DISRPT@NAACL2019
Minneapolis, MN

# GumDrop at the DISRPT2019 Shared Task

## A Model Stacking Approach to Discourse Unit Segmentation and Connective Detection

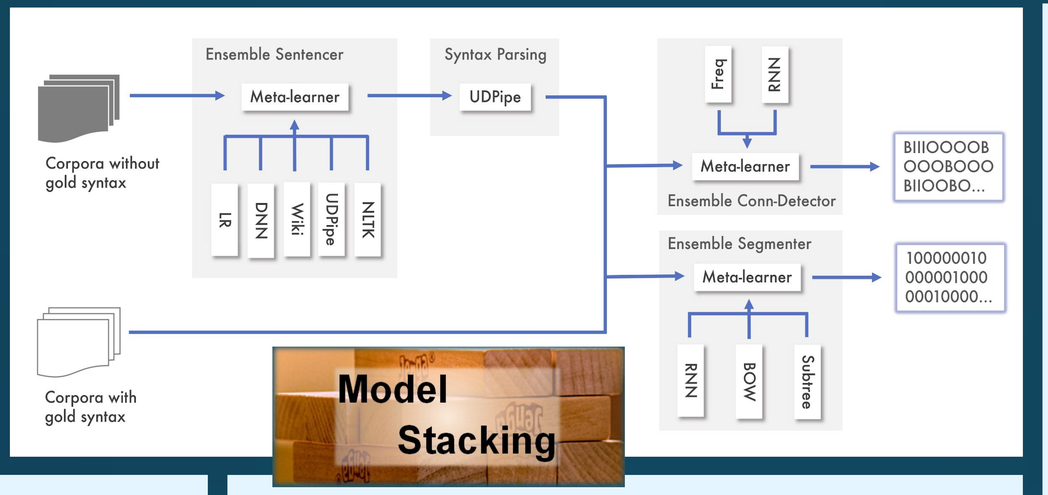
YUE YU, YILUN ZHU, YANG (JANET) LIU, YAN LIU, SIYAO (LOGAN) PENG, MACKENZIE GONG AND AMIR ZELDES
{yy476,yz565,yl879,yl1023,sp1184,mg1745,az364}@georgetown.edu

GEORGETOWN UNIVERSITY
1789
C@GU

PURPLE = best in shared task!

### Overview

- Discourse unit segmentation is still a problem, especially for smaller or less homogeneous corpora
- **GumDrop** relies on *model stacking* (Wolpert 1992), with 3 trainable stacks:
  - sentence splitting (usually: new sent > new EDU)
  - discourse unit segmentation
  - connective detection
- Heterogeneous ensembles of classifiers feed into **meta-learners**
  - 5-fold multitraining prevents submodule overreliance
  - Modules developed as a graduate student seminar project at Georgetown University

