A comparison of selectional preference models for automatic verb classification

Will Roberts and Markus Egg

Institut für Anglistik und Amerikanistik
Humboldt Universität zu Berlin

Sunday, 26 October, 2014

Outline

Selectional preferences

- Predicates can select for their arguments:

  ? My aunt is a bachelor. (McCawley, 1968)

- We model verbs empirically:

  I eat meat
  bread
  fruit
  ⋮
  newspaper

- Evaluate on an automatic verb classification task
- Baseline model clusters verbs based on *subcategorisation*

Selectional preferences

Example

Wir	**benutzen**	Ihre Umfragedaten	nicht	für eigene Zwecke.
We	use	your survey data	not	for own purposes.

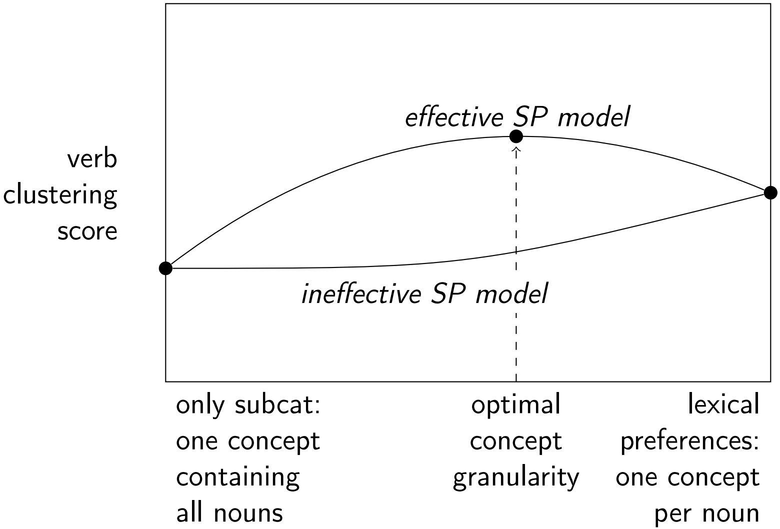
We will not use your survey responses for private purposes.

- We will want to record that this instance of *use* has:

Subject	wir, *we* (pronoun, ignored)
Direct object	Umfragedatum, *survey datum*
PP (für, *for*)	Zweck, *purpose*

- We also include indirect objects (datives)
- A selectional preference model will map noun forms onto concept labels

Hypothesis


verb clustering score
effective SP model
ineffective SP model
only subcat: one concept containing all nouns
optimal concept granularity
lexical preferences: one concept per noun

Subcategorisation

Example

Wir	**benutzen**	Ihre Umfragedaten	nicht	für eigene Zwecke.
We	use	your survey data	not	for own purposes.

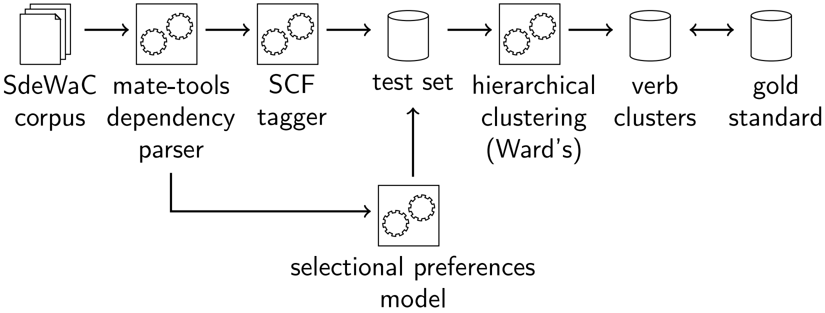
We will not use your survey responses for private purposes.

