USA. ACM.