Lexicalization in Crosslinguistic Probabilistic Parsing: The Case of French

Abhishek Arun and Frank Keller

June 24, 2005

School of informatics

Motivation

- Most statistical parsing models developed for English and trained on Penn Treebank (PTB).
- Broad coverage and High parsing accuracy (around 90% F-Score).
- Can these models generalize to :
  - Other languages e.g languages with different word order.
  - Other annotation schemes e.g flatter treebanks.
- What about French? Statistical parsing not been attempted before.

Typical Approaches to Statistical Parsing

- Lexicalised vs Unlexicalised PCFGs.

- For English, typically unlexicalised PCFGs perform poorly.

- Lexicalise the PCFG by associating a head word with each non-terminal in the parse tree.

- Currently, best results for PTB obtained by lexicalisation and markovization of rules.
  Collins (1997): LR 87.4% and LP 88.1%, Charniak (2000): LR and LP 90.1%

Previous Work

- German: Dubey and Keller (2003). Basic unlexicalised PCFG outperforms 2 different lexicalised models. (70.56% LR and 66.69% LP)
- Hypothesis: Lexicalised models failing due to
  - Flat structure of German treebank (Negra).
  - Flexible word order in German.
- Used sister-head dependency variant of Collins Model 1 to cope with flatness.
- Resulting model (71.32% LR and 70.93% LP).

Research question

- Dubey and Keller's (2003) work does not tell us whether flatness or word order flexibility is responsible for results.

	Annotation	Word Order	Lexicalization
German - Negra	Flat	Flexible	Does not help
English - PTB	Non-Flat	Non-Flexible	Helps
French - FTB	Flat	Non-Flexible	?

