NLP – zpracování přirozeného jazyka

Miloslav Konopík

14. května 2013

Co je to NLP?

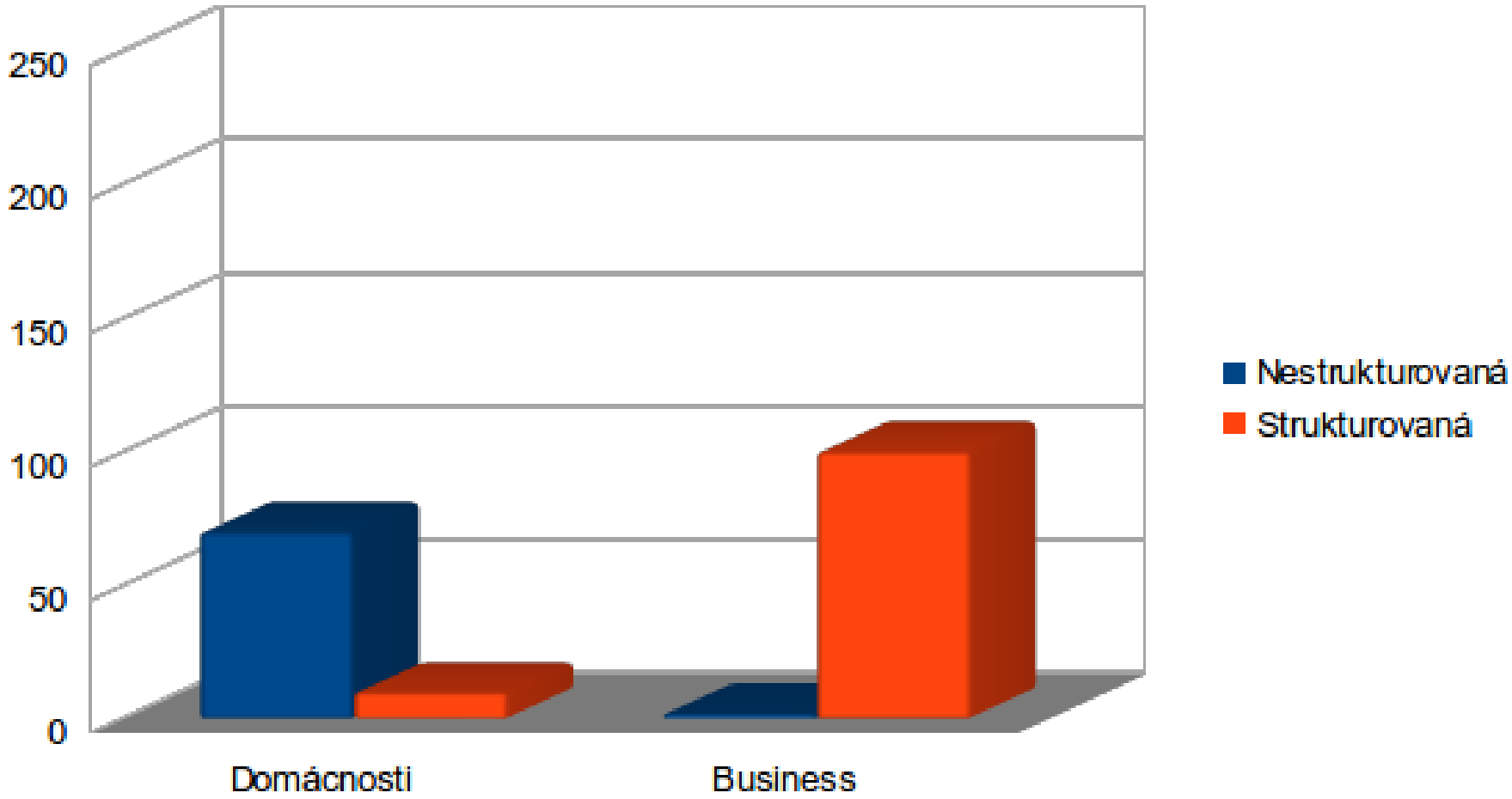
NLP = Natural Language Processing (zpracování přirozeného jazyka)
Computational linguistic (komputační lingvistika)

Aplikační oblasti

- Vyhledávání textů (Google).
- Strojový překlad (IBM word model)
- Podpora marketingu (analýza sentimentu).
- Podpora PR (třídění e-mailů).
- Podpora rozpoznávání (řeči, skenovaných textů).
- Oprava pravopisu.
- Odpovídání otázek (SIRI, IBM Watson).
- další...