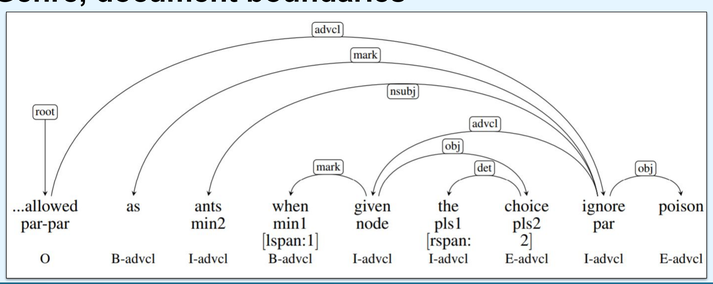



### System modules

- **Sentencer ensemble (metalearner: XGB [Chen & Guestrin 2016])**
  - Wiki-Based Sentencer relies on frequencies and ratios of paragraph-initial tokens extracted from Wikipedia dumps to capture potential sentence split points without punctuation, such as headings
  - DNN Sentencer uses pre-trained 300d FastText embeddings as input to a TensorFlow multilayer perceptron to predict sentence boundaries, with flexible windows: 3-gram for small & 5/7-gram for large corpora
  - LR Sentencer uses chars, POS tag and token length features in a 5-gram logistic regression without word identity to combat sparseness
  - UDPipe+NLTK: pretrained models (Bird et al. 2009, Straka et al. 2016)
- **Segmentation (metalearner: XGB)**
  - Subtree looks for potential split points via dependency subgraphs. *Input:* syntactic features for children & (grand) parents
  - BOW predicts the number of segments in each sentence (and non-segmented abstracts in Russian). *Input:* freqs of top 200 words
  - RNN binary classifier using bi-LSTM/CNN-CRF sequence labeling (Yang & Zhang 2018): word+char embeddings & syntactic features
- **Conn-Detection ensemble (metalearner: random forest)**
  - Frequency-based model looks up sequences seen as connectives with up to 5 tokens and returns their frequencies.
  - RNN predicts BIO label probabilities based on the top 5 optimal paths ranked by the CRF layer in a bi-LSTM/CNN-CRF.
- **Hyperparameter search** Considering the number of hyperparameters and the complexity of models, Bayesian Optimization (using the Tree Parzen Estimator or TPE, Bergstra et al. 2011) is applied for hyperparameter tuning, implemented using *hyperopt*.

### Features

- **Word representations**:
  - word-level: token string, embeddings
  - char-level: left/rightmost/all character, case, char count
  - other: POS, morphology, token length & frequency
- **N-gram window** centering a possible split point; also use parent and grandparent if syntax available
- **Top n words** (usually n=100/200)**:** replace POS with word for n most frequent words
- **Quot/paren:** whether a word occurs in between quotation marks or parentheses.
- **Sent %:** quantile position of current sent/tok (0-1 scale).
- **Syntactic features** (not available for sentencer):
  - Dependency relation and distance to parent
  - Smallest relevant phrase boundary with BIEO encoding
  - Governed token span and 'same head as neighbor'
- **Genre, document boundaries**




### Future plans

- Add contextual embeddings
- Deploy in rstWeb (Zeldes 2016)
- Release standalone sentencer

### Results

#### Sentence splitting

| | Baseline (./!/?) | | | NLTK | | | LR | | | GumDrop | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| corpus | P | R | F | P | R | F | P | R | F | P | R | F |
| deu.rst.pcc | 1.00 | .864 | .927 | 1.00 | .864 | .927 | .995 | .953 | .974 | .986 | .986 | **.986** |
| eng.pdtb.pdtb | .921 | .916 | .918 | .899 | .863 | .880 | .891 | .970 | .929 | .963 | .948 | **.955** |
| eng.rst.gum | .956 | .810 | .877 | .943 | .807 | .870 | .935 | .885 | .909 | .977 | .874 | **.923** |
| eng.rst.rstdt | .901 | .926 | .913 | .883 | .900 | .891 | .897 | .991 | .942 | .963 | .946 | **.954** |
| eng.sdrt.stac | .961 | .290 | .446 | .990 | .283 | .440 | .805 | .661 | .726 | .850 | .767 | **.806** |
| eus.rst.ert | .964 | 1.00 | .982 | .945 | .972 | .958 | 1.00 | 1.00 | **1.00** | 1.00 | .997 | .998 |
| fra.sdrt.annodis | .970 | .910 | .939 | .965 | .910 | .937 | .957 | .943 | .950 | .985 | .905 | **.943** |
| nld.rst.nldt | .991 | .919 | .954 | .983 | .919 | .950 | .951 | .931 | .941 | .980 | .964 | **.972** |
| por.rst.cstn | .984 | .992 | **.988** | .967 | .967 | .967 | .984 | .992 | **.988** | .984 | .984 | **.988** |
| rus.rst.rrt | .867 | .938 | .901 | .737 | .927 | .821 | .948 | .980 | **.964** | .952 | .972 | .962 |
| spa.rst.rststb | .912 | .851 | .881 | .938 | .845 | .889 | .996 | .934 | **.964** | .993 | .934 | .963 |
| spa.rst.sctb | .860 | .920 | .889 | .852 | .920 | .885 | .889 | .960 | **.923** | .857 | .960 | .906 |
| tur.pdtb.tdb | .962 | .922 | .942 | .799 | .099 | 176 | .979 | .979 | .979 | .983 | .984 | **.983** |
| zho.pdtb.cdtb | .959 | .866 | .910 | .– | .– | .– | .954 | .975 | .965 | .980 | .975 | **.978** |
| zho.rst.sctb | .879 | .826 | .852 | .– | .– | .– | 1.00 | .811 | **.895** | .991 | .795 | .882 |
| **mean** | .939 | .863 | .888 | .915 | .790 | .815 | .945 | .931 | .937 | .963 | .933 | **.947** |
| **std** | .046 | 167 | 128 | .079 | .273 | .235 | .055 | .089 | .065 | .046 | .070 | **.050** |