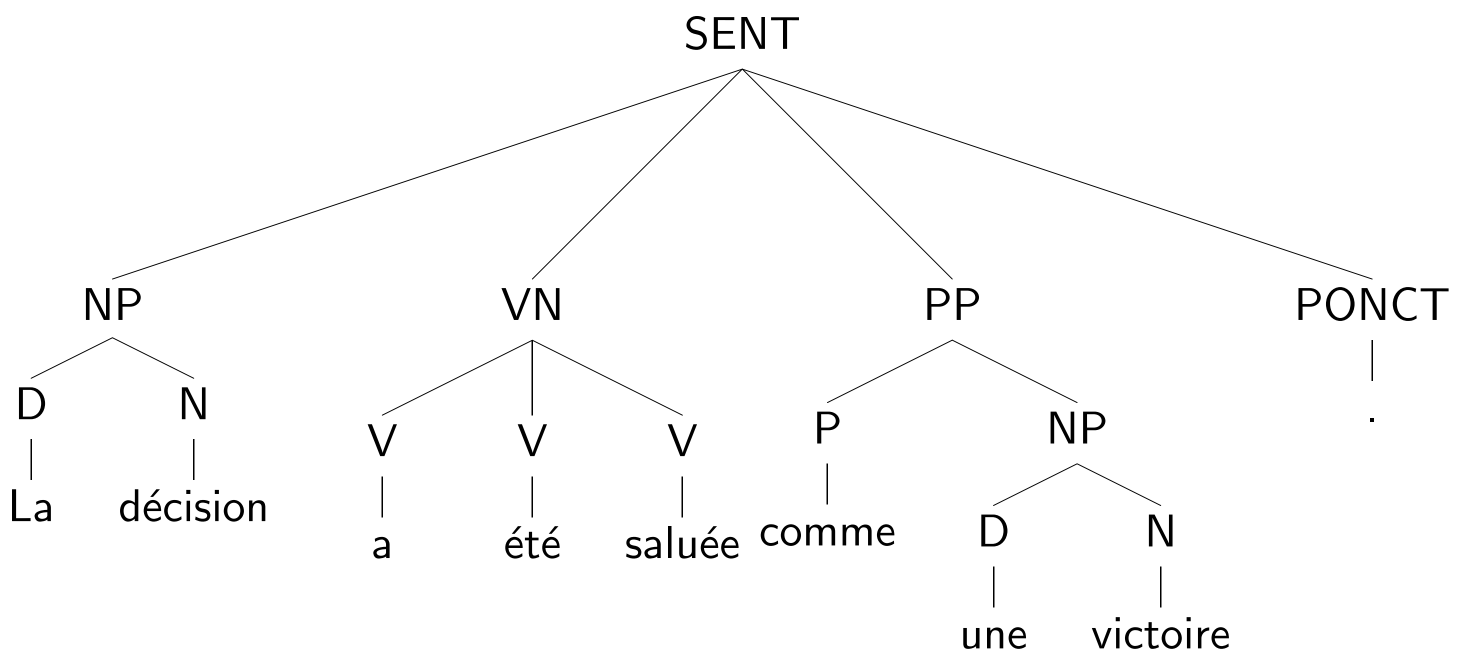
French Treebank - Corpus Le Monde

- French Treebank (FTB; Abeillé et al.2000) Version 1.4, released in May 2004.
- 20,648 sentences extracted from the daily newspaper *Le Monde*, covering a variety of authors and domains (economy, literature, politics, etc.)
- Each token is annotated with its POS tag, inflection (e.g. masculine singular), subcategorization (e.g. possessive or cardinal) and lemma (canonical form).

```
<AP>
  <w lemma="humain" ei="Amp" ee="A-qual-mp" cat="A"
     subcat="qual" mph="mp">humains</w>
</AP>
```

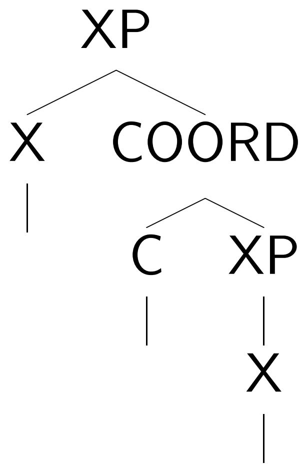
French Treebank - Corpus Le Monde

- No Verb Phrase: only the verbal nucleus (VN) is annotated. VN comprises of the verb and any clitics, auxiliaries, adverbs and negation associated with it.



French Treebank - Corpus Le Monde

- Flat noun phrases, similar to Penn Treebank.
- Coordinated phrases annotated with the syntactic tag COORD.



Dataset

Preprocessing of FTB:

- 38 tokens with missing tag information, 1 sentence with garbled annotation - sentences discarded.
- XML annotated data transformed to PTB-style bracketed expressions.
- Only POS tag kept, rest of morphological information discarded.
- Empty categories removed, punctuation marks assigned new POS tags based on PTB tagset.
- Resulting dataset of 20,609 sentences into into 90% training set, 5% development set and 5% test set.

Tree transformation

A series of tree transformations applied to deal with peculiarities of the FTB annotation scheme.

Compounds have internal structure in the FTB.

```
<w compound="yes" lemma="par ailleurs" ei="ADV" ee="ADV" cat="ADV">
   <w catint="P">par</w>
   <w catint="ADV">ailleurs</w>
</w>
```

Tree transformation

Two different data sets created by applying alternative tree transformations.

1. **Collapsing the compound**: concatenate compound parts, pick up POS tag supplied at the compound level.

