Improving Collocation Correction by Ranking Suggestions Using Linguistic Knowledge

Roberto Carlini[1], *Joan Codina-Filba*[1], *Leo Wanner*[2,1]

(1) Natural Language Processing Group, Department of Information and Communication Technologies Pompeu Fabra University, Barcelona

(2) Catalan Institute for Research and Advanced Studies (ICREA)

Collocations:

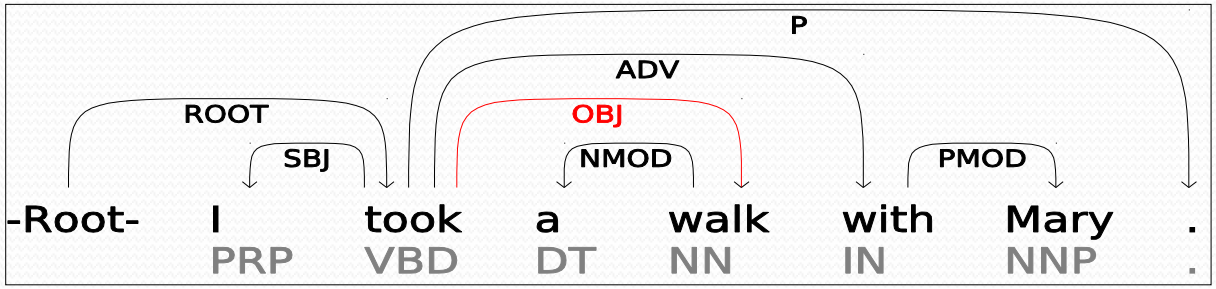
- Many definitions:
- Is a binary idiosyncratic co-occurrence of lexical items between which a direct syntactic dependency holds and where the occurrence of one of the items (the *base*) is subject of the free choice of the speaker, while the occurrence of the other item (the *collocate*) is restricted by the base (Hausmann (1984), Cowie (1994), Mel'ˇcuk (1995))
- "collocations should be defined not just as 'recurrent word combinations', but as 'arbitrary recurrent word combinations' ". Benson (1989)

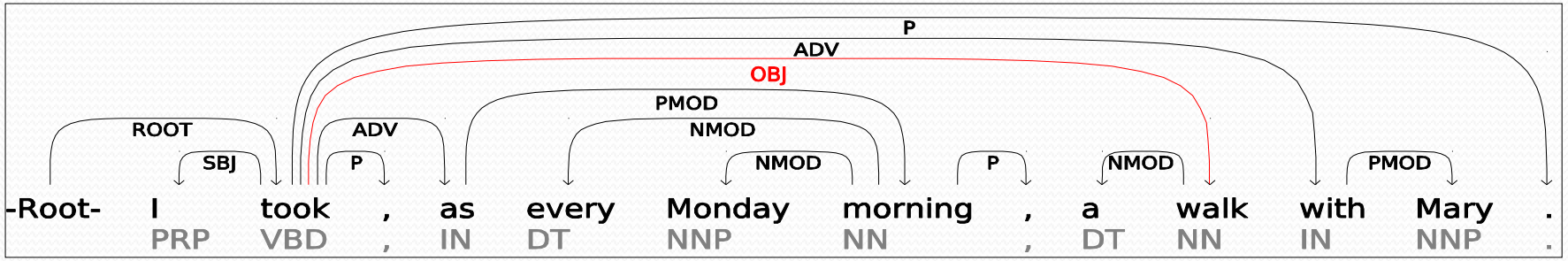
Collocations: Base and collocate

- The base (noun) is free: ***a walk***
- Once the base is chosen, there are some arbitrary recurrent collocates (verbs) to express the corresponding meaning : ***do, go for, go on, have, take***
- Not only verb+noun combinations
  - adj+noun(base): ***~~low~~/short ~~long~~/tall men***
  - verb(base)+adverb: ***walk calmly***

