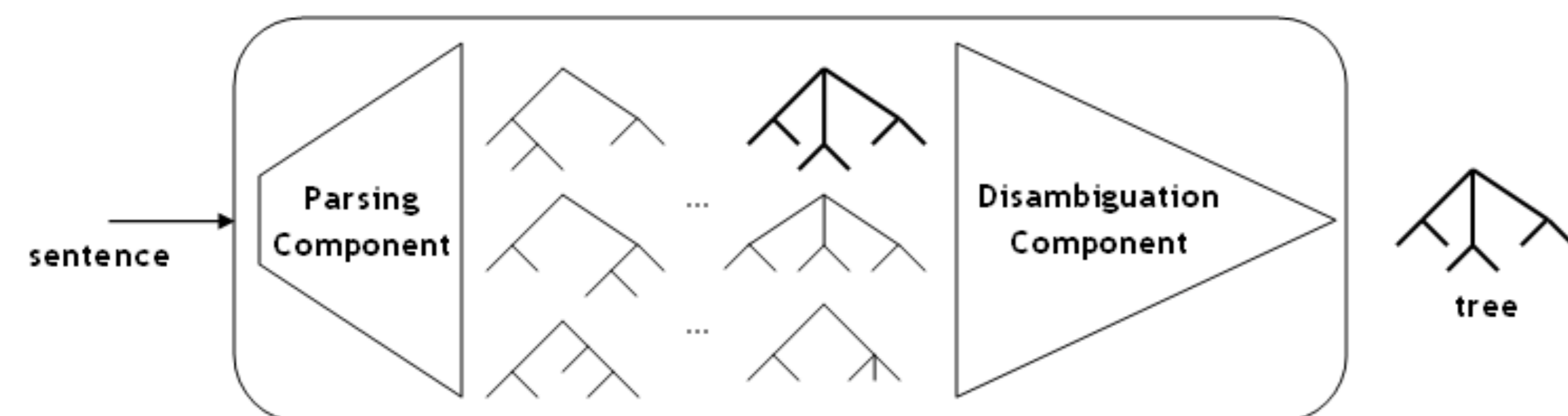


Introduction

Research area: Parsing of Natural Language

- ▶ Definition: Assigning syntactic structure to sentences
- ▶ Challenge: Ambiguity of Natural Language

A parser - Conceptual view



Disambiguation Component

- ▶ Selects the best parse from the (many) alternative hypotheses
- ▶ Statistical in nature; bases its decisions on a hand-parsed treebank → reflects statistical characteristics of the training data

The Problem: Domain Dependence of Parsing

Domains of Language Use

- ▶ Huge variation in vocabulary and style



Problem: Portability

- ▶ A disambiguation component will be successful as long as the treebank it was trained on resembles the input the model gets.
- ▶ Whenever training and test data differ considerably, the performance of such a supervised system *degrades* in an appalling way [Gildea, 2001] (e.g. PCFG parser trained on WSJ: from 89% down to 76% [Lease et al., 2005])

PCFG parsing / English	F-score
WSJ (newspaper)	89.5
Brown (fiction/non-fiction)	83.4
GENIA (biomedical)	76.3

Solution approaches:

1. Build a model for every domain we encounter. Need training data → expensive & unsatisfactory solution
2. Adapt parsers from a *source* domain (e.g., news) to a *target* domain (e.g., biomedical) → **Domain Adaptation**:
 - a. leverage a small amount of labeled target data or
 - b. use unlabeled (source and/or target) data for the adaptation task.

