TiMBL – Shallow parser

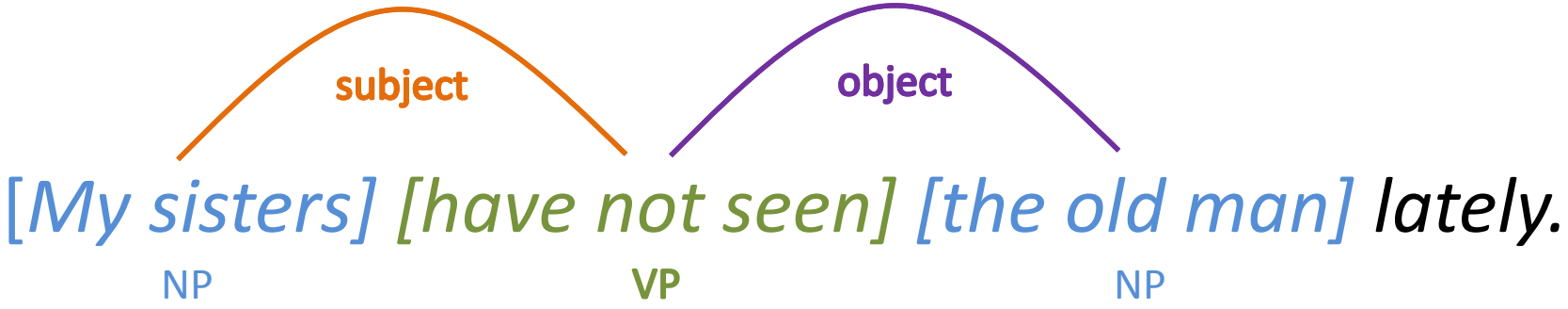
Tomáš Drusa (256167)

TiMBL

- Open-source software z Tilburg University, NL
- Implementace učení z instancí (memory-based learning)
  - IB1-IG (k nejbližších sousedů s váhováním)
  - IGTree (aproximace rozhodovacím stromem)

Shallow parsing

- Neboli chunking; „mělká“ analýza textu
- Určení jmenných, předložkových a slovesných frází a vztahů mezi nimi v textu



Proč učení z instancí?

- Strojové učení
  - minimalizace ruční práce (regulární výrazy…)
  - obecněji použitelné (jiná data, jazyk)
- Učení z instancí
  - umožňuje lépe odlišit jazykové výjimky od šumu
  - podobnostní vyhlazování na neúplných datech

Chunking jako klasifikace

- Analýza textu jako série klasifikačních úloh
- Chunking
  - pro každé slovo na základě lemmat a POS tagů kontextu <2, 1> urči třídu = značku typu fráze

NP_i	NP_i	O	NP_i	NP_i	O	O	VP_i	VP_i	NP_i
Pierre	*Vinken*	*,*	*61*	*years*	*old*	*,*	*will*	*join*	*the*
board	*as*	*a*	*nonexecutive*	*director*	*Nov*	*29*	*.*		
NP_i	O	NP_i	NP_i		NP_i	NP_b	NP_i	O	

Výsledky

- Angličtina, korpus WSJ
- Testování na 1/25 korpusu, trénink na zbytku
- Průměrné hodnoty z 25 běhů

Methods	context	accuracy	precision	recall	$F_{\beta=1}$
	NPs				
IGTree	2-1	97.5	91.8	93.1	92.4
IB1-IG	2-1	98.0	93.7	94.0	93.8
baseline words	0	92.9	76.2	79.7	77.9
baseline POS	0	94.7	79.5	82.4	80.9
	VPs				
IGTree	2-1	99.0	93.0	94.2	93.6
IB1-IG	2-1	99.2	94.0	95.5	94.7
baseline words	0	95.5	67.5	73.4	70.3
baseline POS	0	97.3	74.7	87.7	81.2

Vztahy frází jako klasifikace

- Hledání a určování typů vztahů mezi frázemi
  - pro potenciální dvojici frází (slovesná – jmenná)
  - žádný vztah / je podmětem / je předmětem
  - na základě:
    - lemmatu a značky slovesa slovesné fráze
    - lemmat a značek kontextu <2, 1> jmenné fráze
    - vzdálenosti (počtu mezilehlých slov/frází)
    - počtu mezilehlých čárek (,)
    - počtu mezilehlých slovesných frází

Vztahy frází jako klasifikace

- Které dvojice jsou potenciální?
  - příliš široký záběr -> zahlcení šumem a pomalost
  - praxe (AJ): maximálně 1 mezilehlá slovesná fráze

- Hlasování
  - IGTree lépe zvládá předmětné, lepší precision
  - IB1-IG lépe zvládá podmětné, lepší recall

Výsledky

	Together				Subjects			Objects		
# relations	51629				32755			18874		
Method	acc.	prec.	rec.	$F_{\beta=1}$	prec.	rec.	$F_{\beta=1}$	prec.	rec.	$F_{\beta=1}$
Random baseline		3.9	3.9	3.9	4.5	4.5	4.5	2.7	2.5	2.6
Heuristic baseline		65.9	66.5	66.2	69.3	61.6	65.2	61.6	75.1	67.7
IGTree	96.9	79.5	73.2	76.2	80.9	71.4	75.8	77.2	76.4	76.8
IB1-IG	96.6	74.4	**76.9**	75.6	76.2	**76.9**	76.5	71.5	**76.7**	74.0
IGTree & IB1-IG unanimous	97.4	**89.8**	68.6	**77.8**	**89.7**	67.6	**77.1**	**89.8**	70.4	**79.0**

Závěr

- Memory-Based Shallow Parsing
  - strojové učení s učitelem
  - jednoduchá, efektivní metoda
  - flexibilní (vnořené fráze, …)
  - úspěšností srovnatelná s konkurencí
- TiMBL
  - využívaná open-source implementace
  - paralelní, python a ruby interface, vizualizace, …

Děkuji za pozornost.

Pokud zbývá čas, nyní je vhodná chvíle na vaše dotazy.

- *Zdroje:*
  - *TiMBL: Timburg Memory-Based Learner* [on-line]. 7. října 2012 [cit. 11. 12. 2012]. WWW adresa: <http://ilk.uvt.nl/timbl/>.
  - DAELEMANS, Walter, BUCHHOLZ, Sabine, and VEERNSTRA, Jorn. *Memory-based Shallow Parsing*. In Proceedings of EMNLP/VLC-99, p. 239–246. University of Maryland, USA, June 1999. Dostupné on-line na <http://acl.ldc.upenn.edu/W/W99/W99-0707.pdf>.