Building A Thesaurus Using LDA-Frames

Jiří Materna

Centre for Natural Language Processing
Faculty of informatics, Masaryk University
Brno, Czech Republic

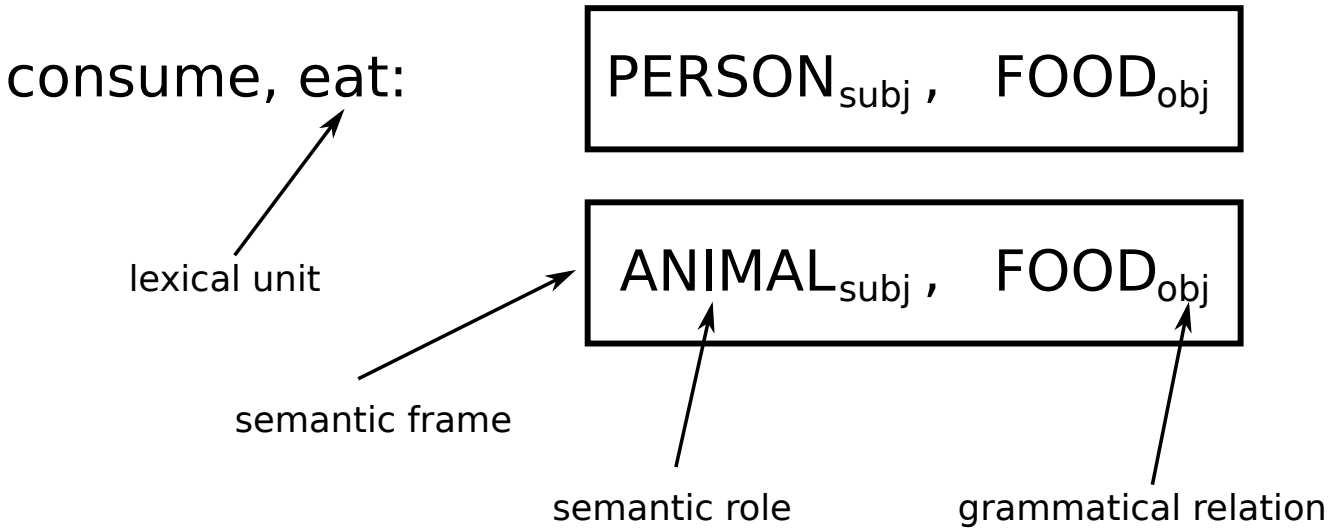
December 8, 2012

home page: http://nlp.fi.muni.cz/projekty/lda-frames/
code: http://code.google.com/p/lda-frames/
e-mail: xmaterna@fi.muni.cz
twitter: @JiriMaterna

Thank you for your attention.

Semantic frames

- terminology adopted from Frame Semantics
- captures selectional preferences of grammatical relations



LDA-frames

- unsupervised method for discovering semantic frames
- generative model based on Latent Dirichlet Allocation
- language independent
- need for a syntactically annotated corpus + number of frames and roles
- for each lexical unit a probability distribution over frames
- semantic roles represented as probability distributions over

LDA-frames – corpus data

Set of frame realizations for each lexical unit

Lexical unit	**subject**	**object**	**frame**
eat	John Mike man	food pizza cake	(Person, Food)
	dog mouse	meat cheese	(Animal, Food)
drink	Jane Mike	coffee tee	(Person, Drink)
teach	teacher professor	student Mike	(Person, Person)
	Peter	dog	(Person, Animal)

Table: Example of grammatical relation realizations.