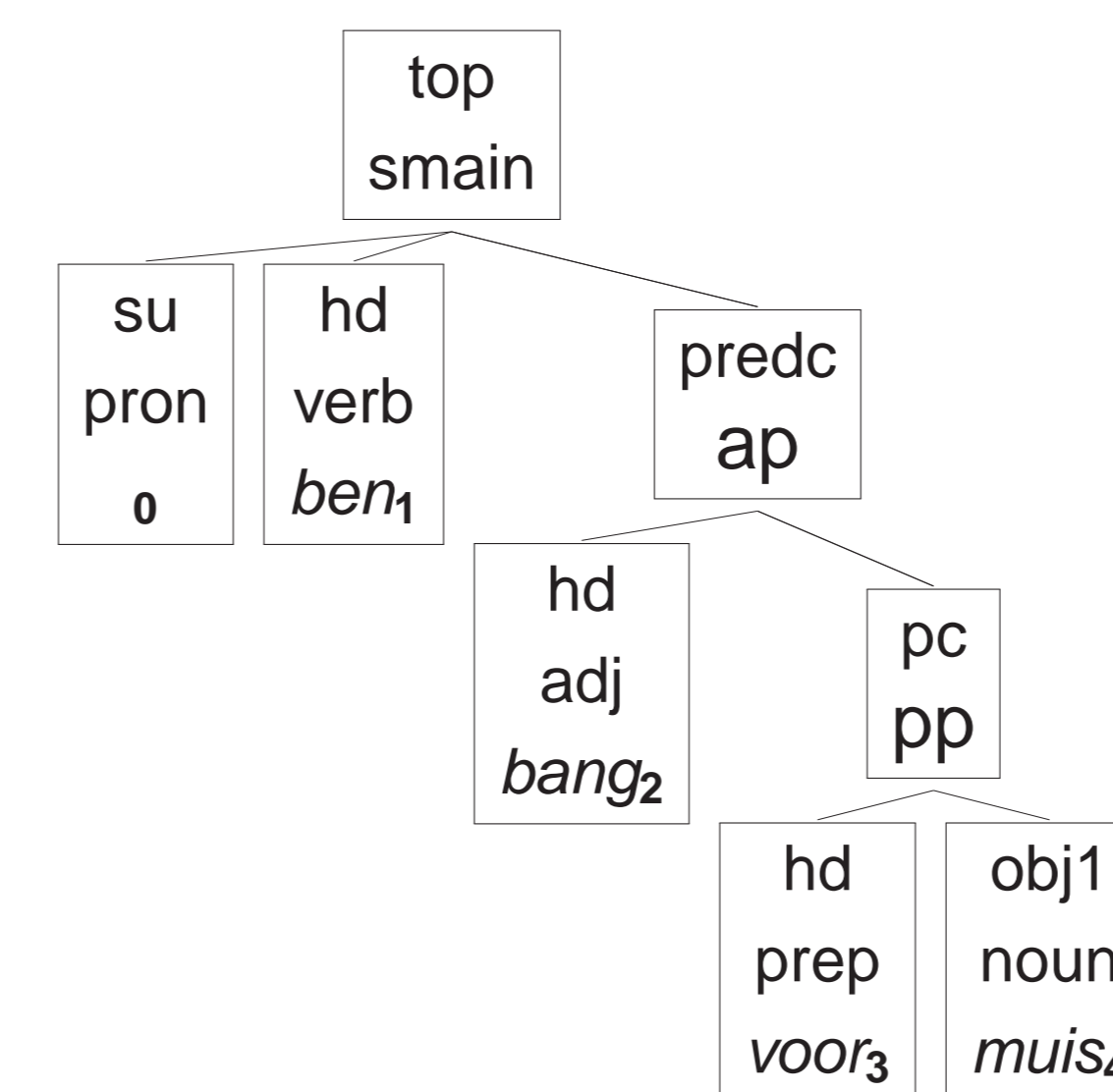
Background: Alpino Parser

- ▶ Wide-coverage dependency parser for Dutch
- ▶ HPSG-style grammar rules, large hand-crafted lexicon
- ▶ Disambiguation Component based on Maximum Entropy; a model is specified by a set of *feature functions* describing properties of the data; training means estimating their *weight*

$$P_{\theta}(\omega|s) = \frac{1}{Z_{\theta}} \exp \sum_{j=1}^m \theta_j f_j(\omega) \quad (1)$$

- ▶ Output: Dependency structure

(Example: 'Hij is bang voor muizen' - He is afraid of mice)



