clarin:el

Segmentation of Argumentative Texts with Contextualised Word Representations

Georgios Petasis

Software and Knowledge Engineering Laboratory,
Institute of Informatics and Telecommunications,
N.C.S.R. “Demokritos”, Athens, Greece

petasis@iit.demokritos.gr

Motivation

- Several approaches exist for detecting argumentative units, either at sentence or clause granularities
  - Park and Cardie, 2014; Goudas et al., 2014, 2015; Sardianos et al., 2015; Stab, 2017; Ajjour et al., 2017; Eger et al., 2017; etc.
  - Proposed approaches exploiting a plethora of features
    - Typically highly engineered and sophisticated, manually constructed, features
    - CRFs have been a popular algorithm for sequential labelling tasks

Motivation

- Deep learning is slowly replacing CRFs for sequence labelling
  - CRFs with manually constructed features
    - Park and Cardie, 2014; Goudas et al., 2014-15; Stab, 2017
  - CRFs with word embeddings
    - Sardianos et al., 2015
  - bi-directional LSTMs on manually engineered features
    - Ajjour et al., 2017
- Missing pieces:
  - CRF layer
  - Contextual embeddings (ELMo, Flair, BERT, etc.)

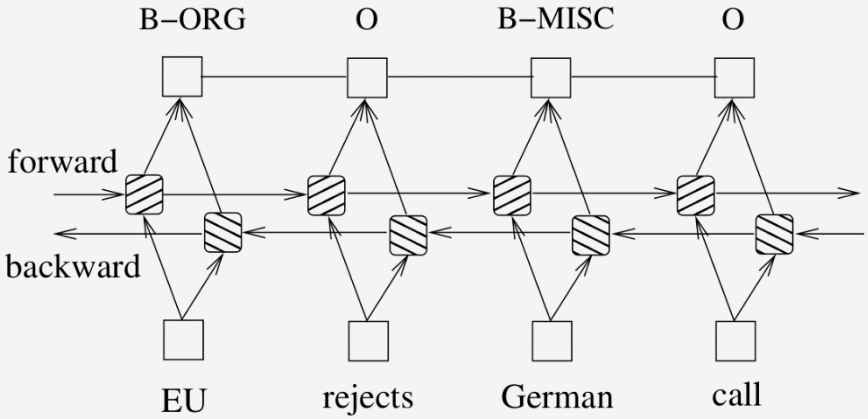
Research Questions

1. Can approaches that do not use manually engineered features achieve performances comparable to approaches that exploit such features?

2. Can contextualised word representations (pre-trained on large corpora) replace manually engineered features in argument mining?

Approach

- We have employed bidirectional LSTM-CRFs (Huang et al., 2015)



- We have replaced manually constructed features with word embeddings
  - Both non-contextual, and contextual
  - Combinations of embeddings
    - Concatenating embeddings into longer vectors

Experimental setting

- Corpus:
  - Stab and Gurevych (2017): 402 persuasive essays

		Number of tokens				
Part	**# Documents**	B-Arg	I-Arg	O-Arg	Total	Average
Train + Development	322	4,823	75,657	38,195	118,675	368.56
Test	80	1,266	18,837	9,442	29,545	369.31

Table 1: Number of documents, tokens per class, and average number of tokens per document.

- Two tasks:
  - Argumentative unit detection as sentence classification
  - Argumentative unit detection as sequential labelling

Task 1: AU detection as sentence classification

- We have applied BERT (Devlin et al., 2018) contextual embeddings with a single feed-forward layer on top of the embeddings
  - With a hidden layer equal to 768 nodes
  - Minimal fine-tuning:
    - A single epoch, learning rate $2e^{-5}$, 32 mini-batch size
- We compared to state-of-art approaches:
  - Bidirectional Sentence-State LSTMs (S-LSTMs) (Zhang et al., 2018), CNNs, bi-LSTM-CRFs
  - Non-contextual word embeddings (GloVe - Pennington et al., 2014)
    - 300 hidden layer size, tuned to $1-8$ layers, max 40 epochs, using 15,000 most frequent words, $1-6$ words window

Task 1: AU detection as sentence classification

- Evaluation results:

Embedding	Architecture	Accuracy
GloVe	CNN	0.8391
GloVe	LSTM	0.8488
GloVe	S-LSTM	0.8619
BERT	Feed Forward	**0.8874**

