Feature Engineering for Data-driven Dependency Parsing

Christian Rishøj Jensen
August 2009

Errata

	Presence			
	% of sentences with...			
Corpus	“	( )	:	-
Law	0.0%	*21.0%*	2.0%	0.0%
~~**Law2**~~	~~0.0%~~	~~*21.0%*~~	~~2.0%~~	~~0.0%~~
Newspaper	*14.0%*	*7.0%*	*11.0%*	*9.0%*
~~**Law+News**~~	~~*14.0%*~~	~~*28.0%*~~	~~*13.0%*~~	~~*9.0%*~~
Both	*7.0%*	*14.0%*	*6.5%*	4.5%

“Presence” metrics for **Italian** (table 3.2)

Errata

- **Bulgarian** (truncated treebank):
  - **Parenthesis** — 5% of $\text{sentences}_{\text{test}}$
    tiny adverse effect [$p \approx 0.49$]
  - **Colon** — 2% of $\text{sentences}_{\text{test}}$
    $\text{LAS}_{\text{overall}}$ *-0.36* [$p \approx 0.12$]

Errata

- **Slovene** (SDT):
  - **Colon** — 2% of sentences
    $LAS_{affected}$ *+1.39* [$p \approx$ *0.26*]
  - **DashApposition** — 2% of sentences
    $LAS_{affected}$ *+1.85* [$p \approx$ *0.09*]

Errata

- **Danish** (DDT):
  - **Colon** — 7% of sentences$_{test}$

    $LAS_{affected}$ *+1.42* [$p \approx$ *0.12*]
  - **DashApposition** — 0.3% (!)

    $LAS_{overall}$ *+0.34* [$p \approx$ *0.10*]

Errata

- **Spanish** (Cast3lb):
- **Colon** — 5% of $\text{sentences}_{\text{test}}$
  $\text{LAS}_{\text{affected}}$ *+1.11* [$p \approx$ *0.23*]

Unused Treebanks

- **BIO** (English)
- **Metu** (Turkish)
- **Alpino** (Dutch)
- **PDT** (Czech)
- **Tiger** (German)
- **Law2** (Italian, revised)

Treebank Selection

- **Objective** selection criteria
  - Not explicitly stated
- Challenges:
  - **Presence** of linguistic phenomena
  - **Impact** on parsing accuracy

Future Treebank Selection

- Gather data:
  - Treebank **characteristics**
  - Augmentation scheme **effects**
- **Automatically** determine applicable augmentations