PARSEME WG 1

Extracting MWE and fixed structures from parsed corpora

EU COST Intitiative Meeting

Athens, Greece, March 11-12

Dr. Gerold Schneider

Institute of Computational Linguistics, University of Zurich

English Department, University of Zurich

gschneid@ifi.uzh.ch

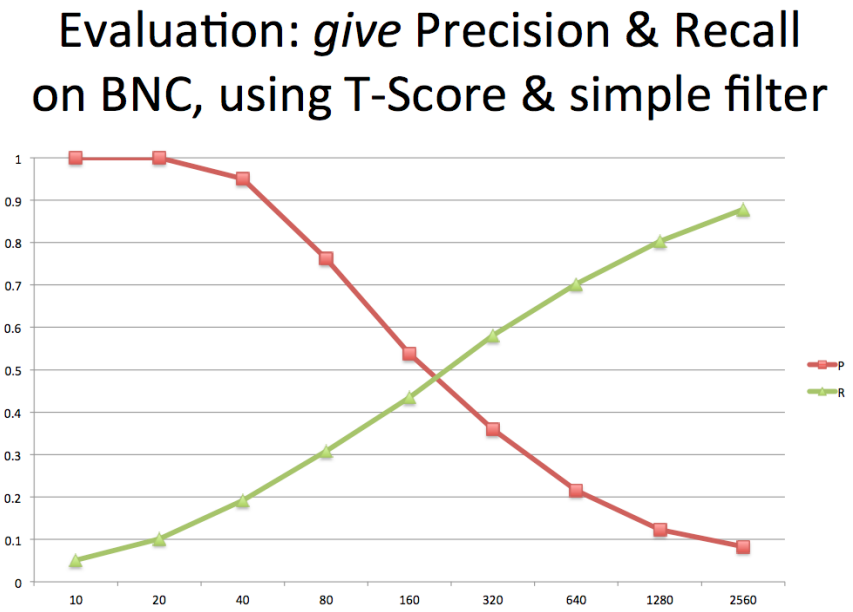
1. We extract MWE from parsed corpora with collocation measure

- Verb-preposition structures: O/E works very well on parsed data

Table 10. VOPN 4-tuples ordered by O/E, filtered by t-score in BNC-W written. (Full table.)

verb	object	prep	desc noun	t-score	O/E
send	shiver	down	spine	5.74456	2.21477×10^{8}
tap	esc	for	escape	6.40312	2.1134×10^{8}
separate	shield	from	plate	6.78233	2.33384×10^{7}
refer	gentleman	to	reply	8.24621	7.8143×10^{6}
obtain	property	by	deception	5.2915	7.60043×10^{6}
ask	secretary	for	affairs	6.40312	5.01529×10^{6}
kill	bird	with	stone	5.38516	3.37917×10^{6}
add	insult	to	injury	6.08276	2.21769×10^{6}
throw	caution	to	wind	5.09902	2.03157×10^{6}
refer	friend	to	reply	7.54983	1.36298×10^{6}
report	loss	on	turnover	7.14142	1.34742×10^{6}

- Light verb constructions: T-score works best

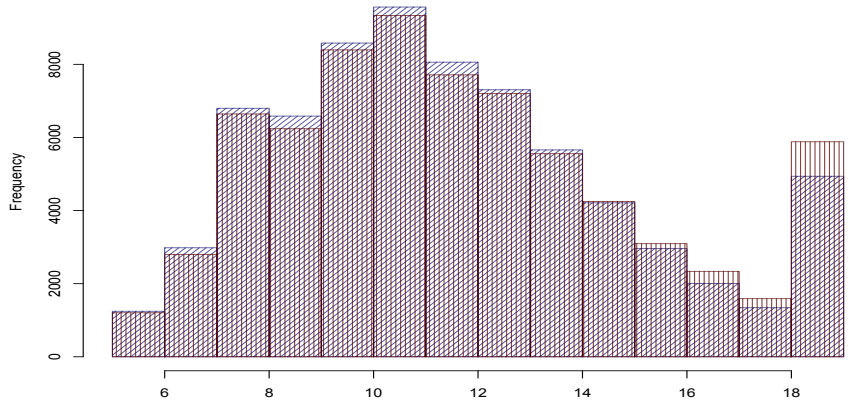


- subject-verb-object and others

Lehmann and Schneider (2011) show whole gradience. Strong but gradient idiomatic and selectional preferences prevail on all levels → probabilistic construction grammar → report lexical priming, information measures like surprisal (Levy and Jaeger, 2007)

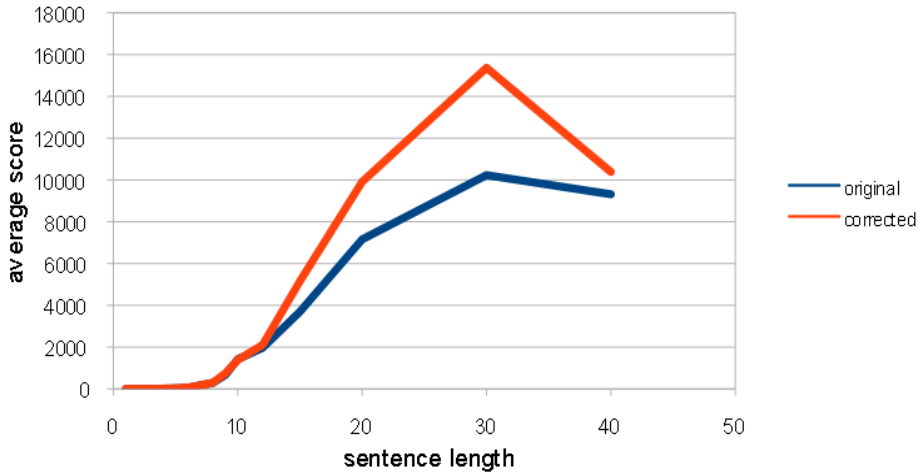
2. We measure MW gradience using surprisal and our parser

Lexical priming Hoey (2005)
Pawley and Syder (1983): native speakers know best how to play the game of *idiom* vs. *syntax* principle (Sinclair)
Language learners produce less fixed, less entrenched structures. We use the NICT Japanese Learner English (JLE) Corpus. Bigram surprisal = $log\frac{1}{p(w_{n-1})} + log\frac{1}{p(w_n|w_{n-1})}$



bigram surprisal = log 1 / p(w1) + log 1 / p(w2 | w1). Original (red), Corrected (blue). Unseen bigrams = 19

We also use our parser as a language model. It uses tri-lexical disambiguation (Collins, 1999) $p(R, dist|w1, w2, w3)$ and other stats



Learner utterances have lower model fit $\rightarrow$ less expected, parser can map them less well to any syntactic analysis.

References

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Hoey, Michael. 2005. *Lexical priming: A New Theory of Words and Language*. Routledge.

Lehmann, Hans Martin and Gerold Schneider. 2011. A large-scale investigation of verb-attached prepositional phrases. In S. Hoffmann, P. Rayson, and G. Leech, editors, *Studies in Variation, Contacts and Change in English, Volume 6: Methodological and Historical Dimensions of Corpus Linguistics*. Varieng, Helsinki.

Levy, Roger and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

Pawley, Andrew and Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. C. Richards and R. W. Schmidt, editors, *Language and Communication*. Longman, London, pages 191–226.

Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.