Parsing with Subdomain Instance Weighting from Raw Corpora

Barbara Plank[1], Khalil Sima'an[2]

[1]University of Groningen, The Netherlands [2]University of Amsterdam, The Netherlands

1. Introduction

Statistical Parsing

- Problem: Ambiguity of Natural Language
- Common approach: Train a parser/model on a hand-parsed treebank, e.g. Penn Wall Street Journal (WSJ)
- Variations: phrase/dependency structure, formal grammar, statistical model and estimator.
- Generative Model: Joint probability P(t,s); t parse tree, s sentence; Markovian process builds <t,s>
- Parse Reranking: Generative Models currently complemented with discriminative rerankers (Charniak & Johnson,2005)

2. Motivation

- Is there more in the treebank that we might exploit?
- View treebank as a mixture of subdomains
  "politics, stock market, financial news etc. can be found in the WSJ" (Kneser and Peters, 1997)
- Implication: Statistics gathered by generative statistical parsers are averages over sentence-tree pairs; may mix 'unrelated' sentence-parse pairs, i.e. different subdomains
- Averages smooth out the differences between subdomains and weaken the biases in the model
- **Research question:** Is there any gain to be had from incorporating subdomain statistics into parse re-ranking?

3. Our Approach

- Exploit unannotated, subdomain specific corpora gathered from the web to weight original treebank trees to reflect subdomain statistics: "Subdomain Instance Weighting"
- Can be seen as instantiation of "Instance Weighting" [3], surfaced in adaptation, but not tested before in parsing

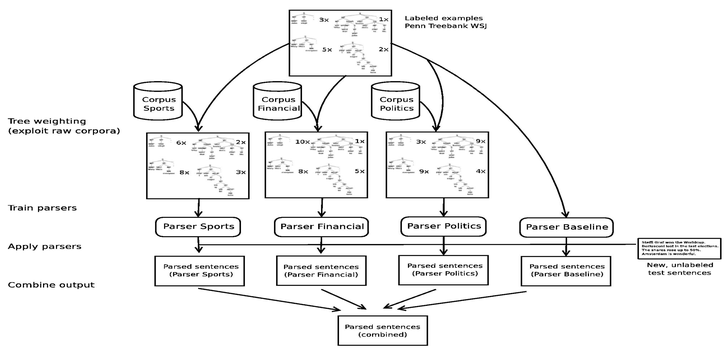


Figure: Summary of Subdomain Instance Weighting from Raw Corpora

4. Subdomain Instance Weighting

- Suppose we know of a subdomain **d** of the WSJ domain **w**
- We would like to scale counts of parses in the WSJ such that those more similar to parses in **d** get higher counts than others
- Had we had access to a subdomain **d** treebank, we could train parser parameters π by maximum-likelihood training:

$$\arg\max_{\pi} \sum_{\langle s,t\rangle \in \mathbf{d}} -P_{\mathbf{d}}(s,t) \log P(s,t;\pi) \qquad (1)$$

- But we do not have $P_{\mathbf{d}}(s,t)$ (subdomain treebanks), and cannot train the parameters π on complete data
- However, given a sentence s, we assume conditional parse probabilities do not change much, i.e. $P_{\mathbf{d}}(t|s) \approx P_{\mathbf{w}}(t|s)$, while $P_{\mathbf{d}}(s)$ deviates from $P_{\mathbf{w}}(s)$.
- Thus: exploit the 'domain difference' [3] and scale the WSJ parses $< s,t >$ by a ratio $\frac{P_{\mathbf{d}}(s)}{P_{\mathbf{w}}(s)}$

Instantiation of Subdomain Instance Weighting:

- We use estimates of n-gram language models trained over raw subdomain data to approximate $P_x(s)$
- Let $D(s) \doteq \log P_{\mathbf{w}}(s) - \log P_{\mathbf{d}}(s)$
- For all $\langle s,t\rangle \in \mathbf{w}$, we scale parse counts in the treebank **w** by multiplying them with $C_{\mathbf{d}}(s,t)$:

$$C_{\mathbf{d}}(s,t) = \begin{cases} \alpha \times D(s) + \beta & D(s) > 0 \\ 1 & otherwise \end{cases} \qquad (2)$$

Where $\alpha = 5$ and $\beta = 500$ set empirically on the dev set data.