Dependency parsing

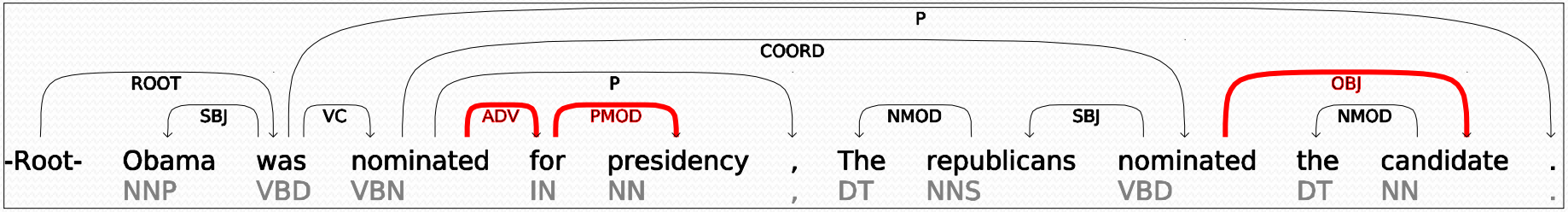
- Current systems: Based on co-occurrence in a corpus:
  - Sentence/Window
  - But what is near is not always what is syntactically related
- Dependency parsing finds direct relations

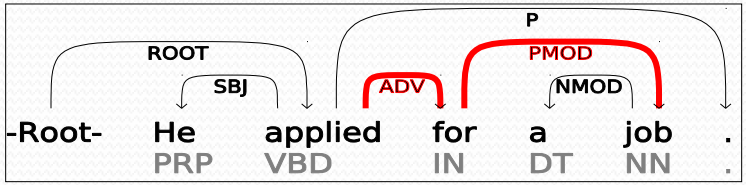


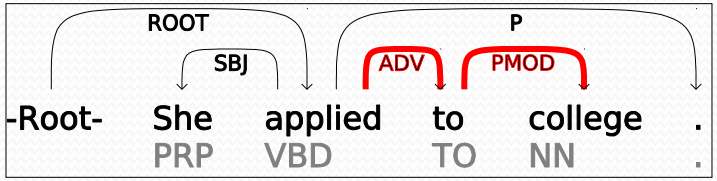


Dependency parsing

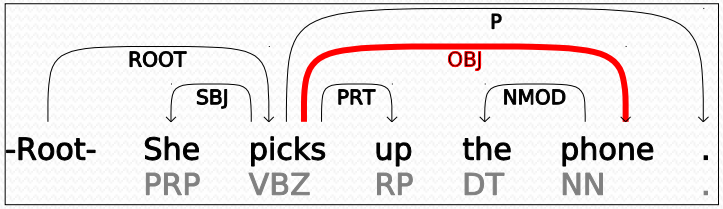
- Detection of governing prepositions

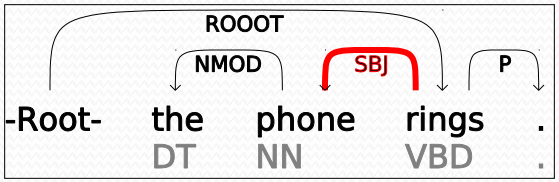






- Each dependency relation is assigned separately





Arbitrary recurrent?

- "collocations should be defined not just as 'recurrent word combinations', but as 'arbitrary recurrent word combinations' ". Benson (1989)
- PMI: Pointwise Mutual information
  - Detects whether two words are found together more often than expected if they were random independent variables.

$$PMI(base, collocate) = \log\left(\frac{P(base \cap collocate)}{P(base)P(collocate}\right)$$

PMI: Pointwise mutual information

The maximum value of PMI is not fixed ➔ making difficult to rank collocates a learner can pick from ➔ NORMALIZE

Collocations must have a positive value of PMI

$$PMI(b,c) = \begin{cases} \log\left(\frac{1}{p(b \cap c)}\right) & if\ p(b \cap c) = p(b) = p(c) \\ 0 & if\ p(b \cap c) = p(b)p(c) \\ -\infty & if\ p(b \cap c) = 0 \end{cases}$$

Normalize PMI

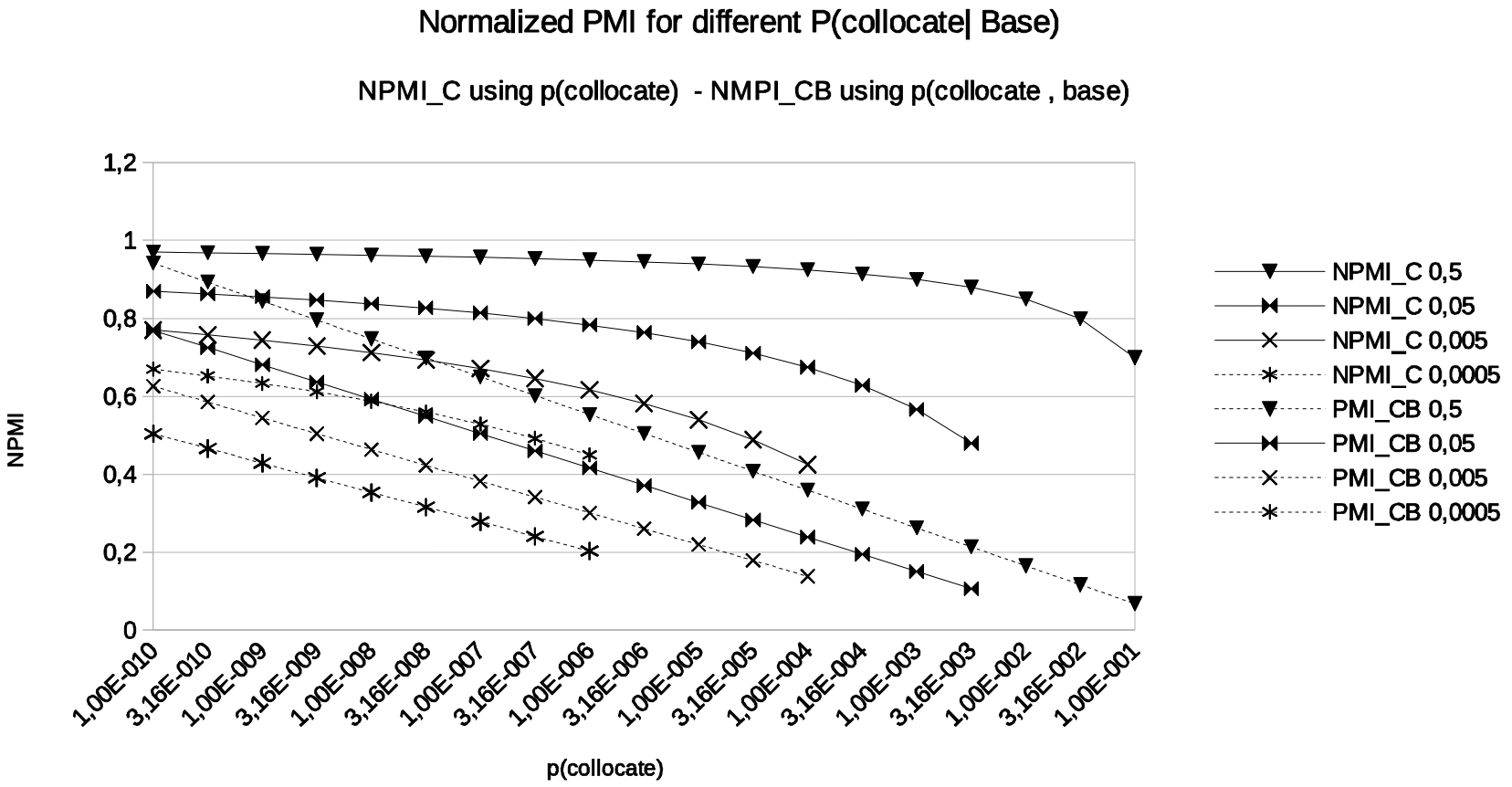
- Bouma, G. (2009). Normalizes PMI using combined probability

$$NPMI_{bc} = -\frac{PMI}{\log(p(b \cap c))}$$

- It penalizes the collocates which are very common words
- We normalize PMI using only probability of the collocate

$$NPMI_{c} = -\frac{PMI}{\log(p(c))}$$

Normalize PMI



Results

Collocates for *telefono* 'phone'

collocate	Freq. *Collocate*	Freq. *base ∩ collocate*	PMI	$NPMI_{CB}$	$NPMI_{C}$
pinchar	403	77	2,789	0,558	**0,652**
descolgar	230	61	2,931	0,575	**0,648**
sonar	2218	105	2,183	0,449	**0,617**
coger	4627	123	1,932	0,403	**0,6**
intervenir	1294	55	2,136	0,415	**0,566**
colgar	1723	61	2,057	0,403	**0,564**
llamar	12957	111	1,44	0,298	**0,519**
desconectar	234	14	2,285	0,398	**0,506**
contestar	3066	41	1,634	0,31	**0,481**
atender	8563	57	1,331	0,259	**0,451**
usar	6286	46	1,372	0,263	**0,444**
habilitar	760	14	1,773	0,309	**0,443**

Conclusions:

- We use dependencies to find the sentences where the base and collocate are syntactically related
- We can differentiate between the different dependency relations and detect governing prepositions
- We changed the Normalized PMI to avoid a penalization of common collocates and take into account that collocations are asymmetric
- As a result, the rank is more natural and suggestions to learners are improved