- The combination of syntactic argument types is assigned a *subcategorisation frame* (SCF) code:

$$\text{benutzen} \Rightarrow \texttt{nap:für.Acc}$$

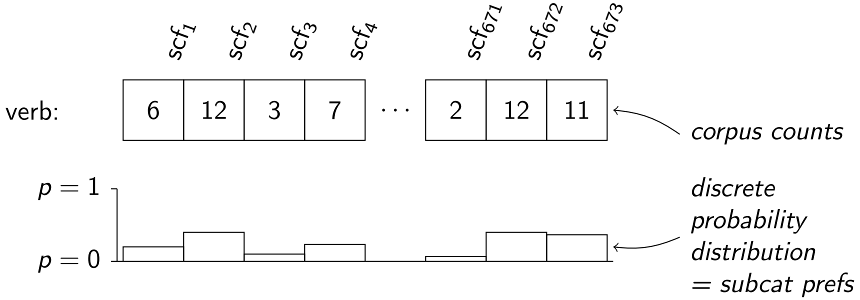
- A verb's distribution over SCF codes is its *subcategorisation preference*

Pipeline



- Test set has 3 million verb instances
- Gold standard: 168 verbs in 43 classes

Verb clustering



- Verb dissimilarity is computed with the Jensen-Shannon divergence

Lexical preferences (LP)

Example

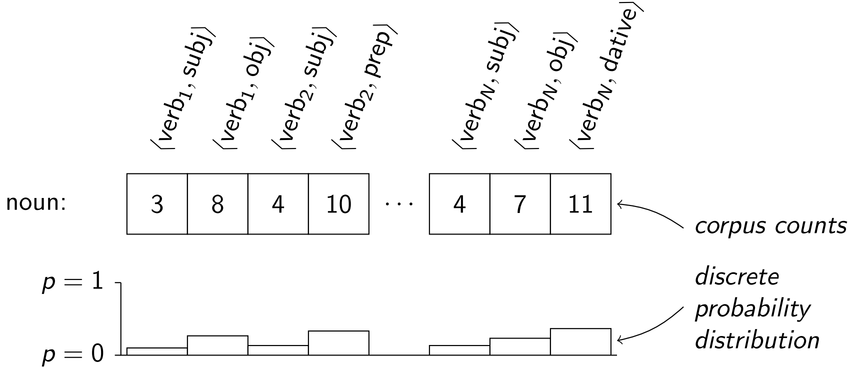
Wir	**benutzen**	Ihre Umfragedaten	nicht	für eigene Zwecke.
We	use	your survey data	not	for own purposes.

We will not use your survey responses for private purposes.

benutzen $\Rightarrow$ `nap:für.Acc*dobj-Umfragedatum*prep-Zweck`

- To control data sparsity, we employ a parameter N: number of nouns included in the lexical preferences model
  - Nouns with rank $> N$ are ignored (as if unseen)

Sun/Korhonen



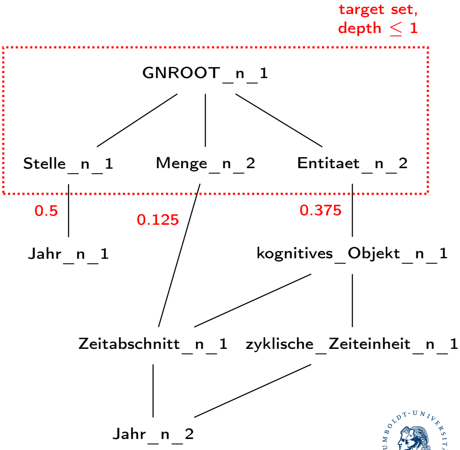
- Partition N nouns into M classes (equivalence relation)

Word space model (WSM)

- Built on lemmatised SdeWaC
- Features are the 50,000 most common words (minus stop words)
- Sentences as windows
- Feature weighting: t-test scheme
- *Context selection* zeroes out infrequent features in the model
- Use cosine similarity and spectral clustering to partition N nouns into M classes

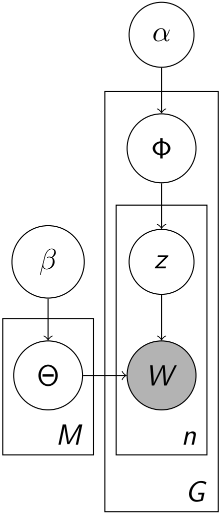
GermaNet

- Granularity is controlled using *depth*, *d*
- Nouns can belong to more than one concept: *soft clustering*