   ```
   (ADV par_ailleurs)
   ```

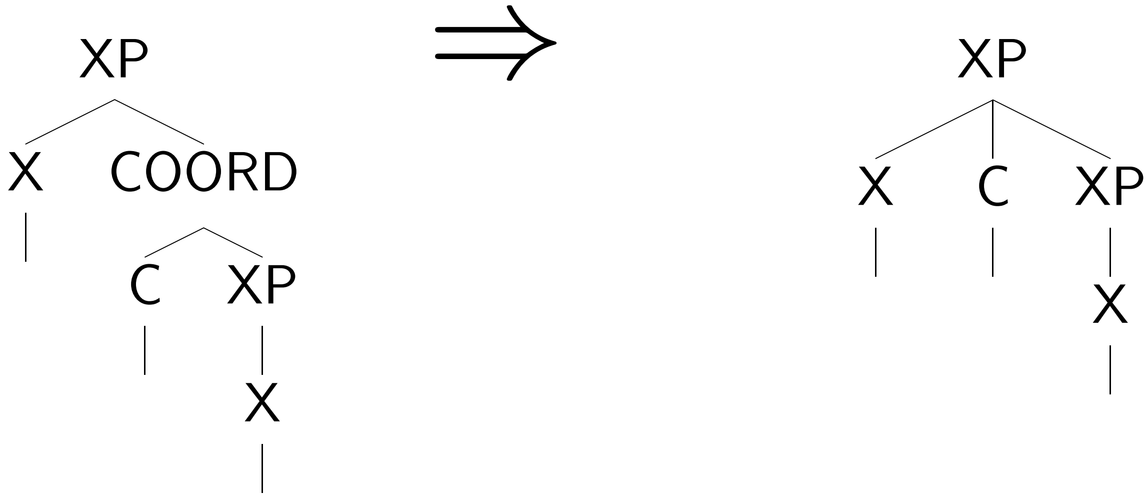
2. **Expanding the compound**: compound parts treated as individual words with own POS tags(from `catint` tag), suffix `Cmp` appended to POS tag of compound.

   ```
   (ADVCmp (P par) (ADV ailleurs))
   ```

Tree transformation

Collins' models, which we will use, have coordination-specific rules, presupposing coordination marked up in PTB format.

New datasets created where a *raising coordination* transformation is applied.



Baseline model - Unlexicalised Parsing - Results

- BitPar (Schmid, 2004): Bit-vector implementation of CKY algorithm.

For sentences of length $\leq$ 40 words.

	LR	**LP**	**CBs**	**OCB**	$\leq$ **2CB**
Expanded	58.38	58.99	2.31	30.00	62.89
Expanded + CR	59.14	59.42	2.25	31.32	64.34
Contracted	63.92	64.37	2.00	35.51	70.05
Contracted + CR	64.49	64.36	1.99	35.87	70.17

Findings

- Raising coordination transformation somewhat beneficial - increases LR and LP by around 0.5%; Contracting compound increases performance substantially - almost 5% increase in both LR and LP.