5. Experiments and Results

Data and Tools:

- Charniak's first-stage generative parser [1]
- Penn Treebank WSJ corpus, standard division (02-21 train)
- Subdomains in WSJ: Financial, Politics, Sports
- Raw corpora: Wikipedia (Financial, Sports), Europarl (Politics)
- SRILM Language Modelling Toolkit to estimate LMs

Subdomain parsers:

- Retrain Charniak parser on subdomain instance weighted TB (not using any heldout set)
- Subdomain parser far less accurate, as expected

Parser	length$\leq$ 40	length$\leq$ 100
WSJ	90.82	89.87
POLITICS	84.75	82.19
FINANCIAL	84.98	82.73
SPORTS	85.30	83.22

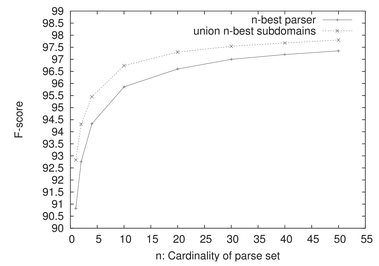
Table: F-scores for subdomain parsers on WSJ section 00

6. Variance: Are subdomain parsers complementary?

Union of n-best lists:

- Oracle F-scores for n-best WSJ parser vs. Union output of four subdomain parsers
- The 50-best union achieves %17 error reduction over 50-WSJ!

n	1	4	10	20	30	40	50
	F-score length $\leq$ 40						
WSJ	90.82	94.33	95.86	96.60	97.00	97.20	97.35
Union	92.83	95.45	96.74	97.30	97.54	97.68	97.80
	F-score length $\leq$ 100						
WSJ	89.87	93.29	94.91	95.70	96.09	96.32	96.50
Union	91.78	94.43	95.77	96.36	96.61	96.79	96.94



7. Is subdomain probability a discriminative feature?

Parse probability for n-best:

- Select the n parses with the highest probabilities given by any of the subdomain parsers
- With this simple selection procedure: up to 0.04 absolute F-score improvement over Charniak's n-best

Adding a single parse to n-best (n+1-best):

- F-score gain by adding the single most probable parse from the FINANCIAL parser to n-best:

n	1	4	10	20	30	40	50
Gain	+1.05	+0.29	+0.13	+0.11	+0.08	+0.08	+0.03

- The probability given by the subdomain parser has reasonable discriminative power (to be further explored)

8. Conclusions, Current & Future Work

- Empirical results warrant the conclusion that subdomain instance weighting is worthwhile further exploration
- Future work: exploring latent subdomains, reranking with new discriminative features from subdomain, examining approach for domain adaptation.

References

Eugene Charniak and Mark Johnson (2005).
Coarse-to-fine n-best parsing and maxent discriminative reranking.
In *Proceedings of ACL 2005*. The Association for Computer Linguistics.

R. Kneser and J. Peters (1997).
Semantic clustering for adaptive language modeling.
In *ICASSP 1997*, volume 02, page 779, Los Alamitos, CA, USA. IEEE Computer Society.

Jing Jiang and ChengXiang Zhai (2007).
Instance weighting for domain adaptation in NLP.
In *Proceedings of ACL 2007*, pages 264–271, Prague, Czech Republic, June. Association for Computational Linguistics.

Sekine, S. (1997).
The Domain Dependence of Parsing.
Washington, DC, USA. The Fifth Conference on Applied Natural Language Processing.

http://www.let.rug.nl/~bplank b.plank@rug.nl,k.simaan@uva.nl http://staff.science.uva.nl/~simaan