#### EDU Segmentation

| Gold syntax | Baseline | | | Subtree | | | RNN | | | GumDrop | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| corpus | P | R | F | P | R | F | P | R | F | P | R | F |
| deu.rst.pcc | 1.0 | .724 | .840 | .960 | .891 | **.924** | .892 | .871 | .881 | .933 | .905 | .919 |
| eng.rst.gum | 1.0 | .740 | .850 | .974 | .888 | .929 | .950 | .877 | .912 | .965 | .908 | **.935** |
| eng.rst.rstdt | 1.0 | .396 | .567 | .951 | .945 | .948 | .932 | .945 | .939 | .949 | .965 | **.957** |
| eng.sdrt.stac | .999 | .876 | .933 | .968 | .930 | .949 | .946 | .971 | **.958** | .953 | .954 | .953 |
| eus.rst.ert | .981 | .530 | .688 | .890 | .707 | .788 | .889 | .754 | .816 | .909 | .740 | **.816** |
| fra.sdrt.annodis | 1.0 | .310 | .474 | .943 | .854 | .897 | .894 | .903 | .898 | .944 | .865 | **.903** |
| nld.rst.nldt | 1.0 | .721 | .838 | .979 | .927 | .952 | .933 | .892 | .912 | .964 | .945 | **.954** |
| por.rst.cstn | .878 | .435 | .582 | .911 | .827 | .867 | .815 | .903 | .857 | .918 | .899 | **.908** |
| rus.rst.rrt | .760 | .490 | .596 | .809 | .745 | .775 | .821 | .710 | .761 | .835 | .755 | **.793** |
| spa.rst.rststb | .974 | .647 | .777 | .921 | .792 | .851 | .759 | .855 | .804 | .890 | .818 | **.853** |
| spa.rst.sctb | .970 | .577 | .724 | .938 | .631 | .754 | .901 | .649 | .754 | .898 | .679 | **.773** |
| zho.rst.sctb | .924 | .726 | .813 | .880 | .744 | .806 | .843 | .768 | .804 | .810 | .810 | **.810** |
| mean | .957 | .598 | .724 | .927 | .823 | .870 | .881 | .841 | .858 | .914 | .853 | **.881** |
| **Pred syntax** | **Baseline** | | | **Subtree** | | | **RNN** | | | **GumDrop** | | |
| corpus | P | R | F | P | R | F | P | R | F | P | R | F |
| deu.rst.pcc | 1.0 | .626 | .770 | .924 | .867 | .895 | .876 | .867 | .872 | .920 | .898 | **.909** |
| eng.rst.gum | .956 | .599 | .737 | .948 | .777 | .854 | .910 | .805 | .854 | .940 | .772 | **.848** |
| eng.rst.rstdt | .906 | .368 | .524 | .916 | .871 | .893 | .883 | .911 | .897 | .896 | .914 | **.905** |
| eng.sdrt.stac | .956 | .253 | .401 | .849 | .767 | .806 | .819 | .814 | .817 | .842 | .775 | **.807** |
| eus.rst.ert | .970 | .543 | .696 | .917 | .705 | .797 | .877 | .747 | .807 | .901 | .734 | **.809** |
| fra.sdrt.annodis | .980 | .285 | .442 | .938 | .824 | .877 | .892 | .915 | **.903** | .945 | .853 | .896 |
| nld.rst.nldt | .991 | .663 | .794 | .951 | .849 | .897 | .938 | .835 | .883 | .947 | .884 | **.915** |
| por.rst.cstn | .879 | .440 | .586 | .935 | .867 | **.900** | .788 | .883 | .833 | .930 | .851 | .888 |
| rus.rst.rrt | .664 | .463 | .545 | .825 | .717 | .767 | .813 | .731 | .770 | .821 | .748 | **.783** |
| spa.rst.rststb | .912 | .566 | .698 | .934 | .772 | **.845** | .820 | .871 | **.845** | .875 | .798 | .835 |
| spa.rst.sctb | .888 | .565 | .691 | .870 | .637 | .735 | .813 | .595 | .687 | .853 | .655 | **.741** |
| zho.rst.sctb | .798 | .589 | .678 | .806 | .643 | .715 | .803 | .607 | .692 | .770 | .696 | .731 |
| mean | .908 | .497 | .630 | .901 | .775 | .832 | .853 | .798 | .822 | .887 | .798 | **.839** |