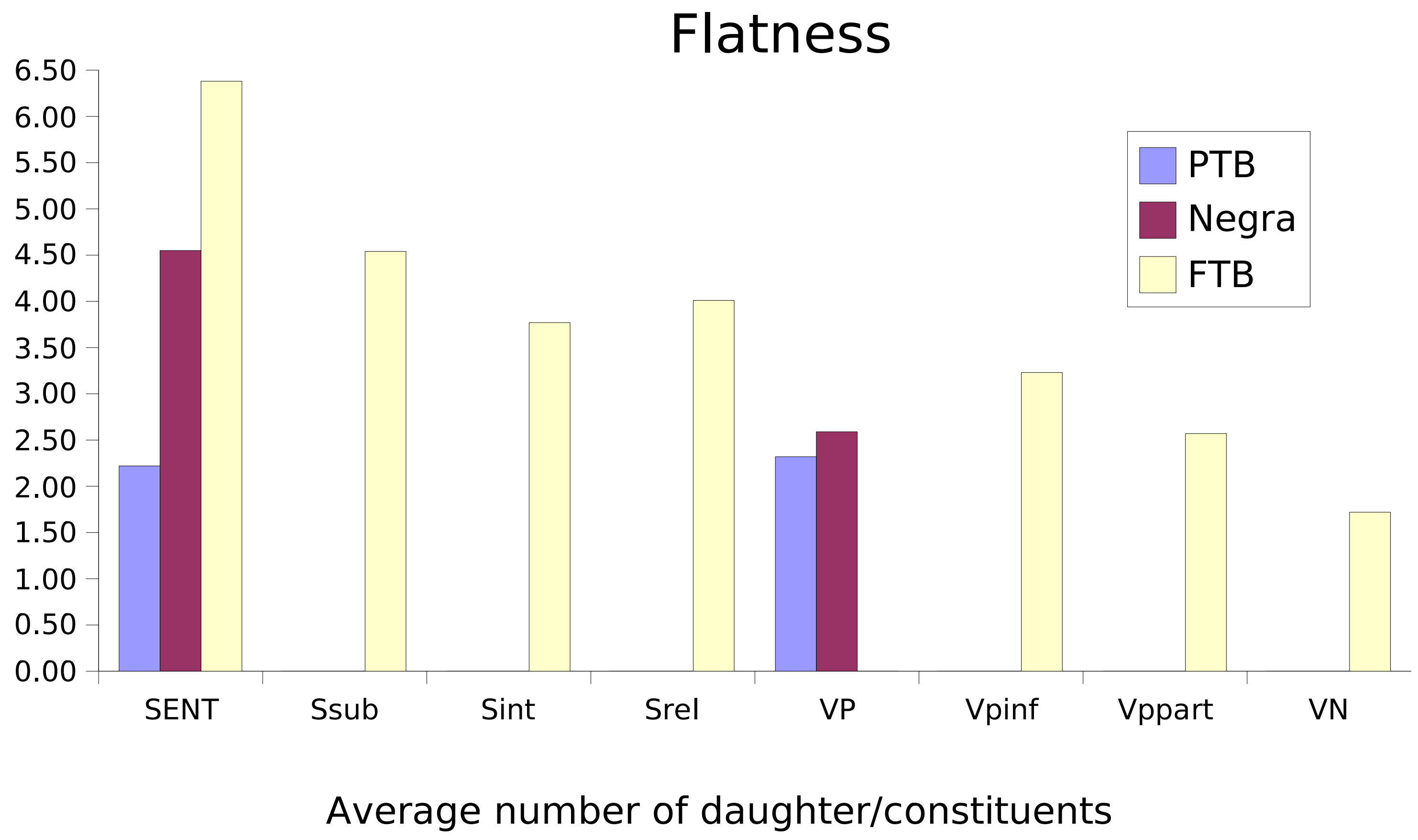
- However, the 2 different compound models do not yield comparable results - expanded compound has more brackets than contracted one.

Lexicalised Parsing models

Experiments run using Dan Bikel's parser (Bikel, 2002) which replicates Collins (97)'s head-lexicalised models, on CONT+CR dataset.

- Magerman style head-identification rules: FTB annotation guidelines and heuristics tuned on the development set.

- Complement/adjunct distinction for Model 2: argument identification rules tuned on dev set.



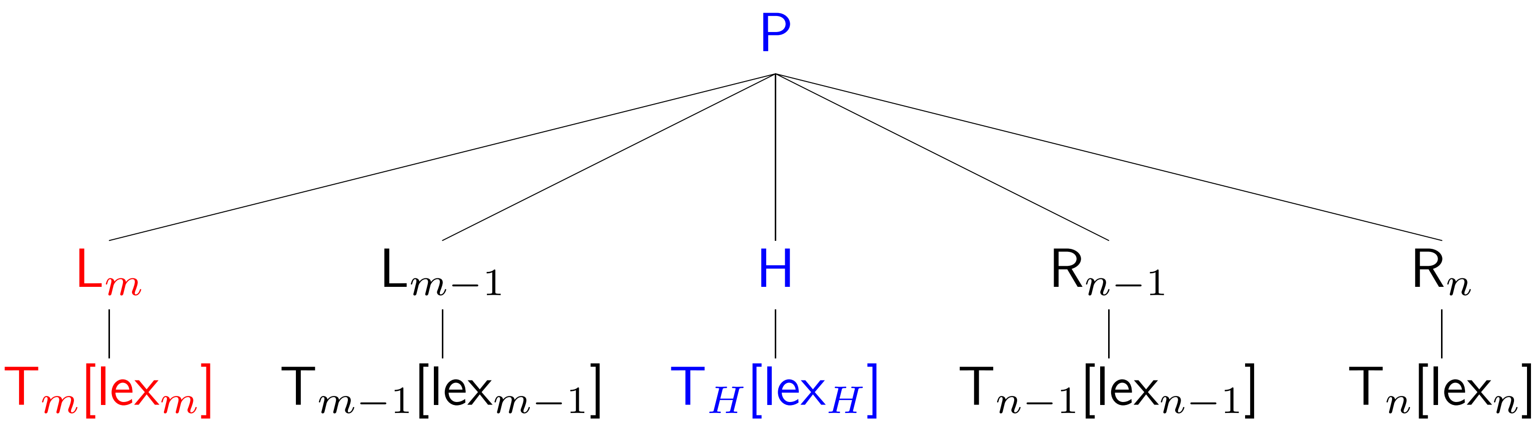
Strategy: Modify Collins model to deal with flat trees.