6 Bi-S-LSTM-CRF layers, with a window of 5 tokens, and after 15 epochs of fine-tuning

Task 2: AU detection as sequence labelling

- We have applied bidirectional LSTM-CRF
  - 2 layers, 256 hidden nodes, 32 mini-batch size
  - GloVe, Character embeddings, ELMo (Peters et al., 2018), Flair (Akbik et al., 2018) and BERT
- We have compared with:
  - (Stab, 2017): CRF with semantic, syntactic and structural features
  - (Ajjour et al., 2017): SVM/CRF/bi-LSTM with semantic, syntactic, structural and pragmatic features

Task 2: AU detection as sequence labelling

- Evaluation results:

Features	Model	Macro F_1
All (Semantic+Syntactic	SVM	61.40
+Structural+Pragmatic)	CRF	79.16
(Ajjour et al., 2017)	BI-LSTM	**88.54**
All		
(Stab, 2017)	CRF	**86.70**
GloVe + Character	BI-LSTM-CRF	85.92
GloVe + Character + Flair	BI-LSTM-CRF	88.17
ELMo	BI-LSTM-CRF	88.62
BERT	BI-LSTM-CRF	89.31
GloVe + Flair + BERT	BI-LSTM-CRF	**90.13**
GloVe + Flair + ELMo + BERT	BI-LSTM-CRF	87.42

89.18±2.45

Task 2: Error Analysis

- 270 sentences (out of 1448 test sentences) were erroneous classified
- 104 sentences were classified as containing argumentative units:
  - In spite of this, the disadvantages of the promotion of a universal language cannot be denied.
  - It is obvious that the benefits of the Internet undoubtedly outweigh its disadvantages.
  - It would be highly unpractical to ask people to adopt a simpler way of life.
  - Some people claim that without this punishment our lives would be less secure and crimes of violence would increase.
  - It is evident that technology promotes economy.

Task 2: Error Analysis

- Argumentative units were missed in 43 sentences:
  - However, it is not sufficient in itself.
  - Some people claim that the prevalent of English brings a great number of benefits for people.
  - In the modern world, computers are used everywhere.
  - There is no end to the evolution of computers.
  - Many people hold the opinion that past behavior determines the future actions, which could be the main reason to support the idea of revealing the record to the jury.

Task 2: Error Analysis

- The rest of the errors (123 sentences) contain various errors, like:
  - Merging argumentative units:
    - For instance, some Asians are seeking individualism, previously denied by many Asian countries, due to the fact that they have gradually identified with such values expressed in American movies, which are imported by the governments as a result of the proliferation of English.
    - First and foremost, sports events are good chances for excellent athletes to meet and learn valuable experiences from one another, so that they can improve their results, break records and bring victories to their own countries.

Task 2: Error Analysis

- The rest of the errors (123 sentences) were various errors, like:
  - Missing parts:
    - From personal level, it fosters a sense of unfairness between the older and younger generations.
    - From social perspective, massively forcing the early retirement would be one of financial burden to the local government.

Conclusions

1. Can approaches that do not use manually engineered features achieve performances comparable to approaches that exploit such features?
   - Manually constructed features can be substituted with standard architectures and word embeddings
2. Can contextualised word representations replace manually engineered features?
   - A small increase in state-of-art
     - Manually engineered features are still relevant and significant at least for this task
     - According to (Ajjour et al., 2017), semantic features appear to be the most significant features

Future work

- Evaluation on more corpora
- Significant optimisation potential, especially through hyperparameter tunning
  - Although computational requirements for some models are high

Thank you!

Εθνική Υποδομή για τις Ψηφιακές Ανθρωπιστικές Επιστήμες και Τέχνες και για τη Γλωσσική Έρευνα και Καινοτομία

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης

Ελληνική Δημοκρατία
Υπουργείο Οικονομίας & Ανάπτυξης
Ειδική Γραμματεία ΕΠΑ & ΤΣ
Ειδική Υπηρεσία Διαχείρισης ΕΠΑνΕΚ

ΕΠΑνΕΚ 2014-2020
Επιχειρησιακό Πρόγραμμα
Ανταγωνιστικότητα Επιχειρηματικότητα Καινοτομία

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης