



Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-Speech Equivalences

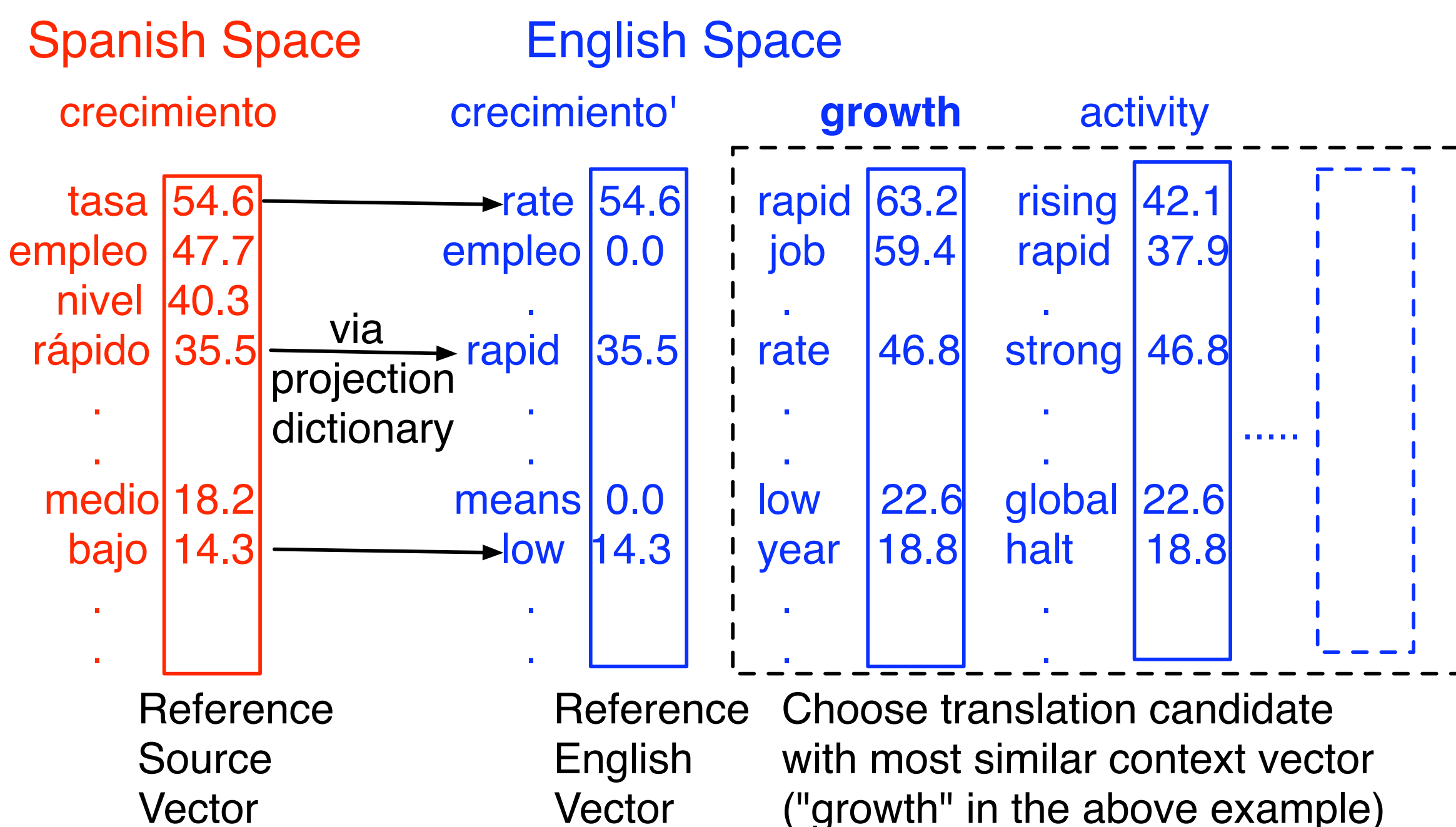
Nikesh Garera, Chris Callison-Burch, David Yarowsky
Johns Hopkins University

Problem

Learn translation lexicons from monolingual corpora given a set of seed translations (projection dictionary)

Context-based Rapp Model

(Rapp, 1999; Koehn and Knight, 2002; Schafer and Yarowsky, 2002; Haghighi et al., 2008)



1. Extract context vectors from monolingual corpora
2. Project reference source vector via projection dictionary
3. Rank vectors of candidate translations by similarity to the projected reference vector

Questions addressed

- ▶ Does a richer model of context help?
- ▶ How to account for changes in word order between languages?
- ▶ Can we extend and improve lexicon induction for all part-of-speech categories?

