seminar für sprachwissenschaft

Exploring the Data-driven Prediction of Prepositions in English

Anas Elghafari, Detmar Meurers, Holger Wunsch
University of Tübingen

Eberhard Karls Universität Tübingen

Aims

Current Aims

- **Predict prepositions** in native English text.
- For predictions, use a **purely surface-based**, "web-as-corpus" approach (e.g., Lapata & Keller, 2005).
- Determine the tradeoff between
  - **Informativeness** of context (→ the size of the context n-gram)
  - **Data-sparseness** (→ can we find the n-gram on the web?)

Mid-Term Aims

- **Correct prepositions** in text written by **learners** of English.

Related Work

Correct use of prepositions is a major problem for learners of English.

→ The computational analysis of preposition usage has attracted significant attention in recent years in the field of **Intelligent Computer Assisted Language Learning**.

(De Felice & Pulman, 2009; De Felice, 2008; Lee & Knutsson, 2008; Gamon et al., 2008; Chodorow et al., 2007; Tetreault & Chodorow, 2008a,b)

Most of the approaches first tackle the task of predicting prepositions in *native language* as a point of reference.

- The task is usually approached as a classification problem:
  - The classes are the prepositions.
  - The instances to be classified are the contexts.
  - Contexts are represented by a *rich set* of linguistic features (POS tags, PP attachment sites, WordNet classes of PP object and modified item, etc.).

Our Approach

Using **purely surface-based predictions**, how much information can we find in the **immediate distributional context of the preposition**?

Data

- Random sample of section J of the BNC XML Edition (8060 sentences)
- We extract one **prediction task** for each occurrence of one of the top nine prepositions in the BNC *of, to, in, for, on, with, at, by, from*

→ Same corpus section and prepositions as used by De Felice (2008).

Prediction Tasks

Source Sentence

But **for** the young, it is rather a question **of** the scales falling **from** their eyes, and having nothing to believe **in** any more.

Prediction Task

```
But for the young, it is rather a question
-*-MASKED-*-of the scales falling from their
eyes, and having nothing to believe in any more.
```

Cohort

```
1. rather a question of the scales falling
2. rather a question to the scales falling
3. rather a question in the scales falling
...
9. rather a question from the scales falling
```

Prediction of Prepositions Using Web Counts

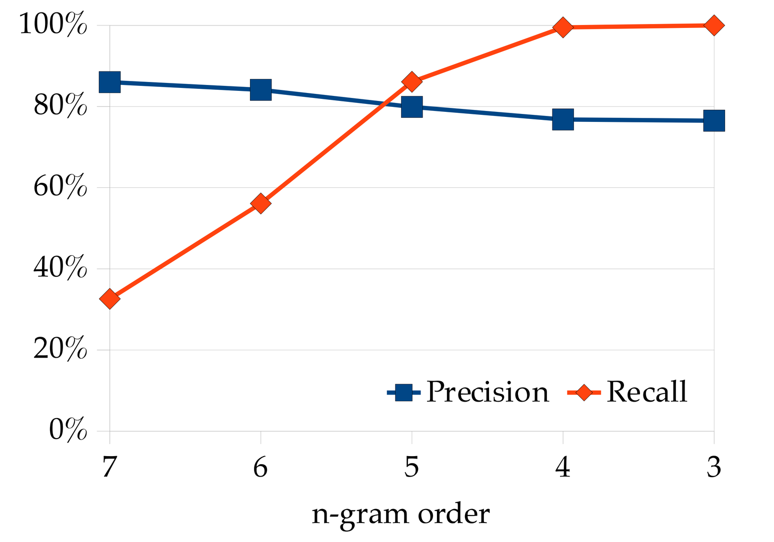
- Determine the **number of hits** for each 7-gram in a cohort using Yahoo's BOSS service.
- Pick the preposition for which the **most hits** were found.
- In case of **no hits**, back off to **lower n-gram orders using overlap backoff**.

Results

Summary

Approach	Accuracy
Gamon et al. (2008)	62.32%
Tetreault and Chodorow (2008a)	79.00%
Bergsma et al. (2009)	75.50%
De Felice (2008) system	70.06%
Majority baseline (*of*)	26.94%
Human agreement	88.60%
This paper – surface-based approach	**76.50%**

