Multi-lingual and Cross-genre Discourse Unit Segmentation

Peter Bourgonje and Robin Schäfer
Applied Computational Linguistics
University of Potsdam, Germany
{bourgonje,rschaefer}@uni-potsdam.de

Shared Task Description

- Attempt at cross-formalism (RST, SDRT, PDTB) discourse unit segmentation for multi-lingual and cross-genre data sets.
- Differences in representation of coherence relations among the different theories may be too large to gap, but the comparatively simple segmentation task could be a first step toward convergence.
- Implementing this from the available corpora is not straightforward, approach followed in the shared task:
  - Segmenting into EDUs for RST and SDRT corpora.
  - Identifying spans of explicit connectives for PDTB corpora.
- Languages: EN, RU, ZH, PT-BR, ES, DE, EU, FR, NL, (TR).
- Genres: Scientific writing, news, news commentary, encyclopedia texts, fund-raising letters, commercial ads, etc.
- Annotations available in CoNLL format (consisting of segmentation information and dependency parses).

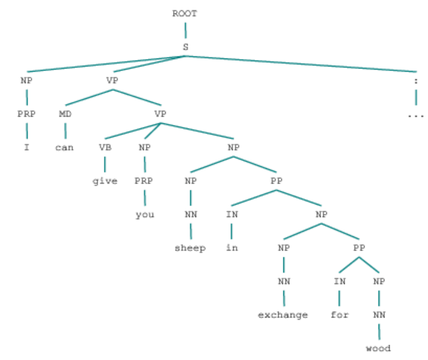
Data Sets

Corpus name	Language	Framework	Size (·1k tokens)
PDTB	English	PDTB	1,101
RRST	Russian	RST	244
RSTDT	English	RST	184
GUM	English	RST	83
CDTB	Chinese	PDTB	63
CSTN	Brazilian PT	RST	51
RSTSTB	Spanish	RST	51
STAC	English	SDRT	42
PCC	German	RST	30
RSTBT	Basque	RST	29
ANNODIS	French	SDRT	25
NLDT	Dutch	RST	21
SCTB	Spanish	RST	13
SCTB	Chinese	RST	11

Methods

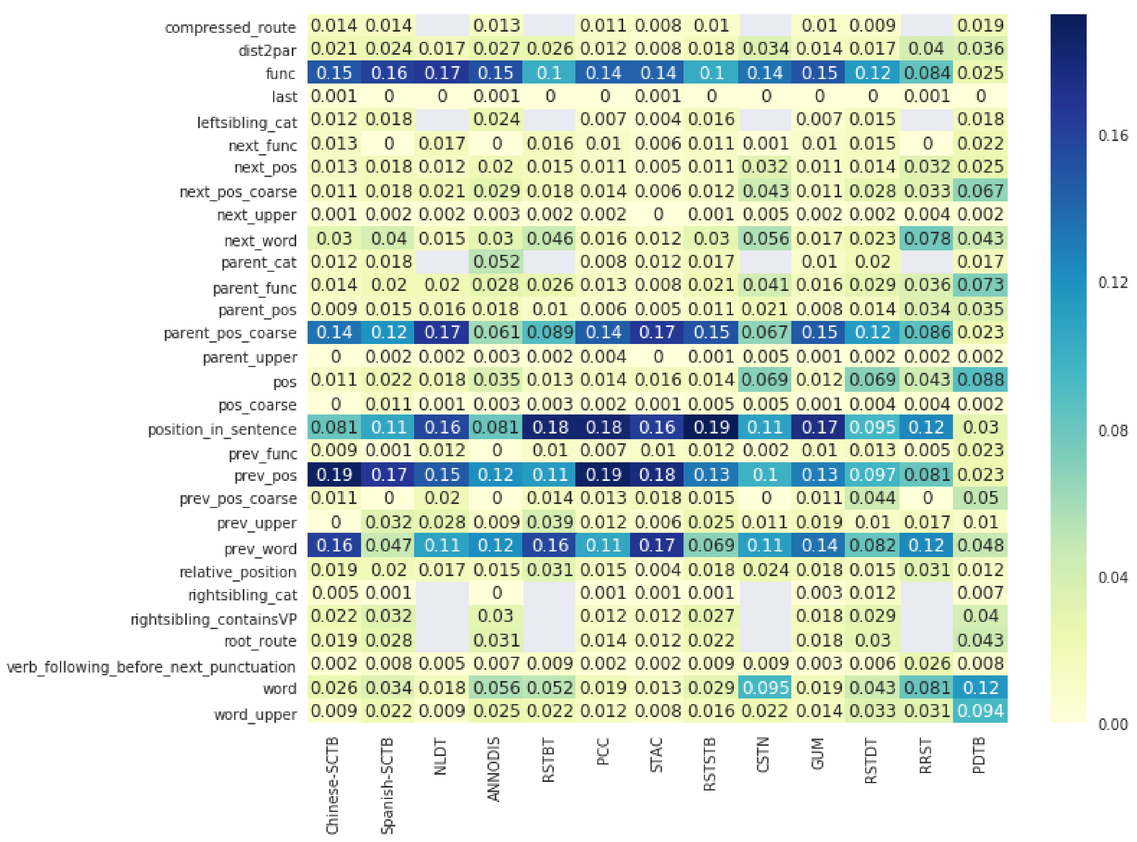
- Baseline: Segment start at the beginning of each sentence, optionally splitting on commas.
- Random forest: Surface (word and n-gram) features, distance to parent, dependency functions*, pos-tags*, orthographical feature, position in sentence (absolute and relative), verb in between token and next punctuation token. Constituency features (for languages supported by Stanford CoreNLP): category of parent, category of left and right siblings in tree, path to root node. (* of token, neighbours and parent)
- LSTM: Embedding of token using pre-trained embeddings or embeddings from corpus itself (if large enough), distance to parent, function and pos-tag of parent and token, orthographical feature, relative sentence position.

```
1     I          I          PRON     PRP    Case=Nom|Number=Sing|Person=1|PronType=Prs    3     nsubj    _    BeginSeg=Yes
2     can        can        AUX      MD     VerbForm=Fin                                  3     aux      _    _
3     give       give       VERB     VB     VerbForm=Inf                                  0     root     _    _
4     you        you        PRON     PRP    Case=Acc|Person=2|PronType=Prs                3     iobj     _    _
5     sheep      sheep      NOUN     NNS    Number=Plur                                   3     obj      _    _
6     in         in         ADP      IN     _                                             7     case     _    _
7     exchange               exchange        NOUN       NN     Number=Sing                 5     nmod     _    _
8     for        for        ADP      IN     _                                             9     case     _    _
9     wood       wood       NOUN     NN     Number=Sing                                   7     nmod     _    _
10    ...        ...        PUNCT    .      _                                             3     punct    _    _
```



Feature Gains

- Overall most informative features for random forest: previous pos-tag, position in sentence, (dependency) function, parent pos-tag (coarse), previous word (surface form).



Conclusion

EDU Segmentation - treebanked

	System	Mean F
1	ToNy (Inria)	90.11
2	GumDrop (Georgetown)	88.11
3	IXA (University of the Basque Country)	87.84
4	DFKI RF (Deutsches Forschungszentrum für Künstliche Intelligenz)	84.58

- Found large genre differences, but no indicators one particular language is *a priori* more difficult than another.
  - More research into genre differences with respect to discourse segmentation.
- EDU ≠ connective token span.
  - Re-visit existing work on (PDTB) argument/EDU convergence.
- More refined, unified notion of minimal segment and how to apply this to PDTB corpora is needed before cross-theory segmentation can succeed and corpora can benefit from each other for machine learning approaches.
- Code: https://github.com/PeterBourgonje/discourse_segmentation