Argumentation ranking semantics as a feature for classification

on automatic evaluation of argumentative essays

R. Barile[a]

[a]*Dipartimento di Informatica – Università di Bari*

September, 2022

Abstract

Writing is crucial for success. In particular, argumentative writing fosters critical thinking and civic engagement skills, and can be strengthened by practice. The goal of this work is to classify argumentative elements in student writing as "effective", "adequate", or "ineffective". Here we propose to use as an additional feature, in the training process of a classifier, the ranking score obtained by performing argumentative reasoning on the different argumentative elements of an essay. We show that the introduction of this feature lead to an improvement in the performances of both Ada boost classifier and biLSTM neural network.

1. Introduction

With automated guidance, students can complete more assignments and ultimately become more confident, proficient writers so we explore the task of automatic evaluation of argumentative essays[1]. Each argumentative essay can be split into several parts called "discourse elements". Each discourse element can play one of the following roles:

- **Lead**: an introduction that begins with a statistic, a quotation, a description, or some other device to grab the reader's attention and point toward the thesis
- **Position**: an opinion or conclusion on the main question
- **Claim**: a claim that supports the position
- **Counterclaim**: a claim that refutes another claim or gives an opposing reason to the position
- **Rebuttal**: a claim that refutes a counterclaim
- **Evidence**: ideas or examples that support claims, counterclaims, or rebuttals.
- **Concluding statement**: a concluding statement that restates the claims.

[1]This task was a competition on Kaggle, details on the evaluation metric and the data used in this work can be found at https://www.kaggle.com/competitions/feedback-prize-effectiveness

The evaluation is performed predicting the quality of each discourse element, quality can be, in order of increasing quality, one of:

- Ineffective
- Adequate
- Effective

The approach used in this work is text classification, we try to improve performances introducing additional features, the discourse type and a number obtained by performing argumentative reasoning to compute the strength propagation ranking semantics (sp-ranking) of a Bipolar Weighted Argumentation Framework (BWAF) [8]. The intuition behind this experiment is that the quality of a discourse element in an argumentative essay depends also on how it "attacks" or "supports" other discourse elements, so not only on textual features or on the grammatical or syntactical quality detected by individually and independently analyzing the discourse element.
We evaluate the usefulness of this feature by analyzing how two state of the art models behave with and without this feature.
The metric used for evaluation is multi-class logarithmic loss, defined as follows:

$$log\_loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij})$$

where $\mathcal{N}$ is the number of rows in the test set, $\mathcal{M}$ is the number of class labels, log is the natural logarithm, y_{ij} is 1 if observation i is in class j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to class j.

2. Related works

Several strategies exist for automated evaluation of student argumentative writing, in particular many of them are summarized and compared in the survey proposed in [9] which explores methods belonging to three main categories:

- **feature-based** where off-the-shelf algorithms are used with additional hand-crafted features, in particular features can be:
  - **lexical** features which aim to capture word-level information and common lexical features
  - **syntactic** features which commonly rely on parse trees
  - **structural** which generally describe the position and frequency of a piece of text
  - **embedding** which are based on word vectors that represent words in a continuous space
  - **discourse** which captures how sentences or clauses are connected together
- **neural-based** where neural architectures such as long short-term memory (LSTM) networks and convolutional neural networks (CNN) are adopted
- **unsupervised** methods which use heuristics for bootstrapping a small set of labels and then train the model in a self-training fashion.

The main difference between the approaches summarized in the cited survey and our work is about the dataset, the data on which we focused has the peculiarity of split essays which allows to build argumentation frameworks as we will explain in detail in the following section.
Furthermore, the purpose of this work is not to compare our solution with existing approaches, but to test if the proposed feature can be an improvement in any given classifier, for instance the ones we evaluated.

3. Methodology

The main contribution of this work is the introduction of a numerical feature in addition to the textual data, so in order to understand how it is computed we briefly explain how the argumentation reasoning works and then we present how it is applied to this task.

