Discourse Connective Identification

Phyllicia Leavitt, and Srilakshmi Balard
1st year Msc. NLP

Supervisor:
Chloé Braud

Discourse Parsing

Penn Discourse TreeBank (PDTB): Largest corpus annotated for discourse relations

Shallow Discourse Parsing (SDP): Automatic detection of predicate-argument relation between spans text (called shallow as it does not correspond to document-level tree-like structures).

Discourse connective: The semantico-pragmatic link between two spans of text

Connective identification: The identification of whether a word form is in discourse use.

Objectives: Train predictive models in absence of syntactical information derived from tree-parses, making use of the following features:

- Lexical information: word form which lexicalizes the connective (Connective token)
- Gold POS information from PDTB annotation
- Lexical and POS information of window of 3 tokens prior to and following connective token

Our Approach

- Accumulate a set of 100 connective token-types annotated in PDTB
- Label positive or negative according to PDTB annotation
- Train a binary classification model on the following features:
  1. Connective token itself
  2. Connective POS tag
  3. Trigrams of the 3 previous tokens from the connective
  4. Trigrams of the 3 tokens following the connective
  5. POStags of the 3 previous tokens from the connective
  6. POStags of the 3 next tokens following the connective
- Stack feature sets 1 and 2 only, as well as the ensemble of the 6 features on the four models:
  - Multinomial Naïve Bayes (MultiNB)
  - Perceptron (Percep)
  - Passive Aggressive (PA)
  - Logistic Regression (LR)

Data

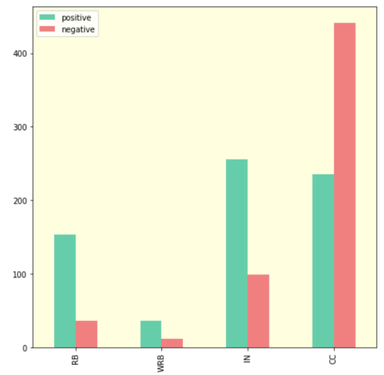
Preprocessing

PBTB 2.0 Annotation

Data Set	PDTB Sections	Connectives Annotated
Train	2-21	14,719
Dev	22	680
Test	23	923
	Total	16,322

Our Counts

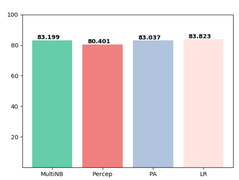
Data Set	Negative	Positive	Total
Train	12,933	14,737	27,670
Dev	588	680	1,268
Test	733	924	1,657
Total	14,254	16,341	30,595



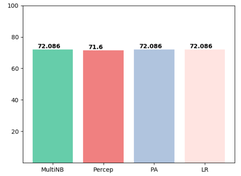
Dev set POS tag Distribution

Results

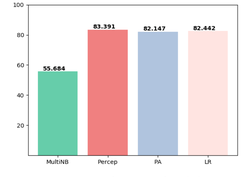
Experiments on Feature-sets



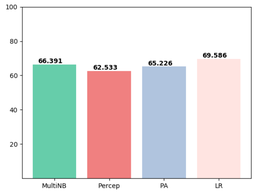
Token Only



Connective POS

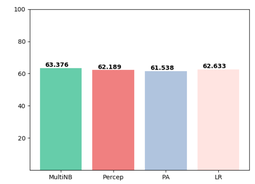


Token and POS



Previous Tri-grams

Feature-Set	MultiNB	Percep	PA	LR
Conn Token	83.20	80.40	83.04	**83.82**
Conn POS	72.09	71.60	72.09	72.09
Tok, POS	55.68	83.39	82.15	82.44
Prev 3 tri-grams	66.39	62.53	65.25	69.59
Next 3 tri-grams	72.02	56.95	64.28	67.52
Prev 3 POS	64.35	64.44	58.43	76.59
Next 3 POS	69.70	69.25	65.91	70.08
Global Results	63.38	62.20	61.54	62.63
Averaged f1 by classifier	63.35	68.85	69.08	**73.10**



Global Results

Settings and Evaluation

Tuned Hyper-parameters:

- Multinomial NB: Alpha increments of 0.04 from 1 to 3
- Passive Aggressive: C parameter at 1, 10, 100, 1000
- Logistic Regression: C parameter increments of 0.04 from 1 to 4
- Perceptron: Alpha increments of 0.04 from 1 to 3

Evaluation:

- Ten-fold cross-validation using GridSearchCV
- Scores reflect best f1 of the grid

Conclusion

Conclusions about results

- Connective token itself performs best compared to other isolated features or global feature-set
- POS information, when added, significantly increases Perceptron performance, but decreases Multinomial NB
- In the specific context of our modeling of the task, n-gram and POS information of 3 previous and following tokens only introduce noise to the model
- Logistic Regression performs best overall

Conclusions about approach:

- Performance may have been compromised due to difficulties in conforming PBTB counts during pre-processing
- More sophisticated feature engineering needed to increase performance

Future Work

- Re-work approach to pre-processing
- Do similar experiments on a predicted setting with the use of different POS taggers and syntactic parsers to train system for out-of–domain application
- Develop a point-wise (per-connective) classifier
- Introduce a neural network
- Experiment on other sub-tasks of Shallow Discourse Parsing

Acknowledgements

We would like to thank our supervisor, Chloé Braud, for her availability and guidance.

And thanks to the university of Lorraine department IDMC for affording us the opportunity to carry out the project.