Modifying Collins' model

Standard modifier context: In the expansion probability for the rule:

$$P \rightarrow L_m \dots \; L_1 \; H \; R_1 \dots R_n$$

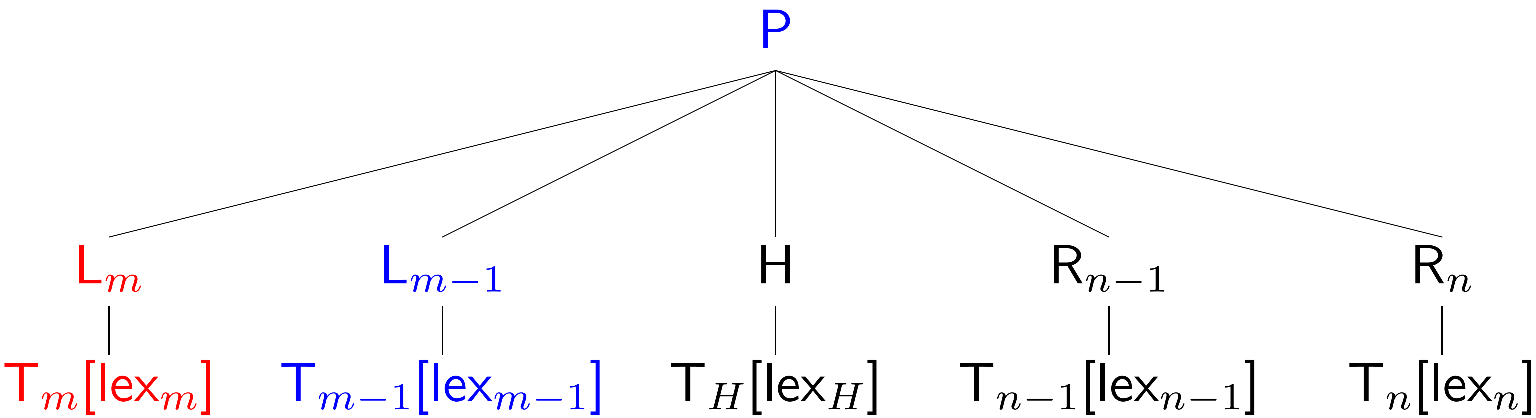
Modifier $\langle L_m, T_m, lex_m \rangle$ is conditioned on P and head $\langle H, T_H, lex_H \rangle$:



Modifying Collins' model

Sister-head model:

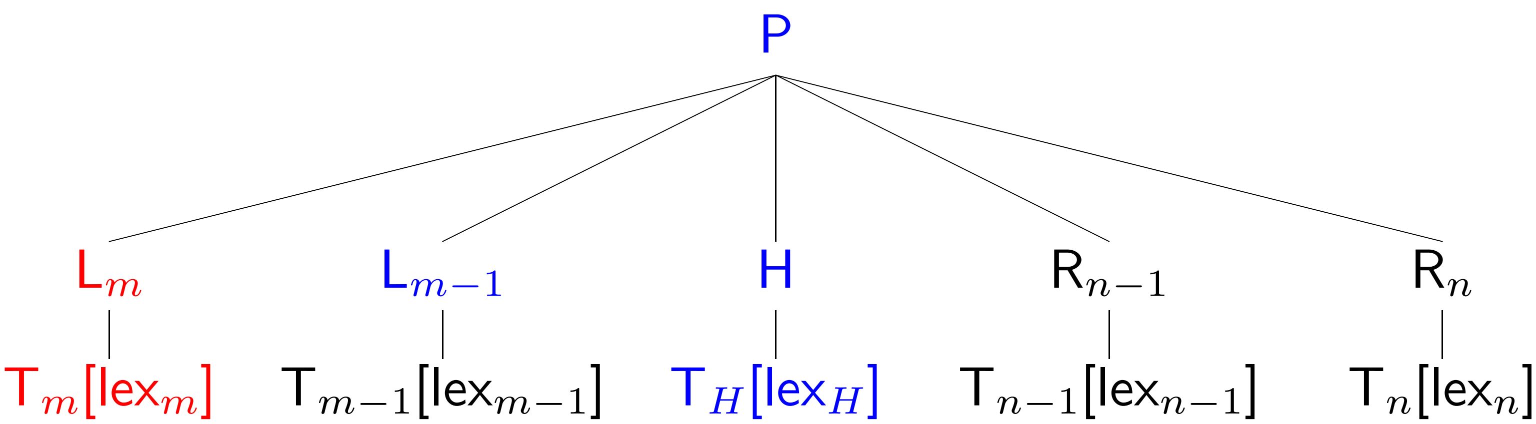
Modifier $\langle L_m, T_m, lex_m \rangle$ is conditioned on P and previous sister $\langle L_{m-1}, T_{m-1}, lex_{m-1} \rangle$:



Modifying Collins' model

Bigram model:

Modifier $\langle L_m, T_m, lex_m \rangle$ is conditioned on P, head $\langle H, T_H, lex_H \rangle$ and previous sister L_{m-1}:



Results

For sentences of length $\leq$ 40 words.

	LR	**LP**	**CBs**	**OCB**	$\leq$ **2CB**
Best unlex	64.49	64.36	1.99	35.87	70.17
Model 1	79.80	79.12	1.11	55.70	84.39
Model 2	79.94	79.36	1.09	56.02	83.86
SisterHead	77.68	76.62	1.26	51.70	81.31
Bigram	80.66	80.07	1.05	55.96	85.68
BigramFlat	80.65	80.25	1.04	56.85	85.58

Note: Bigram-flat model applies bigram model only to categories with high degrees of flatness (SENT, Srel, Ssub, Sint, VPinf and VPpart).

Lexicalised models - Results

Main Findings:

- Lexicalised models achieve performance almost 15% better than best unlexicalised model.
- Consistent with English parsing findings.
- Model 2 with complement/adjunct distinction and subcat frames, gives only slight improvement over model 1: FTB annotation scheme unsuitable?
- SisterHead performs poorly - maybe overfitting Negra?

Dependency Evaluation

Dependency evaluation argued to be more annotation-neutral than PARSEVAL, and less susceptible to cascading errors (Lin, 1995).

Model	Unlabeled Dependency	F-score
Cont+CR	75.20	64.42
Model2	85.20	79.65
SisterHead	83.33	77.15
Bigram	85.91	80.36
BigramFlat	85.75	80.45

- Dependency accuracies higher than constituency F-Scores across the board.
- Effect of lexicalization same on both measures.

Conclusions

- First probabilistic, treebank-trained parser for French.
- Unlexicalised baseline model achieved accuracy of about 64%.
- Lexicalised model, based on Collins Model 1 and 2, achieved accuracy of around 79%.
- Implementing a bigram model to account for flatness of treebank, increases accuracy by 1%.
- Sister-head model fares poorly.

Conclusions

- Poor results on German (Dubey and Keller, 2003): Due to flexible word order.
- Prediction

	Annotation	**Word Order**	**Lexicalization**
German - Negra	Flat	Flexible	Does not help
English - PTB	Non-Flat	Non-Flexible	Helps
French - FTB	Flat	Non-Flexible	Helps
?	Non-flat	Flexible	Will not help

- Test prediction on Korean.

Crosslinguistic comparison

Parsing results for corpora of same size as FTB datasets (Sent length $\leq$ 40).

Corpus	Model	LR	LP	CBs	0CB	$\leq$2CB
FTB	Cont+CR	64.49	64.36	1.99	35.87	70.17
	Model2	79.24	78.59	1.12	55.96	83.51
PTB	Unlex	73.97	76.63	2.30	33.55	63.20
	Model2	88.35	88.34	1.00	61.89	85.34
Negra	Unlex	70.56	66.69	1.03	58.21	84.46
	Model1	67.91	66.07	0.73	65.67	89.52

Negra: Training set 18,600 sentences; Testing set: 1,000 sentences.

PTB: Sections 00-09 (18,318 sentences); Testing set: first 1,000 sentences from section 23.

Perfect tagging - Results

Upper bound on parsing results - correct POS tags provided.

	LR	LP	CBs	OCB	$\leq$ 2CB	Tag	Coverage
Exp+CR	64.11	63.44	11.10	33.82	65.92	100.00	99.08
Cont+CR	67.78	67.07	1.84	36.42	71.99	100.00	98.32
Model 1	80.65	80.03	1.08	56.25	84.62	98.22	99.76
Model 2	80.79	80.23	1.07	56.44	83.39	98.25	99.64
SisterHead	78.22	77.24	1.26	50.79	81.00	97.94	98.56
Bigram	81.43	81.90	1.02	55.96	86.16	98.25	99.64
BigramFlat	81.26	80.88	1.02	56.37	85.94	98.22	99.64

Note: Bikel parser uses provided POS tags only for words in the test set that were seen fewer than 6 times during training.

Further research

Additional crosslinguistic analysis:

- Investigate impact of markovization and distance feature.
- Further treebank transformations.

Improve parsing performance for French:

- Better treatment of coordination and punctuation.
- Alternative treatment of compounds.
- Grandparent annotation.

Previous Work

- Czech: Collins et al.(1999). Using a refined version of the lexicalised Model 1 of (Collins 97), accuracy of 80%.

- Chinese: Chiang and Bikel.(2002). Using a refined version of the lexicalised Model 2 of Collins (1997), accuracy of 80%.

Overview

- Generalizing existing parsing models to new languages.
- French: annotation scheme and syntactic properties.
- Treebank transformation.
- Experiment 1: Unlexicalised models.
- Experiment 2: Lexicalised models.
- Crosslinguistic comparison.
- Conclusions.

French Treebank - Corpus Le Monde

- Small POS tag set (13 tags vs 36 tags in the PTB).
- Punctuation marks are represented as the single PONCT tag, no separate tags for modal verbs, *wh*-words and possessives.
- Verbs, adverbs and prepositions coarsely defined.
- Presence of a separate clitic tag(CL) for weak pronouns.

Modifying Collins' model

1. Collins Model 2

$$P_m(M_i(m_i)|P, H, w_h, t_h, d(i), subcat_{side})$$

2. Sister-head model: Extend Collins' base NP model to all syntactic categories.

$$P_m(M_i(m_i)|P, M(w,t)_{i-1})$$

3. Bigram model: Condition generation of non-modifying terminal on parent, head and previously generated modifier.

$$P_m(M_i(m_i)|P, H, w_h, t_h, d(i), M_{i-1}, subcat_{side})$$

4. Bigram-flat model: Apply bigram model only to categories with high degrees of flatness (SENT, Srel, Ssub, Sint, VPinf and VPpart).