Dependency Contexts

... el camino para el crecimiento y la prosperidad económica ...
(the) (path) (to) (the) (growth) (and) (the) (prosperity) (economic)

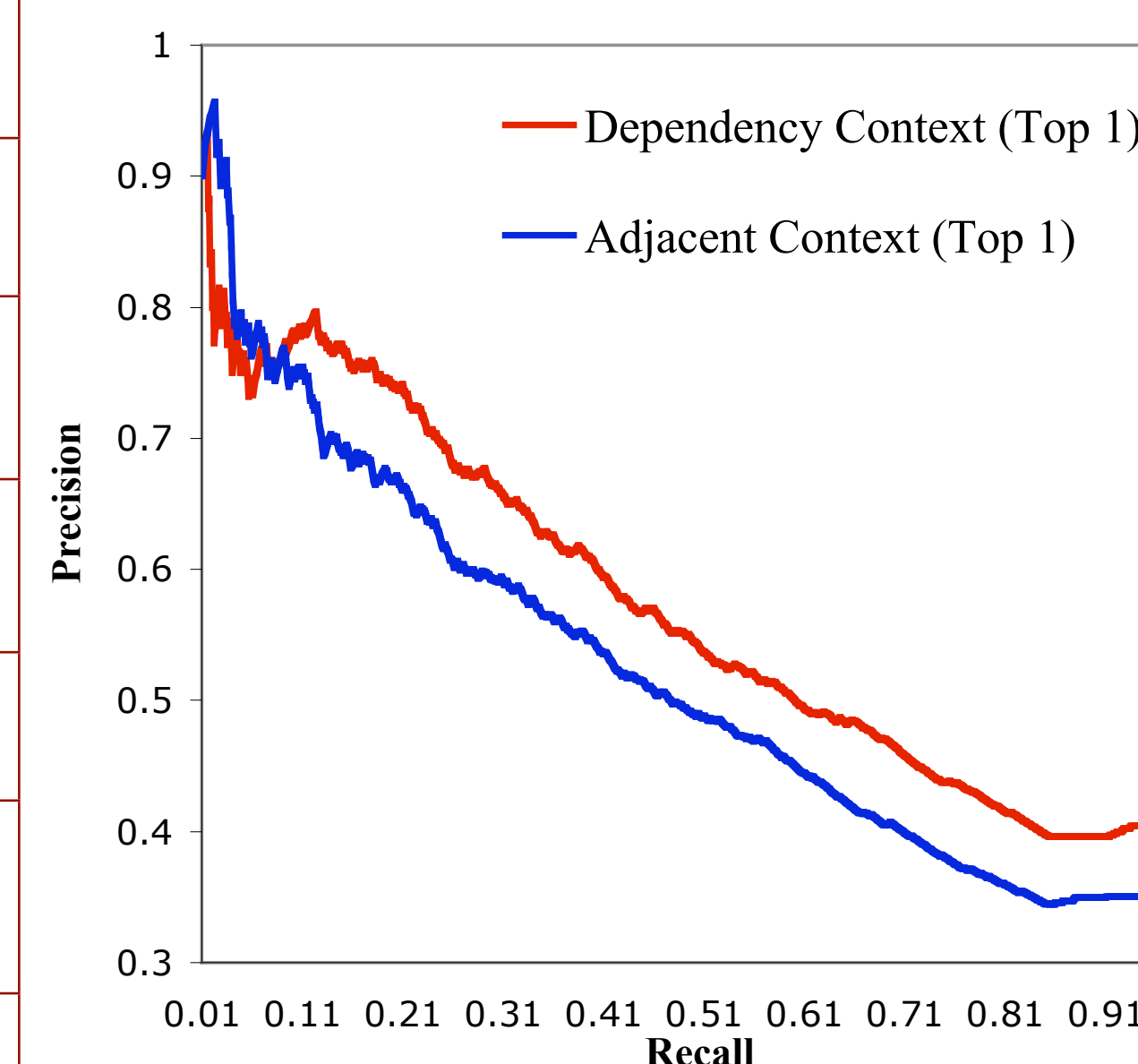
Position	Adjacent Context	Dependency Context
-2	para	camino
-1	el	para
+1	y	prosperidad, y, el
+2	la	económica

- Dynamic context size (nodes can have more than one children)
- Same dependency positions in case of reordering (E.g: Noun Adj => Adj Noun)

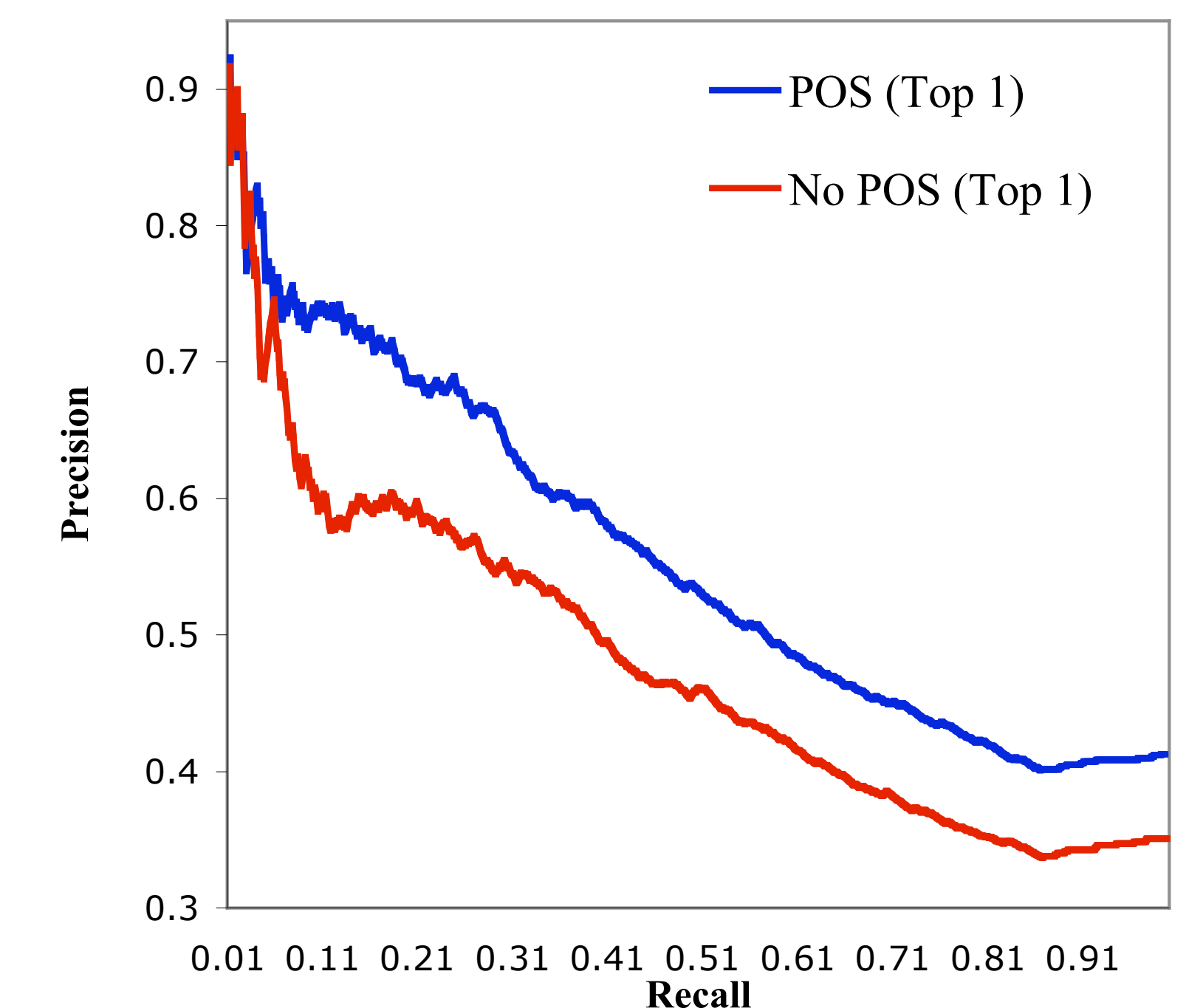
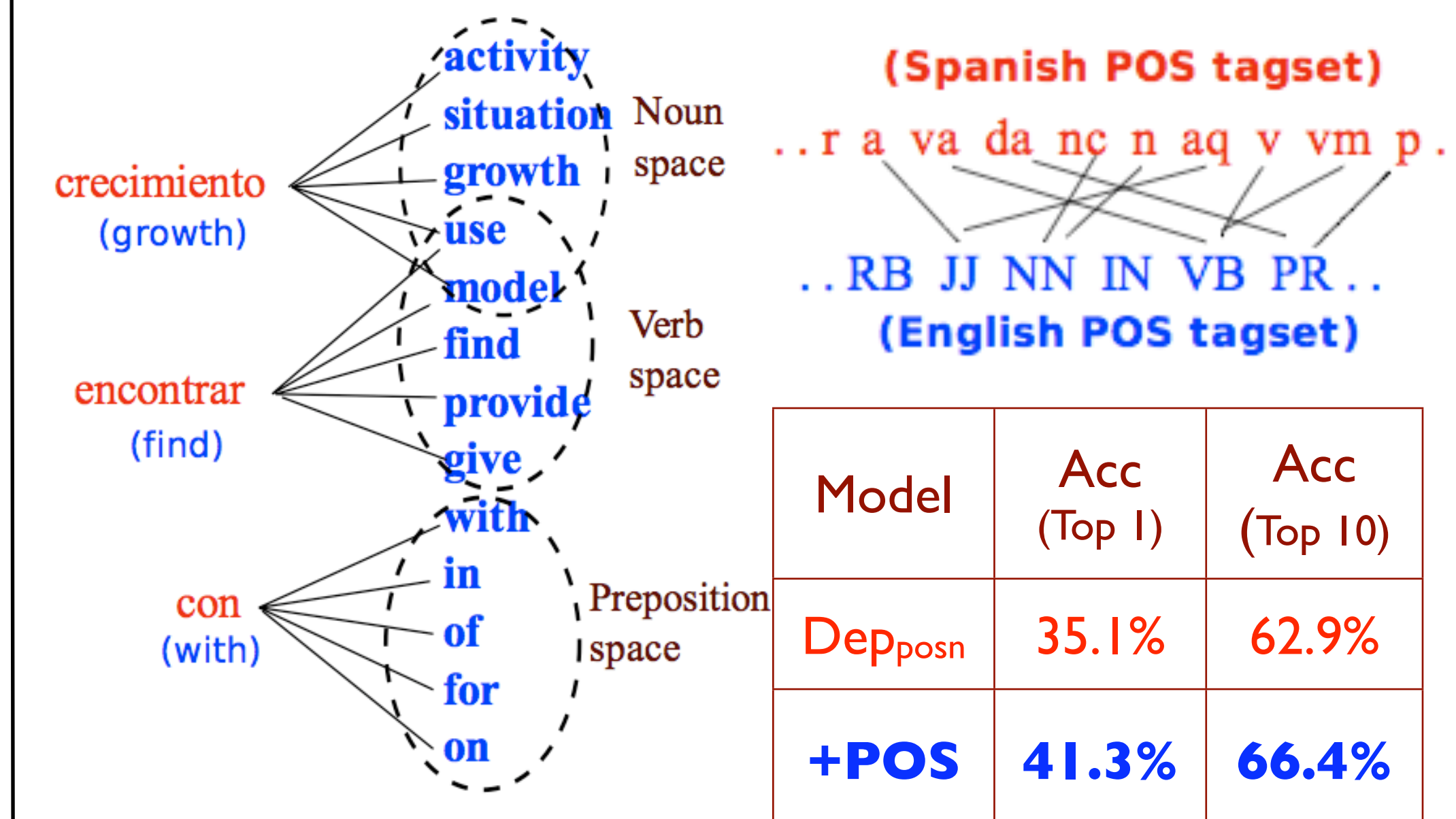
Translation candidates for "camino"			
	Adjacent Context Model	Dependency Context Model	
intentions	0.22	way	0.12
way	0.21	solution	0.10
idea	0.20	steps	0.09
thing	0.20	path	0.09
faith	0.18	debate	0.08
steps	0.17	account	0.08
example	0.17	means	0.08
news	0.16	work	0.08

Results

Model	Acc _{Top 1}	Acc _{Top 10}
Adj _{bow}	35.3%	59.8%
Adj _{posn}	20.9%	46.9%
Dep _{bow}	41.0%	62.0%
Dep _{posn}	41.0%	64.1%
Dep _{posn+rev}	42.9%	65.5%
Moses _{en-es-100k}	56.4%	62.7%



Generalizing to other word types via tagset mapping



Challenges and Future Work

- Learning lexicons from unrelated corpora (news, blogs, etc.). Accuracy decreases when unrelated news corpus of same size was utilized (59.8 => 48.8)
- Learning lexicons for rare words. Accuracy decreases when evaluated on rare words (59.8 => 24.9)