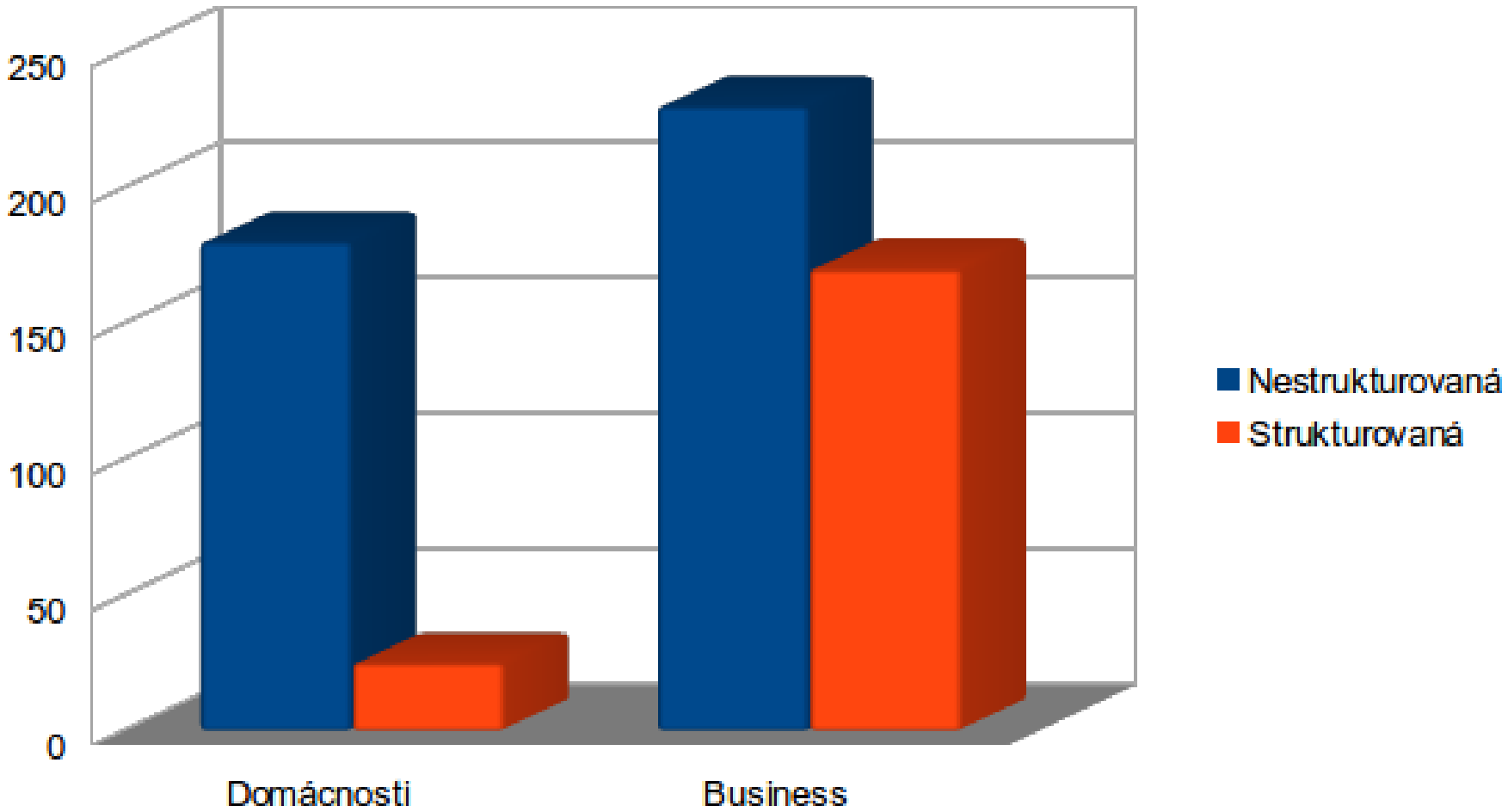
Nestrukturovaná data

Množství dat v osmdesátých létech.



