UniMelb Site Report

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Overview

- Melbourne is a little town on a small island far, far away from anywhere that it takes days to travel to in big mechanical birds ...

Overview

- "Phenomenon corpus", determination of the constructional coverage of a given grammar, identification of relevant sections of grammar files that relate to a given phenomenon (Ned)
- Harnessing the lexical type hierarchy in supertagging (Andrew Chester)
- Social media text analytics (Tim, Andy MacK)

Supertagging with Hierarchical Tagset: Background

- There has been plenty of work on supertagging for various reasons (robustness, efficiency, ...), but the standard assumption has been that that tagset is "flat"
- With HPSGs, there is, of course, lots of structure to the supertags (= lexical types) that, intuitively, it would appear we should be able to use to good effect
- **Research question:** can we improve the accuracy of supertagging via cleverer user of the type hierarchy?

Supertagging with Hierarchical Tagset: Basic Approach

- Extract type hierarchy from a grammar (focusing exclusively on the ERG for now), and Rebecca-style supertag data
- Also experiment with Penn POS tagging, and shallow-fragmented hierarchy defined by Penn POS tags
- In the first instance focus on supervised learning
- Evaluate in terms of both supertagger accuracy and (ultimately) the impact on parse selection accuracy
- Method 1: supertag "backoff" using the type hierarchy and trigram HMM (interpolate up the type hierarchy)

Supertagging with Hierarchical Tagset: Methodology

- Current method: trigram HMM with class “backoff” using the type hierarchy (interpolate transition probabilities up the type hierarchy)
  - different smoothing methods
  - different context sizes
  - different levels of class backoff

Supertagging with Hierarchical Tagset: Open Questions

- How far up the type hierarchy/how aggressively should we be backing off?
- Is all of the type hierarchy “fair game” for class smoothing?
- Do all classes equally require class smoothing, or should we be adapting a dynamic smoothing approach?
- Does class smoothing improve the quality of the sequence probabilities/ranking of tag sequences any?

Supertagging with Hierarchical Tagset: Other Ideas

- Pseudo-likelihood?
- Hierarchical HMMs?
- Same basic approach with MEMMs/CRFs?
- Ultimately interested in moving to unsupervised learning, but want to "concept-prove" in a supervised context first

Social Media Analytics

- Ultimately interested in (very) robust “semantic parsing” of social media text
- Some preliminary work on applying the ERG to social media text from different sources, to do Beauty and the Beast-style profiling of the parsing difficulty of different social media sources (to appear at IJCNLP)

How Noisy Social Media Text?

1. Collect (English) text data from a variety of social media sources (Twitter [×2], YouTube comments, web user forums, blogs, Wikipedia, in addition to BNC)
2. Language-filter, sentence tokenise, and strip meta-linguistic tokens (e.g. hashtags and mentions) based on Twitter-POS tagger
3. Parse the resultant sentences with the ERG v1111 (with robustness rules turned on, using unknown word handling based on Twitter-POS tags and generic lexical types, and with re-tokenisation)

“Parsability” Results

Corpus	Parseable				Unparseable
	strict		informal		
	full	frag	full	frag	
TWITTER-1	13.8	23.9	22.2	2.5	37.4
TWITTER-2	13.9	23.8	22.8	1.7	37.6
COMMENTS	18.0	22.2	26.4	1.4	31.9
FORUMS	23.9	14.1	24.7	1.5	35.6
BLOGS	25.6	17.5	18.8	2.7	35.3
WIKIPEDIA	48.7	4.5	18.9	1.5	26.2
BNC	38.4	12.0	24.0	2.2	23.2

Causes of Parse Failure

Corpus	Frag.	Pre-proc error	Res. limit	Ungram. inputs	Extra-gram.	Grammar gaps
Twitter-1	0.16	0.24	0.00	0.32	0.09	0.18
Twitter-2	0.19	0.22	0.00	0.31	0.10	0.17
Comments	0.13	0.32	0.00	0.31	0.04	0.20
Forums	0.05	0.31	0.01	0.36	0.03	0.24
Blogs	0.09	0.22	0.11	0.11	0.22	0.25
Wikipedia	0.08	0.11	0.10	0.06	0.06	0.59
BNC	0.15	0.05	0.15	0.04	0.05	0.56

Summary

- "Phenomenon corpus", determination of the constructional coverage of a given grammar, identification of relevant sections of grammar files that relate to a given phenomenon
- Harnessing the lexical type hierarchy in supertagging
- Social media text analytics