3.1. Argumentation

In this section, we briefly review Dung's [3] argumentation framework and his extensions including bipolar weighted argumentation framework on which we will focus.

Definition 1 An *argumentation framework* (or AF) is a pair $\mathcal{F} = \langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}$ is a finite set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is an attack relationship (meaning that, given $\alpha, \beta \in \mathcal{A}$, if $\alpha \mathcal{R} \beta$ then α attacks β).
An argumentation semantics is the formal definition of a method ruling the argument evaluation process. Then, standard acceptability semantics, introduced by Dung [5], characterize admissible sets of arguments.
There are different kind of extension of the Dung's AF that has been implemented:

- **Bipolar AFs** (or BAFs) [2] : allow two kinds of interactions between arguments, expressed respectively by the attack relation and the support relation;
- **Weighted AFs** (or WAFs) [4] : consent to specify a numeric weight for each attack between arguments, indicating its relative strength;
- **Bipolar Weighted Argumentation Framework** (or BWAF) [8]: embed the notions of attack and support into the weights.

Definition 2 A *BWAF* is a triplet $\mathcal{G} = \langle \mathcal{A}, \mathcal{R}, w_R \rangle$, where $\mathcal{A}$ is a finite set of arguments, $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ and w_R: $\mathcal{R} \mapsto [-1, 0[\cup]0, 1]$ is a function assigning a weight to each relation. Attack relations are defined as

$$\mathcal{R}_{att} = \{ \langle a, b \rangle \in \mathcal{R} \,|\, w_R(\langle a, b \rangle) \in [-1, 0[\,\}$$

and support relations as

$$\mathcal{R}_{sup} = \{ \langle a, b \rangle \in \mathcal{R} \,|\, w_R(\langle a, b \rangle) \in]0, 1] \,\}$$

For some applications, it can be problematic to have only two levels of evaluations (arguments are either accepted or rejected). In order to fix these problems, a solution consists in using semantics that distinguish arguments with a larger number of levels of acceptability. Ranking-based semantics [1] aim at determining such a ranking among arguments. We decided to use, as a ranking semantics, suitable for BWAFs the strength propagation semantics proposed in [8] and defined in the following.

Definition 3 Let $\mathcal{G} = \langle \mathcal{A}, \mathcal{R}, w_R \rangle$ be a BWAF and $a, b \in \mathcal{A}$ be two arguments such that there

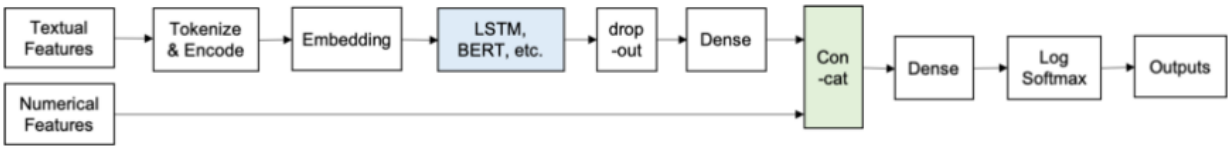

Figure 1: biLSTM additional layer for integration of numeric features

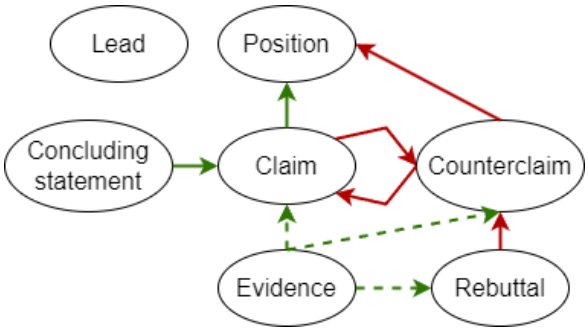

Figure 2: General structure of a BWAF built on the discourse elements of an argumentative essay. Green arrows represent supports, red arrows represent attacks, the green arrows starting from evidences are dashed because an evidence can express support only to one discourse element (of type claim, counterclaim or rebuttal), in particular, the most similar one.

exists a simple path $\langle a \dots b \rangle$. The strength propagation (sp) from a towards b is defined as:

$$sp(a,b) = \sum_{\langle a \dots b \rangle} (pw(\langle a \dots b \rangle)) \times \prod_{c \in \langle a \dots b \rangle} infl(c))$$

Function pw$(\cdot)$ (path weight) computes the strength of a simple path by multiplying every weight relation in it, while function infl$(\cdot)$ (influence) computes the influence of a node within the simple path on the basis of cycles to which it belongs.

