CA Tutorial: Argumentative Unit Segmentation

21st May 2019

Denis Kuchelev, Enri Ozuni & Nikit Srivastava

Abstract. In this tutorial, we will accomplish the task of training classifiers on argument annotated data to enable them to classify argument-units and their non-argumentative counterparts in a given text (argumentative) document. We will also compute the performance of our classifiers.

1 Introduction

Argumentative unit segmentation is an important part of Argument Mining pipeline. Hence, in this tutorial we will use the Annotated Argumentative Student Essays[1] corpus to train a classifier on its tokens. The classifier will be provided tokens with labels in form of IOB-Tags. To define features for each token we will use/implement functions to extract linguistic features. Upon successful completion of our model training, we will evaluate our model on the test dataset and compare the results between different implementations.

2 Pre-requisites

Before we proceed, please make sure you have the following items downloaded / installed:

1. Make sure you have Jupyter and Python installed

2. Python Libraries to install:

```
pip install pprint
pip install python-crfsuite
pip install nltk
pip install scikit
pip install nltk
pip install gensim
```

[1] https://www.informatik.tu-darmstadt.de/ukp/research_6/data/argumentation_mining_1/argument_annotated_essays/index.en.jsp

```
pip install numpy
pip install stanfordnlp
```

3. Pre-Trained word embeddings model to download:

   a. **Word2Vec** Google News model
      - Trained on: 100 Billion tokens
      - Vocab size: 3 Million (words and phrases)
      - Memory Requirement: 3.5 Gigabytes
      - Listed at: https://code.google.com/archive/p/word2vec/
      - Directlink: https://bit.ly/1R9Wsqr

   b. **GloVe** Wikipedia 2014 + Gigaword 5 model
      - Trained on: 6 Billion tokens
      - Vocab size: 400k (words)
      - Memory Requirement: 680 Megabytes
      - Listed at: https://github.com/stanfordnlp/GloVe
      - Direct link: http://nlp.stanford.edu/data/wordvecs/glove.6B.zip

3 Introduction to Word embeddings

We have created a separate notebook named "Word-Embeddings-Playground". In this notebook you could find the code snippets of following items :

- Model conversion to Word2Vec
- Model loading
- Miscellaneous model function calls

4 Training a token based argumentative unit classifier

Our main task for this tutorial is to implement an argumentative unit classifier using the existing python libraries. For simplicity we have divided this task into further steps that will help accomplish the main task.

4.1 Steps to follow

1. Load the annotated data
2. Perform IOB-Tagging for tokens
3. Prepare feature extraction for each token
4. Split dataset into training and test
5. Select and train the classifier
6. Perform predictions on the test set

7. Evaluate the model

4.2 Initial Setup

We have created a class “GenModel” in “main.py” that contains the implementations of basic functions that you will to need to complete the task.

The class also contains the unimplemented parts that are to be completed by you.
In your text editor search for “# TODO” to find all the parts that are to be completed by you.

Please Note:

To save time, do not frequently rerun the cell with word embeddings model loading logic. The word embeddings model once loaded in the variable can be re-used.

In order for your code changes to take effect, you should execute cell where the model is being imported. The statement “importlib.reload(main)” makes sure that the latest version of “main.py” is being loaded.

4.3 Next Steps

Starting with “Step 1” in the Jupyter notebook, check the function calls being made in the cells and make sure they are implemented fully in order to progress to the next step.

References

1. Christian Stab and Iryna Gurevych Annotating Argument Components and Relations in Persuasive Essays. IProceedings of the the 25th International Conference on Computational Linguistics (COLING 2014), p.1501-1510, 2014.

2. Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth and Benno Stein "Unit Segmentation of Argumentative Texts " Proceedings of the 4th Workshop on Argument Mining, September 2017