Latent Dirichlet Allocation (LDA)

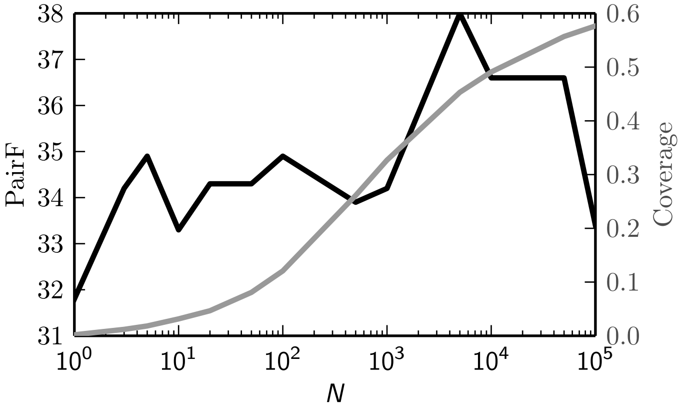
- Built with the same data used by the Sun/Korhonen model
- Each $\langle \text{verb}, \text{grammatical relation} \rangle$ pair has a distribution Φ over concepts
- Each concept z has a distribution Θ over the N nouns
- Number of concepts M is 50 or 100



Results

SP model	Parameters	Granularity	F-score
SUN	10K nouns	1,000 noun classes	39.76
LDA (hard)	10K nouns	50 topics	39.09
LP	5K nouns		38.02
WSM	10K nouns	500 noun classes	36.92
LDA (soft)	10K nouns	50 topics	35.91
GermaNet	depth = 5	8,196 synsets	34.41
Baseline			33.47

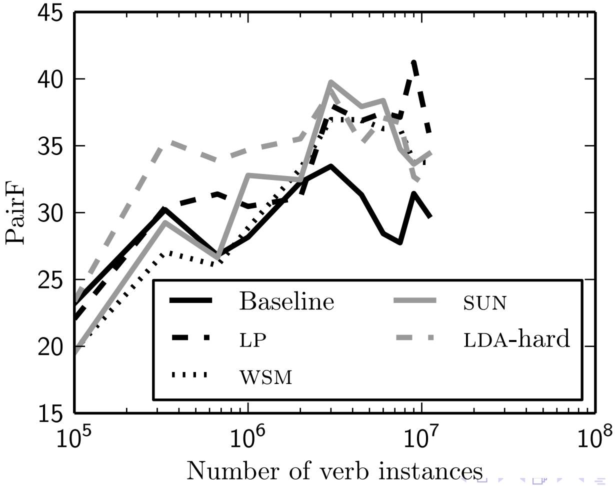
Sparsity effects in LP



Qualitative differences in noun partitions

SUN	WSM
F-score 39.76	*F*-score 36.92
syntagmatic information	paradigmatic information
synonym/co-hyponym structure	thematic structure
class size variance 37	class size variance 2800
semantically consistent	large classes inconsistent

Test set size



Conclusions

1. Selectional preferences help automatic verb classification
2. Optimal concept granularity is relatively fine
   - Lexical preferences works very well if it is properly tuned
   - Classification of proper names is useful: given names, corporations, medications, etc.
3. Syntagmatic information works better than paradigmatic

Summary

- Selectional preference models have been compared before
  - Almost always under a plausibility or pseudoword paradigm!
- We are interested in semantic verb clustering
- We evaluate several selectional preference models, comparing them using a manually constructed semantic verb classification

- We show that modelling selectional preferences is beneficial for verb clustering, no matter which selectional preference model we choose
- Other findings:
  - Capturing syntagmatic relations seems to work better than paradigmatic
  - A simple lexical preferences model performs very well; data sparsity does not seem to be more of a problem for this model than for others

References

James D. McCawley. The role of semantics in a grammar. In Emmon Bach and Robert Harms, editors, *Universals in Linguistic Theory*, pages 124–169. Holt, Rinehart and Winston, 1968.