Definition 4 Let $\mathcal{G} = \langle \mathcal{A}, \mathcal{R}, w_R \rangle$ be a BWAF, $a \in \mathcal{A}$ an argument, $sp(\cdot, a)$ the strength propagation of path branch of path branch ending to a, SP $= \{sp(x_1, a), \dots, sp(x_m, a)\}$ the set of all the strength propagations ending to a and $\mathcal{P} = \{p_1, \dots, p_n\}$ the set of all directed paths towards a in $\mathcal{G}$, with $p_i = \langle x, \dots, a \rangle \in \mathcal{P}, \forall i \leq n$. The spr function $spr : \mathcal{A} \mapsto [0, 2]$ is defined as:

$$spr(a) = \begin{cases} 1 \\ \text{if } \forall x \in \mathcal{A} : \langle x, a \rangle \notin \mathcal{R} \\ \frac{1}{n} \sum_{sp(x_i,a) \in SP}^{m} 1 + sp(x_i, a) \quad \text{otherwise} \end{cases}$$

3.2. Build a BWAF from an argumentative essay

As explained in the introductory section the different discourse elements have specific roles in the whole essay, according to such roles we define a general structure, reported in figure 2, to follow in order to build a BWAF for each argumentative essay.
The weights of attacks and supports are obtained computing a similarity function between document embeddings [6] which represents the semantics of the text in a continuous vector space, in this task the documents are the discourse elements. After downloading a pretrained model (e.g. word2vec [7]) which maps an embedding vector to words in the vocabulary, each discourse element is translated into an embedding vector with the following procedure:

1. extract an embedding vector from the word2vec model for each word in the discourse element
2. compute the weighted average of the embedding vectors using as weighting factors the tf - idf score of each word.

The sp-ranking semantics can now be computed on each BWAF obtaining the ranking of each discourse element with respect to other discourse elements in an argumentative essay.

3.3. Classification method

As explained in the introduction, the evaluation task is solved using text classification, so we explore two classification methods:

- **Ada Boost classifier** trained on a bag of words representation of discourse elements, in which each word is associated to its tf - idf score.
- **biLSTM neural networks** with the embedding layer initialized from a pretrained model [2]. The loss function used in the training process is the negative log likelihood loss, this choice is justified by the similarity in the behaviour between the loss function and the evaluation metric adopted in the testing phase.

[2] In our experiments we downloaded the word2vec model trained on Google news and with embedding size equal to 300

We slightly modified both models in order to include the additional features. In the standard classification the integration is straightforward, the sp-ranking and the discourse type are just other numerical values, analogously to the tf - idf scores. For the biLSTM network, instead, we need to add a concatenation step as represented in figure 1.

4. Evaluation

In the evaluation phase we adopted the dataset described in the introduction, but we needed to filter out some essays, in particular we kept only essays composed of less than 15 discourse elements. This choice is required to keep the experiments computationally feasible because the complexity of the argumentation semantics increases with the number of arguments in the graph.
We separately discuss the evaluation of the two approaches.

