

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

29. října 2013

Word Sense Disambiguation

úkolem WSD je zjistit, jaký význam (z inventáře významů) má slovo ve vstupním textu

ukázali jsme si dva reprezentanty metod pro WSD: Leskův algoritmus pracující se slovníkovými definicemi a příklady užití a Yarowského algoritmus strojového učení

Word Sense Disambiguation: slabiny

největší slabinou je inventář významů

proto existují jednak snahy vytvořit dobré inventáře, jednak snahy úplně se inventářím vyhnout (HyperLex, [Véronis, 2004])

WordNet jako inventář významů?

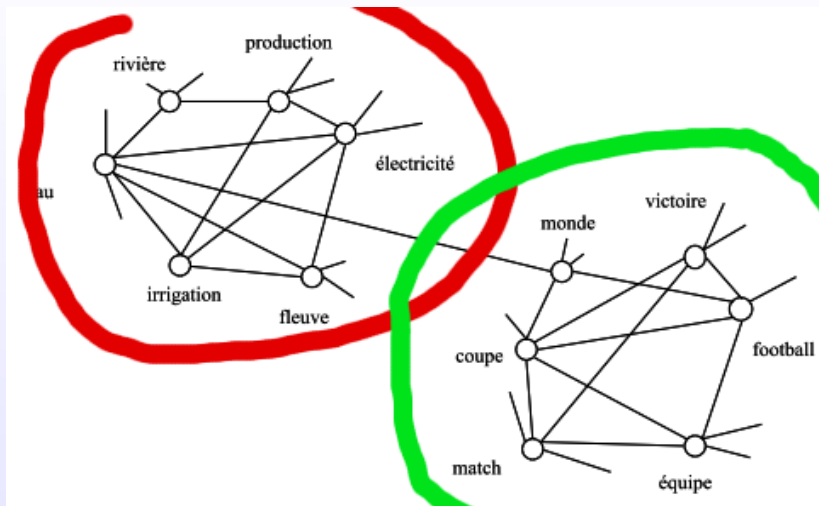
Princeton WordNet – ukázka

český WordNet – ukázka

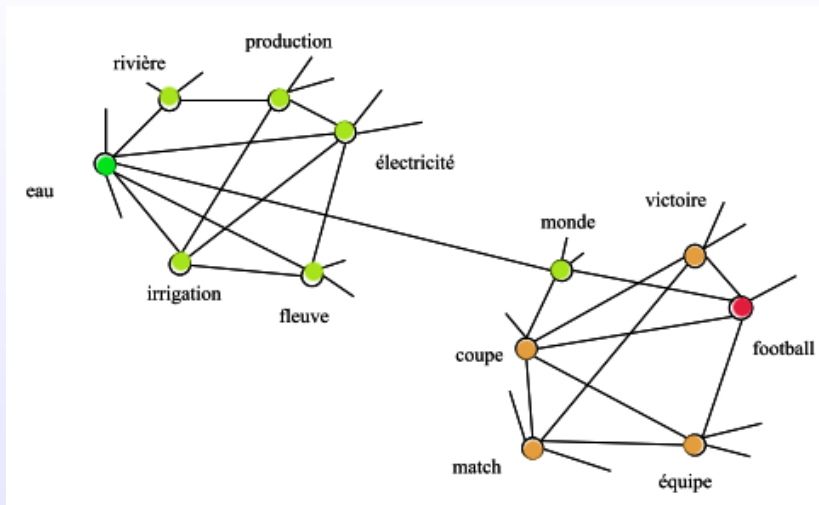
HyperLex, grafy

- „malé světy“ (Milgram, 1967)
- graf
- vážené hrany $A-B$:
 - $w = 0$, pokud se slova vyskytují vždy spolu
 - $w = 1$, pokud se nikdy spolu nevyskytují
 - $w_{AB} = 1 - \max[p(A|B), p(B|A)]$
- rozdělení grafu na podgrafy (NP-těžký problém)

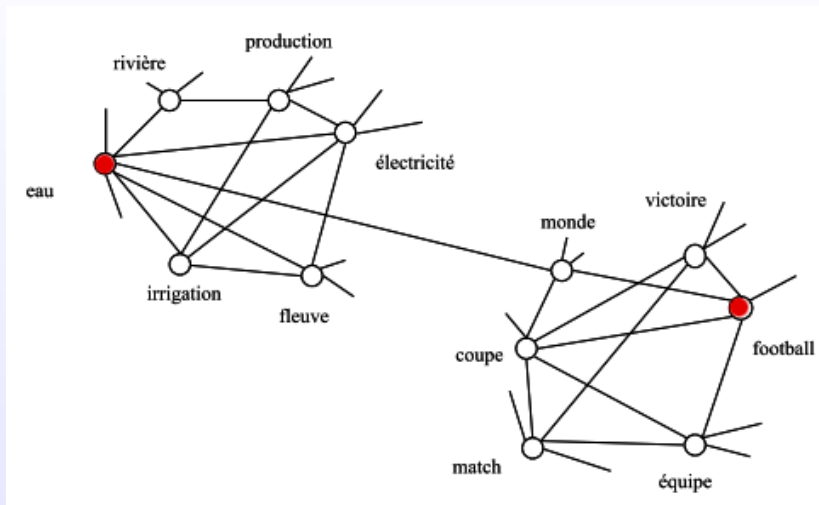
HyperLex: příklad „barrage“



HyperLex: nalezení kořenového uzlu



HyperLex: nalezení minimální kostry



Word Sense Disambiguation: shrnutí

- všechny algoritmy pro WSD pracují s kolokacemi
- všechny pracují s určitým oknem, ve kterém kolokace sledují

Word Sense Disambiguation: měření kvality

soutěž SENSEVAL (www.senseval.org)

- vyhodnocení systémů pro WSD
- od roku 1998 (Senseval-1, -2, -3, Semeval-2007, -2010)
- od Semeval-1 jsou úkoly různé (např. přiřazení emoce ke krátkému textu, detekce metonymie ...)
- čeština (zatím) chybí
- data z proběhlých kol jsou k dispozici



Véronis, J. (2004).

Hyperlex: Lexical cartography for information retrieval.

In *Computer Speech and Language: Special Issue on Word Sense Disambiguation*, page 23.