Nestrukturovaná data

Množství dat nyní.

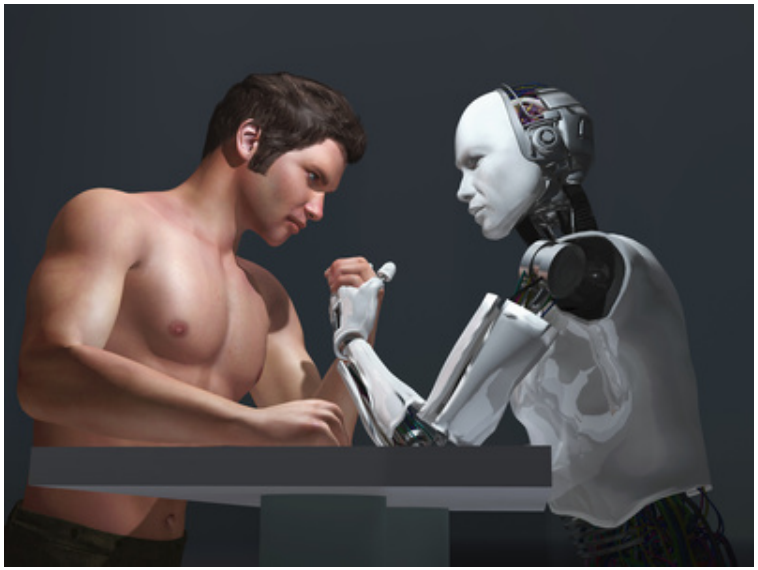


IBM Watson

14. února 2011 vyhrál vědomostní soutěž o 15-ti otázkách proti Kenu Jenningsovi a Bradu Rutterovi.

Člověk VS počítač

NLP není soutěž, ale nástroj, jakým může počítač pomáhat.



Například filtrování spamu.

Tokenizace

Začněme něčím jednodušším...

Tokenizace = rozdělení textu na:

- **slova**,
- **věty, resp. souvětí**,
- **dokumenty**,
- odstavce,
- věty souvětí,
- slabiky,
- další jednotky.

Tokenizace

Ale ani to není vždy zcela jednoduché.

在北京，如果迷失方向，
完全不必着急。
北京是个大城市，
北京人对外国人都很热情。

- zkratky (s.r.o.),
- data, hodiny (12. března 2013, 12:30, 3.3), ...

Rozpoznávání pojmenovaných entit

NER - Určení významu slov a slovních spojení, která mají určitý předem definovaný význam. Například:

- Jména osobností.
- Názvy měst.
- Názvy států.
- Názvy společností.
- Data.
- Čísla.
- ...

Datum Společnost

První systém pro strojový překlad byl představen 7. ledna 1954 v ústředí firmy IBM .

Nasazeno v ČTK.

Latentní sémantika

Předpoklad: význam slova je určen jeho okolím.
Nástroj: strojové učení a velká data.

> *You shall know a word by the company it keeps (Firth, J. R. 1957:11)*

Okolí:

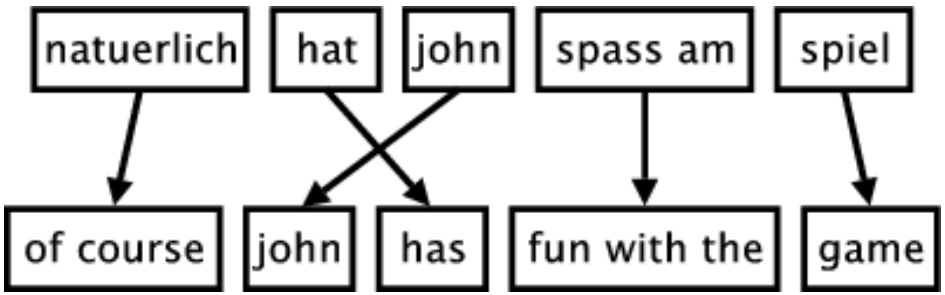
- Globální – LSA, PLSA, LDA.
- Lokální – HAL, COALS, RI.