4.1. Ada Boost classifier

The setup adopted for this approach consists in:

1. Perform $n(= 10)$ iterations of cross-validation in which the model is trained **without** taking into account the sp-ranking
2. Perform $n(= 10)$ iterations of cross-validation in which the model is trained taking into account the sp-ranking
3. Perform a corrected resampled t-test to check if there is a significative difference between the two approaches

To perform the statistical test we compute the value of t as follows:

$$t = \frac{m_d}{\sqrt{(\frac{1}{k} + \frac{n_2}{n_1})\sigma^2}}$$

where:

- m_d is the mean of the differences between the log losses of the two setups
- $k = 100$ since we perform a 10-fold cross-validation for 10 times
- $n2 = 0,1$ and $n1 = 0,9$ because the set is split in 10 folds
- σ^2 is the variance of the differences between the log losses of the two setups

The value of t is -6.466, we fix as confidence level for the statistical test $c = 5\%$ so we look out for the value z corresponding to $\frac{c}{2}$ on the Student's distribution with $k - 1$ degrees of freedom, so we have $z = 1.984$; since t is less than $-z$ we can reject the null hypothesis of the test.

Table 1: biLSTM 5-fold cross-validation with hp-tuning

batch size	learning rate	epochs number
64	0.005	9
128	0.0005	10
64	0.005	8
32	0.0005	9
32	0.005	7
use sp-ranking	**log loss**	
No	0.859	
Yes	0.853	
No	0.840	
Yes	0.848	
No	0.861	

4.2. biLSTM

During the training process of this model we adopted a train set - validation set split in order to perform the tuning of the hyper-parameters, in particular to choose an appropriate value of batch size, learning rate and number of epochs. For the batch size the possible values are $\{32, 64, 128\}$ and for learning rate the possible values are $\{0,0005, 0,005\}$; while for the number of epochs, instead of using a set of candidate values, we fix the values to 10 and we adopt an early stopping if the log loss on the validation set increases due to overfitting. In addition, to decide whether to introduce the sp-ranking or not we consider this choice as a boolean hyper-parameter to tune contextually to the other parameters. So we report in table 1, for each test fold of a 5-fold cross validation the parameters selected after the tuning and the performances obtained after training the model with such parameters.
We show that in 2 out of 5 folds the sp-ranking is taken into account.

5. Conclusion and future work

We explored the task of automatic evaluation of argumentative essays, proposing a text classification approach, we proposed to compute a new feature using argumentative reasoning on a BWAF built on each essay following a general structure. To evaluate the usefulness of this feature we experimented with a standard classification algorithm and a biLSTM neural network comparing the performances with and without the sp-ranking feature.
As already mentioned in the evaluation phase an important drawback of this approach is the complexity of argumentation which limits the computation of the ranking to small graphs.

To further investigate this topic we can use a more complex reasoning strategy, for instance a general argumentation framework [5] in which we can express intrinsic weight of a node (e.g. based on the grammatical and syntactical correctness of the discourse element) or experiment with different measures of similarity as weights of the relations in the AF.

References

[1]L. Amgoud and J. Ben-Naim, ≪Ranking-based semantics for argumentation frameworks≫, in International conference on scalable uncertainty management (Springer, 2013), pp. 134–147.

[2]C. Cayrol and M.-C. Lagasquie-Schiex, ≪On the acceptability of arguments in bipolar argumentation frameworks≫, in European conference on symbolic and quantitative approaches to reasoning and uncertainty (Springer, 2005), pp. 378–389.

[3]P. M. Dung, ≪On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games≫, Artificial intelligence **77**, 321–357 (1995).

[4]P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, and M. Wooldridge, ≪Weighted argument systems: basic definitions, algorithms, and complexity results≫, Artificial Intelligence **175**, 457–486 (2011).

[5]S. Ferilli, ≪Introducing general argumentation frameworks and their use≫, in International conference of the italian association for artificial intelligence (Springer, 2020), pp. 136–153.

[6]Q. V. Le and T. Mikolov, *Distributed representations of sentences and documents*, 2014, 10.48550/ARXIV.1405.4053.

[7]T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, 2013, 10.48550/ARXIV.1301.3781.

[8]A. Pazienza, S. Ferilli, and F. Esposito, ≪Constructing and evaluating bipolar weighted argumentation frameworks for online debating systems.≫, in Ai^3@ ai* ia (2017), pp. 111–125.

[9]X. Wang, Y. Lee, and J. Park, *Automated evaluation for student argumentative writing: a survey*, May 2022.