LDA-frames

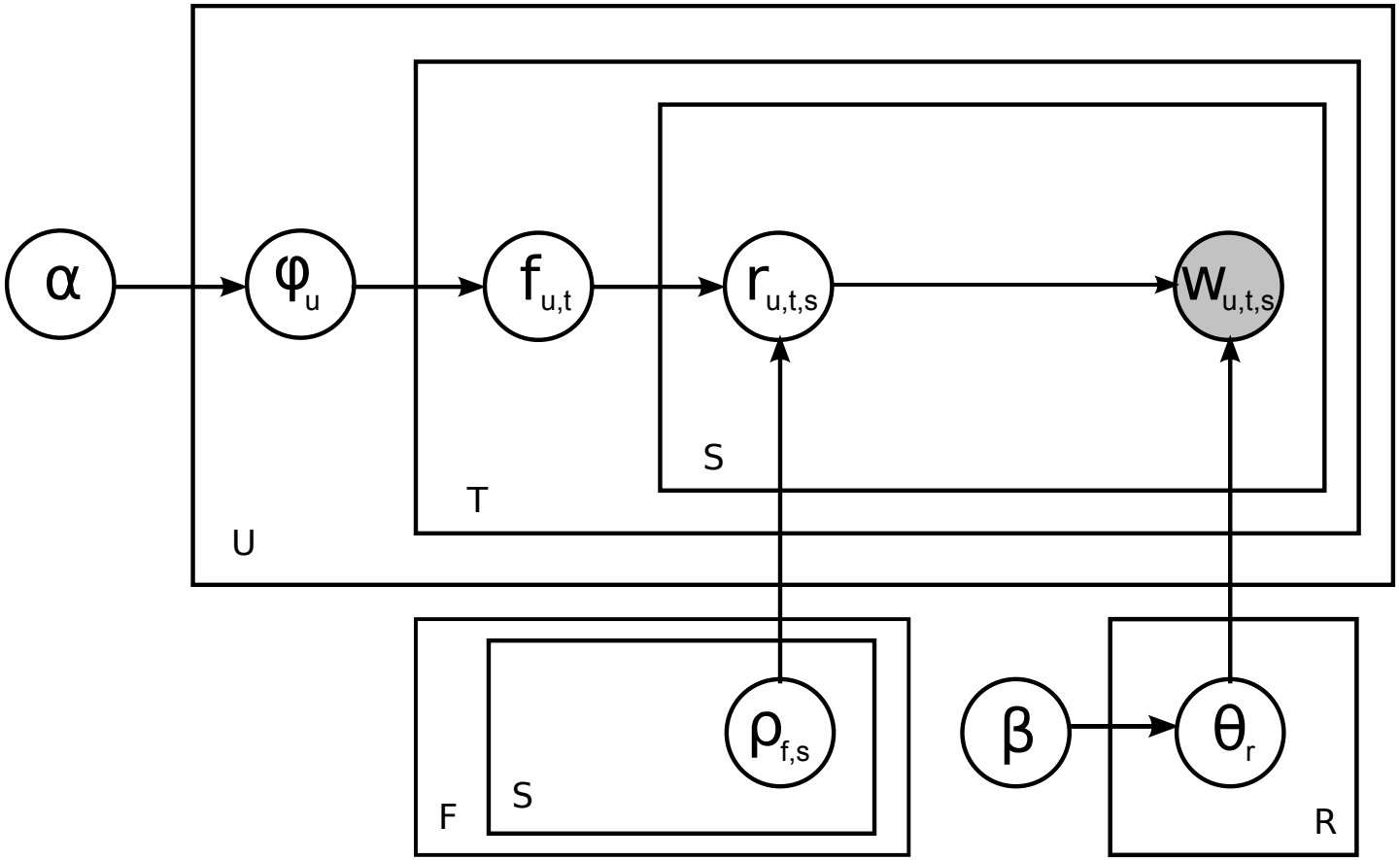


Figure: Graphical model of LDA-frames.

Measuring Semantic Relatedness

Measured as the similarity of φ distributions.

$$H(a,b) = \sqrt{\frac{1}{2}\sum_{f=1}^{F}\left(\sqrt{P(f|a)} - \sqrt{P(f|b)}\right)^2}$$

Lexical unit	drink	eat	ingest	gulp	smoke	sip	devour	slurp
Distance	0.619	0.622	0.658	0.661	0.666	0.681	0.691	0.691