Příklad:

- Globální – loď, plout, plavidlo, voda, výletní, posádka, kotvit.
- Lokální – vyplout, akcelerovat, potopit, miniaturizovat, 921, kotvit, dokař, odplout, manévrovat, plout, připlout, pilotovat.

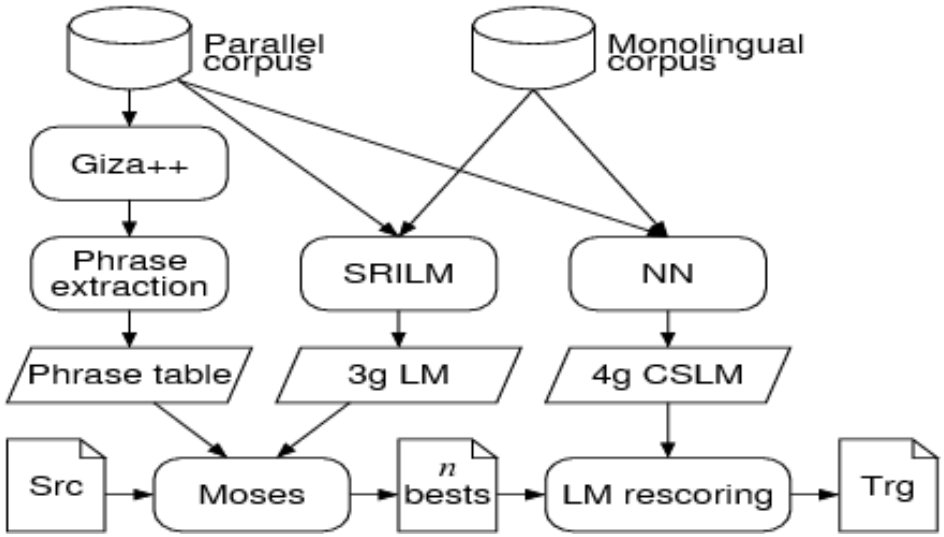
Strojový překlad

Předpoklad: Slova z přeložené věty a překládané věty by si měla odpovídat.
Nástroj: strojové učení a velká data – opět :)
Modely: IBM, frázový model.