#### Connective Detection

| Gold syntax | Baseline | | | Freq | | | RNN | | | GumDrop | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| corpus | P | R | F | P | R | F | P | R | F | P | R | F |
| eng.pdtb.pdtb | .964 | .022 | .044 | .836 | .578 | .683 | .859 | .871 | .865 | .879 | .888 | **.884** |
| tur.pdtb.tdb | .333 | .001 | .002 | .786 | .355 | .489 | .759 | .820 | .788 | .766 | .816 | .790 |
| zho.pdtb.cdtb | .851 | .259 | .397 | .715 | .618 | .663 | .726 | .628 | .674 | .813 | .702 | **.754** |
| mean | .716 | .094 | .148 | .779 | .517 | .612 | .781 | .773 | .776 | .819 | .802 | .809 |
| **Pred syntax** | **Baseline** | | | **Freq** | | | **RNN** | | | **GumDrop** | | |
| corpus | P | R | F | P | R | F | P | R | F | P | R | F |
| eng.pdtb.pdtb | .964 | .022 | .044 | .836 | .578 | .683 | .811 | .798 | .805 | .846 | .828 | **.837** |
| tur.pdtb.tdb | .333 | .001 | .002 | .786 | .355 | .489 | .761 | .821 | .790 | .768 | .817 | **.792** |
| zho.pdtb.cdtb | .851 | .259 | .397 | .715 | .618 | .663 | .705 | .590 | .642 | .806 | .673 | **.734** |
| mean | .716 | .094 | .148 | .779 | .517 | .612 | .759 | .736 | .746 | .806 | .773 | **.788** |

References
- Bergstra, J. S./Bardenet, R. Rémi/Bengio, Y. Yoshua/Kégl, B. 2011. Algorithms for hyper-parameter optimization. *NIPS 24*, 2546–2554.
- Bird, S./Klein, E./Loper, E. 2009. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly.
- Chen, T./Guestrin, C. 2016. XGBoost: A scalable tree boosting system. In *Proceedings ACM SIGKDD 2016*. San Francisco, CA, 785–794.
- David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.
- Straka, M./Hajič, J./Straková, J. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of LREC 2016*. Portorož, Slovenia, 4290–4297.
- Yang, Jie/Zhang, Yue 2018. NCRF++: an open-source neural sequence labeling toolkit. In *Proceedings of ACL 2018*. Melbourne, 74–79.
- Zeldes, A. 2016. rstWeb - A browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of NAACL-HLT 2016 System Demonstrations*. San Diego, CA, 1–5.

For error analysis see our paper!
https://aclweb.org/anthology/papers/W/W19/W19-2717/