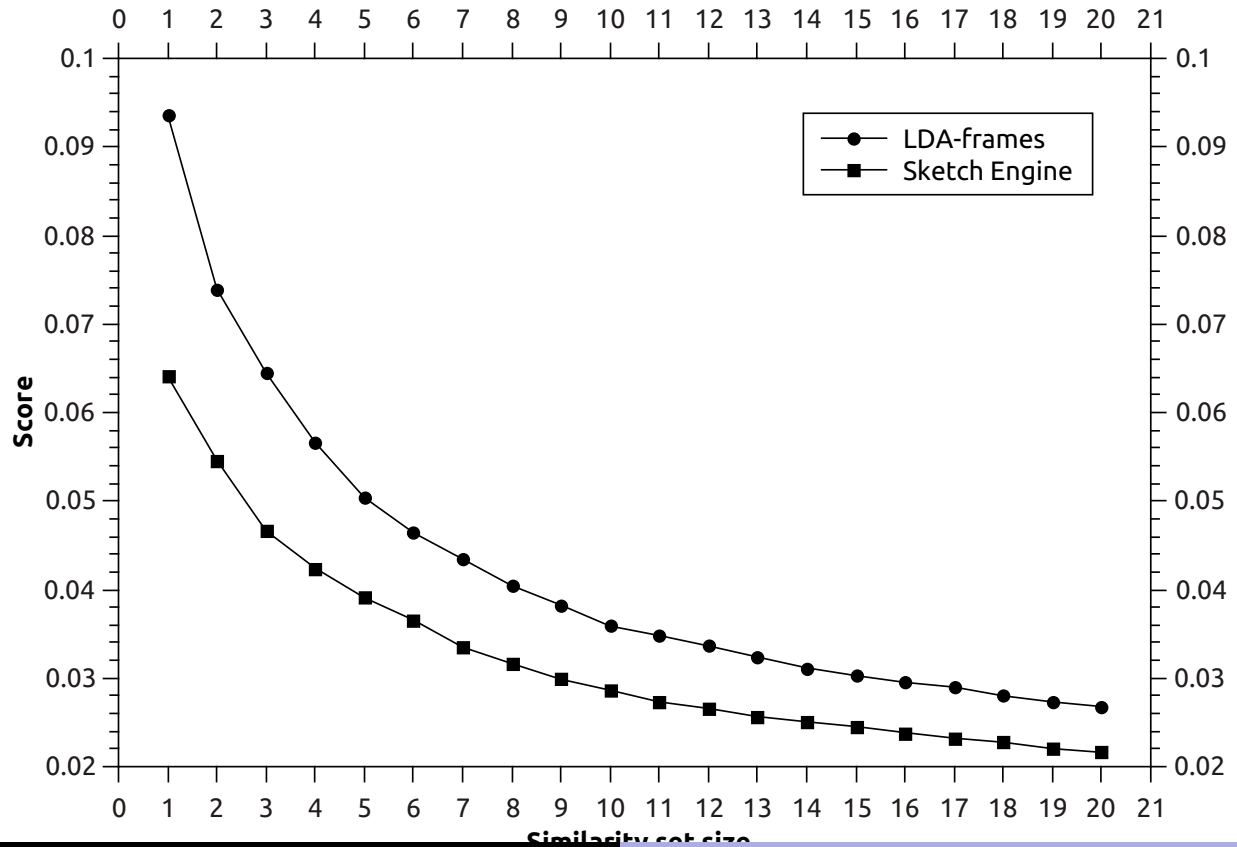
Table: The most similar lexical units to *consume*.

The experiment I

- LDA-Frames generated using 1.4 millions of (*verb*, *subject*, *object*) tuples
- Word Sketches generated on BNC just using *subject* and *object* grammatical relations
- thesauri for similarity set sizes: $1 \leq n \leq 20$
- comparison with WordNet 3.0

The experiment II

$$Score(T) = \frac{1}{V} \sum_{v=1}^{V} \frac{|T(v) \cap W(v)|}{|T(v)|}$$



Conclusions

- LDA-frames algorithm outperforms a similar approach from the Sketch Engine
- only two grammatical relations have been taken into consideration
- enhancing train data by other grammatical relations should lead to significantly better results

home page: http://nlp.fi.muni.cz/projekty/lda-frames/
code: http://code.google.com/p/lda-frames/
e-mail: xmaterna@fi.muni.cz
twitter: @JiriMaterna

Thank you for your attention.