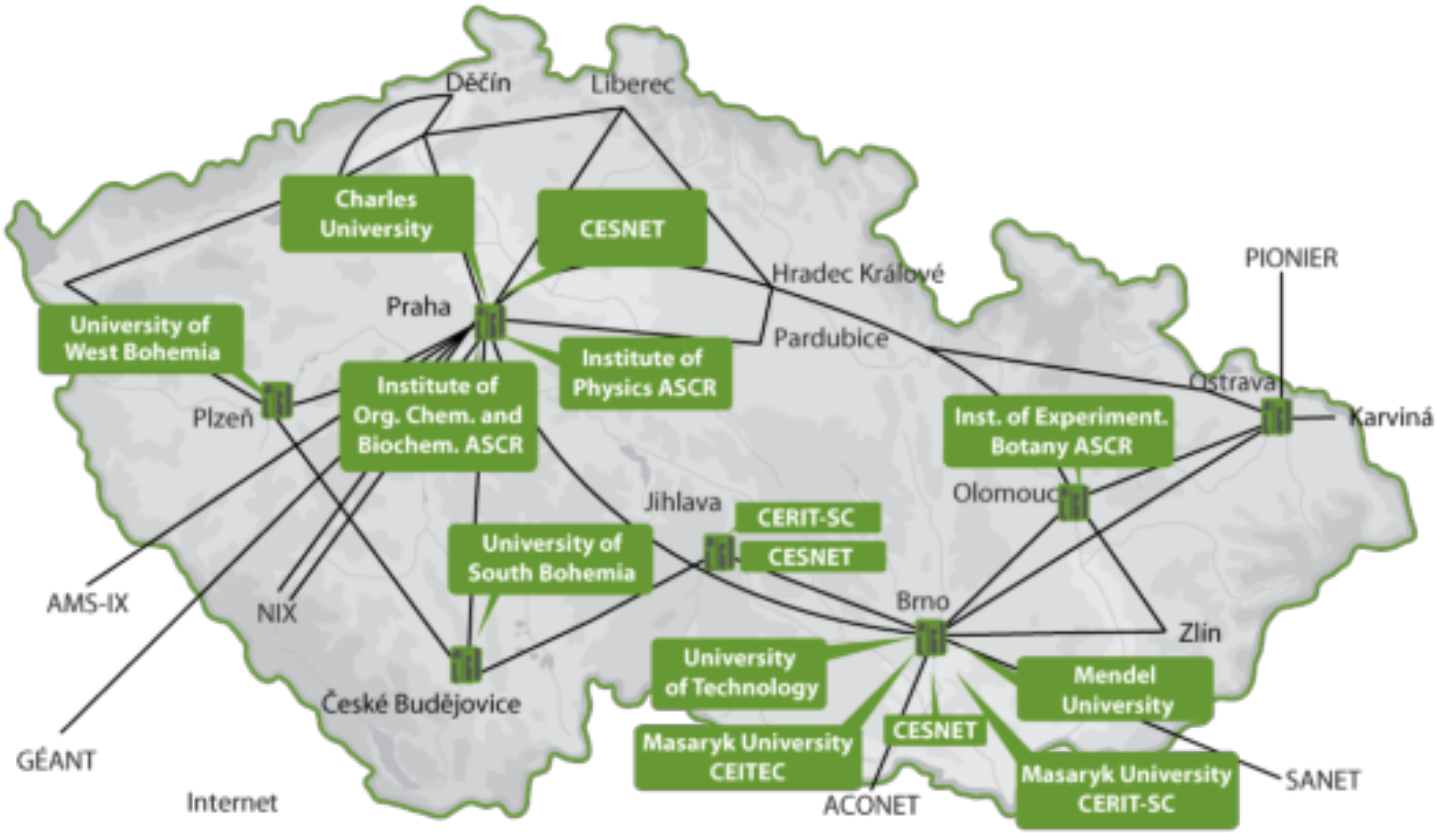
Strojový překlad

Softwarový nástroj: Moses
Zdroje dat: Europarlament, evropské zprávy, titulky, systémové hlášení, ...



Výpočetní prostředky

Metacentrum.



Jeden cluster z centra Cerit

Cluster zewura.cerit-sc.cz - 1 600 CPU

cluster SMP strojů s 80 CPU a 512GB RAM (Brno)

První SMP cluster pořízený CERIT-SC

Cluster zewura.cerit-sc.cz obsahuje 20 uzlů, každý z nich má následující hardwarovou specifikaci:

CPU	8x 10-core Intel Xeon E7-2860 2.26 GHz
RAM	512 GB
disk	20x 900GB v RAID-10 v celkové kapacitě 8 TB v každém uzlu
net	2x InfiniBand 4xQDR, 1x 10 Gbit/s Ethernet, 4x 1 Gbit/s Ethernet
poznámka	
vlastník	CERIT-SC/MU

zewura1 (80 CPU)	**zewura2 (80 CPU)**	**zewura3 (80 CPU)**	**zewura4 (80 CPU)**	**zewura5 (80 CPU)**	**zewura6 (80 CPU)**	**zewura7 (80 CPU)**
zewura8 (80 CPU)	**zewura9 (80 CPU)**	**zewura10 (80 CPU)**	**zewura11 (80 CPU)**	**zewura12 (80 CPU)**	**zewura13 (80 CPU)**	**zewura14 (80 CPU)**
zewura15 (80 CPU)	**zewura16 (80 CPU)**	**zewura17 (80 CPU)**	**zewura18 (80 CPU)**	**zewura19 (80 CPU)**	**zewura20 (80 CPU)**	