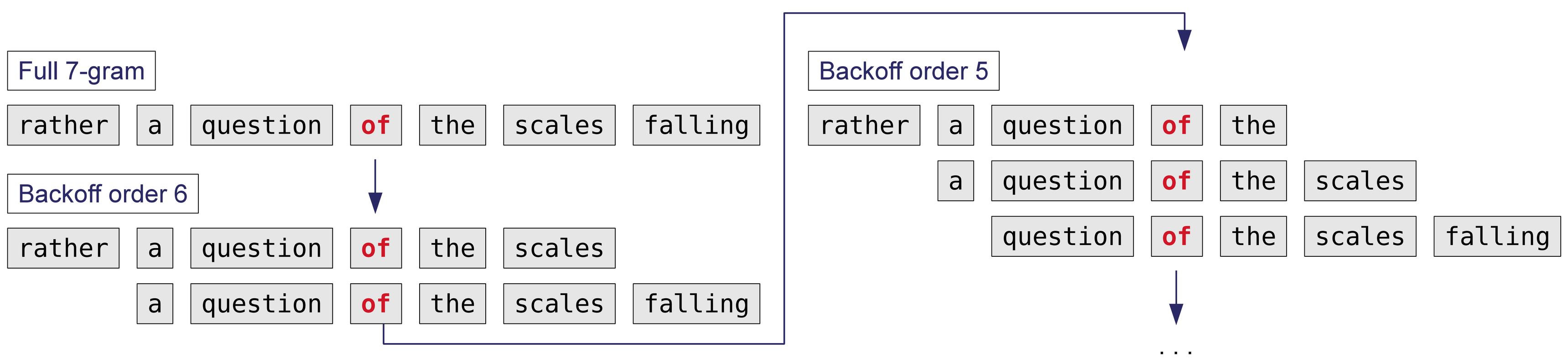
Informativeness/Data Tradeoff



Detailed Results

	Sum	7-grams (3 + prep + 3)	6-grams (truncated 7-gram)	5-grams (truncated 7-gram)	4-grams (truncated 7-gram)
Total	8060	6999	656	182	223
Predictions	2900	2195	379	119	207
correct	*2495*	*1931*	*326*	*91*	*147*
incorrect	*405*	*264*	*53*	*28*	*60*
Requiring back-off	5160	4804	277	63	16
Precision	86%	88%	86%	76.5%	71%
Recall	32.6%	28.7%	79.6%	59.1%	90.2%
		Back-off order 6			
Predictions	2028	2028			
correct	*1620*	*1620*			
incorrect	*408*	*408*			
Still requiring back-off	2776	2776			
Predict. orders 7+6	4223	4223			
correct	*3551*	*3551*			
incorrect	*672*	*672*			
Precision	84.1%	84.1%			
Recall	56.1%	56.1%			
		Back-off order 5			
Predictions	2180	2020	160		
correct	*1542*	*1411*	*131*		
incorrect	*638*	*609*	*29*		
Still requiring back-off	873	756	117		
Predict. orders 7 – 5	6782	6243	539		
correct	*5419*	*4962*	*457*		
incorrect	*1363*	*1281*	*82*		
Precision	79.9%	79.5%	84.8%		
Recall	86.1%	86.8%	79.6%		
		Back-off order 4			
Predictions	905	743	106	56	
correct	*488*	*382*	*68*	*38*	
incorrect	*417*	*361*	*38*	*18*	
Still requiring back-off	31	13	11	7	
Predict. orders 7 – 4	7806	6986	645	175	
correct	*5998*	*5344*	*525*	*129*	
incorrect	*1808*	*1642*	*120*	*46*	
Precision	76.8%	76.5%	81.4%	73.7%	
Recall	99.5%	99.8%	97.9%	94.9%	
		Back-off order 3			
Predictions	47	13	11	7	16
correct	*21*	*5*	*7*	*3*	*6*
incorrect	*26*	*8*	*4*	*4*	*10*
Still requiring back-off	0	0	0	0	0
Predict. orders 7 – 3	**8060**	6999	656	182	223
correct	***6166***	*5349*	*532*	*132*	*153*
incorrect	***1894***	*1650*	*124*	*50*	*70*
Precision	**76.5%**	76.4%	81.1%	72.5%	68.6%
Recall	**100%**	100%	100%	100%	100%

Overlap Backoff



Future Work

- Use Google's Web 1T 5-gram corpus data source instead of web counts.
- Explore backoff strategies based on linguistic generalizations.
- Vary the size of the context window based on **linguistic information**.
- Apply to **learner language** to identify and correct incorrect usage.

References

Chodorow, M., J. Tetreault & N.-R. Han (2007). Detection of Grammatical Errors Involving Prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*. Prague, Czech Republic: Association for Computational Linguistics.

De Felice, R. (2008). Automatic Error Detection in Non-native English. Ph.D. thesis, St Catherine's College, University of Oxford.

De Felice, R. & S. Pulman (2009). Automatic Detection of Preposition Errors in Learner Writing. *CALICO* 26(3).

Gamon, M., J. Gao, C. Brockett, A. Klementiev, W. Dolan, D. Belenko & L. Vanderwende (2008). Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. In *Proceedings of IJCNLP*. Hyderabad.

Lapata, M. & F. Keller (2005). Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing* 2(1).

Lee, J. & O. Knutsson (2008). The Role of PP Attachment in Preposition Generation. In A. Gelbukh (ed.), *Proceedings of CICLing 2008*. Haifa.

Tetreault, J. & M. Chodorow (2008a). Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection. In *Proceedings of COLING-08*. Manchester.

Tetreault, J. & M. Chodorow (2008b). The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of COLING-08*.