Initial experiment: Exploring auxiliary distributions

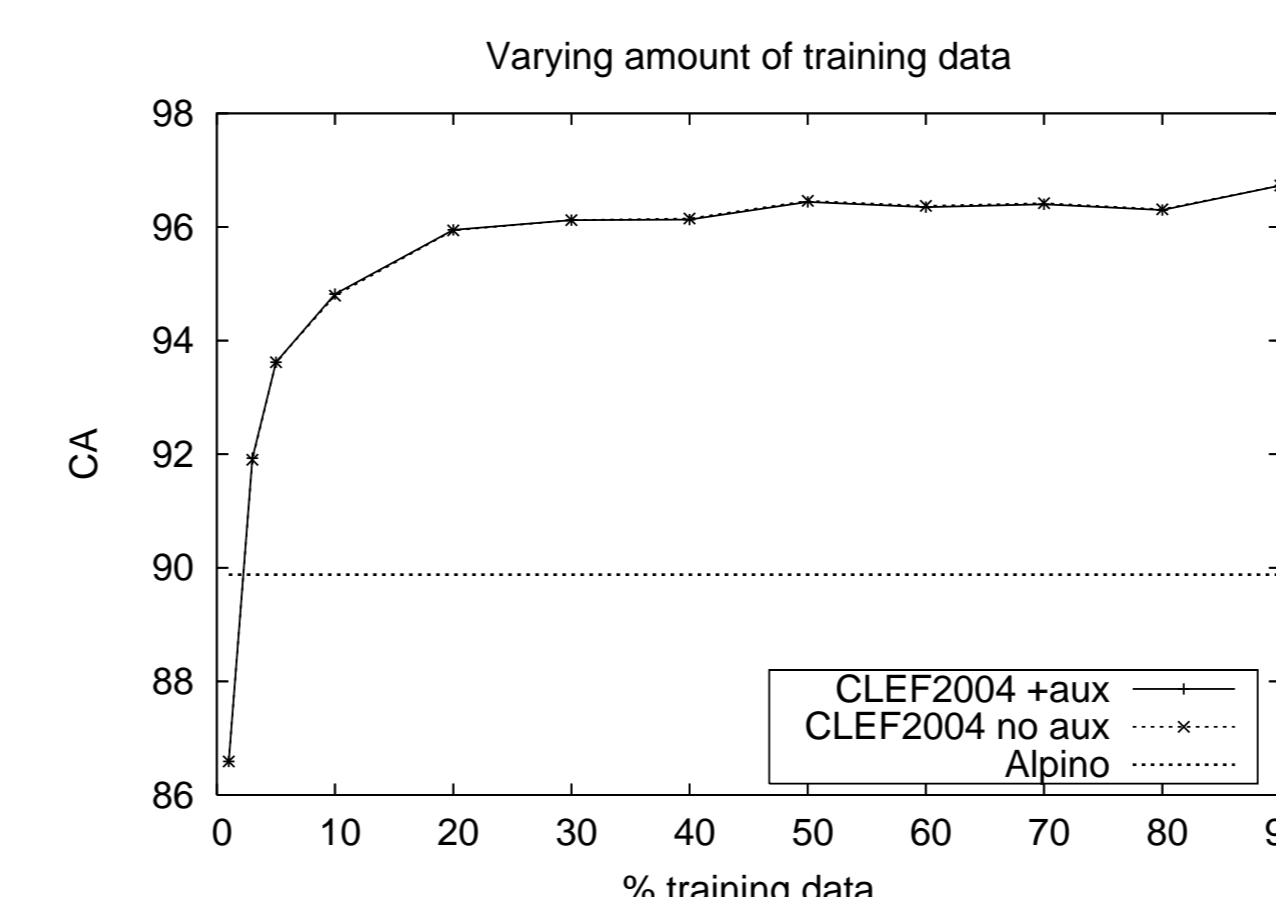
- ▶ Assume a limited amount of *target* training data
- ▶ Exploit an already trained out-of-domain/more general model to overcome the limited amount of target training data
- ▶ *Idea*: incorporate the more general model into the specific by exploiting auxiliary distributions [Johnson and Riezler, 2000]
 - ▶ In more detail: the logarithm of an auxiliary distribution is considered an additional, real-valued feature.
- ▶ We leverage the information from the general model Θ (the probability it assigns to a given parse) and thus add an auxiliary feature: $f_{m+1} = -\log P_{\Theta}(\omega|s)$

Data & Results:

- ▶ Source: General Alpino model (newspaper text)
- ▶ Target: CLEF (corpus of questions)

	CLEF	Alpino	Alpino +CLEF	CLEF +aux
CLEF 2003	97.01	94.02	97.21	97.01
CLEF 2004	96.60	89.88	95.14	96.60
CLEF 2005	97.65	87.98	93.62	97.72
CLEF 2006	97.06	88.92	95.16	97.06
CLEF 2007	96.20	92.48	97.30	96.33

Table: Results on the CLEF test data



Conclusions:

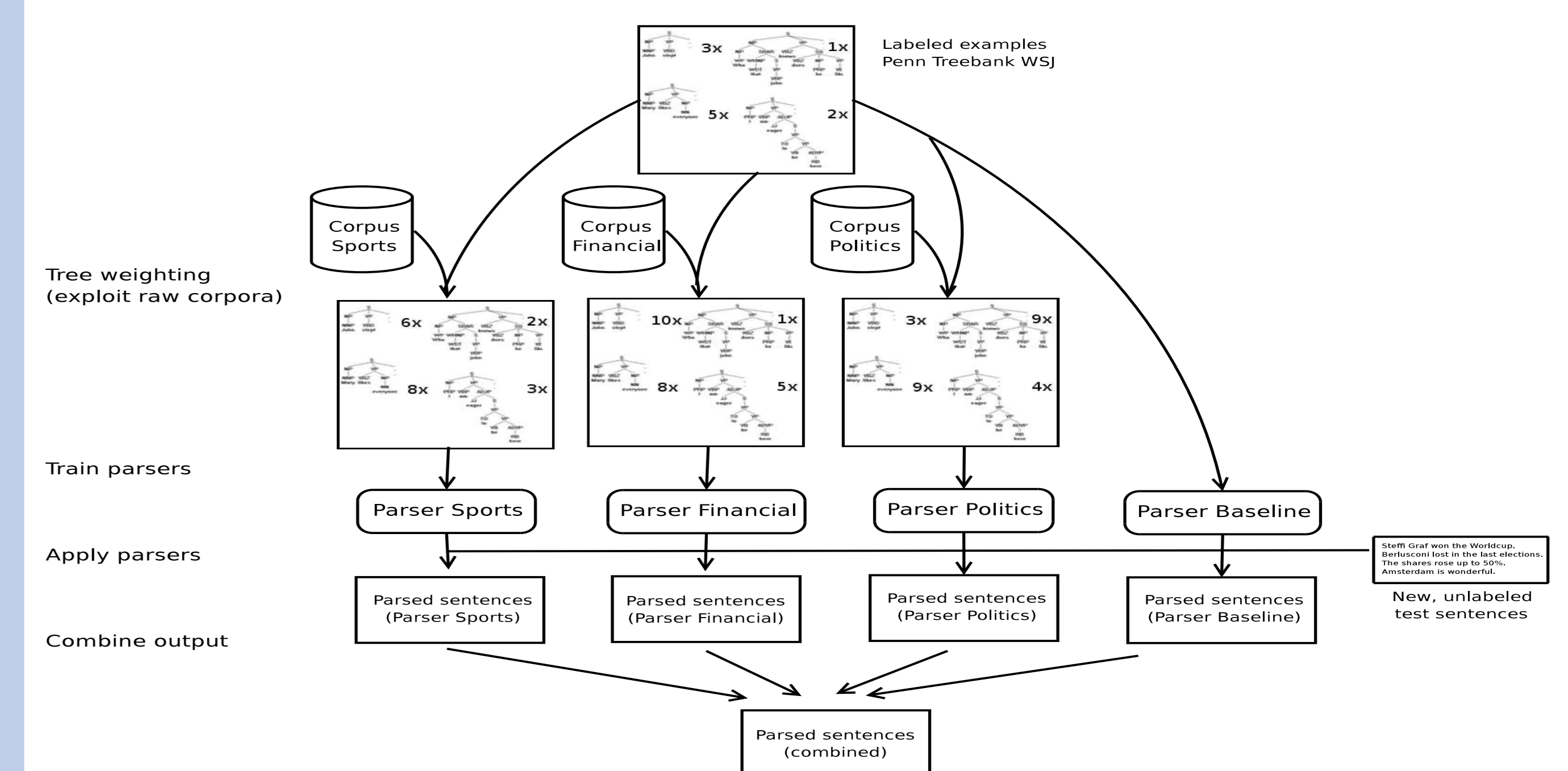
- ▶ Approach not suitable for leveraging limited training data
- ▶ Even on models trained on very little in-domain training data the approach does not work; instead hurts performance
- ▶ The auxiliary approach needs training data to estimate the weight(s) of the auxiliary feature(s).
 - ▶ little training data → weight cannot be estimated appropriately
 - ▶ more training data → just use that (less useful to incorporate the feature)

Conclusions, Current & Future Work

- ▶ Auxiliary based approach does not help for domain adaptation; as soon as there is labeled target training data just use it!
- ▶ Using unlabeled data: Structural Correspondence Learning (SCL) [Blitzer et al., 2006]
 - ▶ Finds correspondences between features of different domains
 - ▶ Has been successfully applied to PoS tagging and Sentiment analysis
 - ▶ We will examine SCL on dependency parsing for Dutch
- ▶ Exploit Wikipedia and its category system as data source
- ▶ Investigate the question: What precisely is meant by domain? E.g. Blogs → politics blog, tech blog, sports blog,...
- ▶ Other approaches/techniques for domain adaptation

Related Work - Parsing and Subdomains

- ▶ Subdomain Sensitive Statistical Parsing using Raw Corpora
- ▶ Existence of subdomains in WSJ (sports, financial, politics)
- ▶ Using domain dependent raw corpora to train subdomain PCFG parsers for the classical WSJ parsing task in English (by subdomain instance weighting)
- ▶ Approach has shown to have potential [LREC08]: F-score: 89.53 → 90.62 (89.62); Future work: extend it to *n*-best parsing



References

- 1 Daniel Gildea (2001). *Corpus variation and parser performance*. In 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- 2 Johnson and Riezler (2000). *Exploiting auxiliary distributions in stochastic unification-based grammars*. In Proceedings of the 1st NAACL, Seattle, WA.
- 3 Matthew Lease and Eugene Charniak. *Parsing Biomedical Literature*. In Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'05), Korea.
- 4 Gertjan van Noord (2007). *Using Self-Trained Bilexical Preferences to Improve Disambiguation Accuracy*. In Proceedings of IWPT 2007, Prague.
- 5 John Blitzer, Ryan McDonald, and Fernando Pereira (2006). *Domain Adaptation with Structural Correspondence Learning*. Empirical Methods in Natural Language Processing - EMNLP 2006.