Using Context to Predict the Purpose of Argumentative Writing Revisions

Fan Zhang
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA, 15260
zhangfan@cs.pitt.edu

Diane Litman
Department of Computer Science and LRDC
University of Pittsburgh
Pittsburgh, PA, 15260
litman@cs.pitt.edu

Abstract

While there is increasing interest in automatically recognizing the argumentative structure of a text, recognizing the argumentative purpose of revisions to such texts has been less explored. Furthermore, existing revision classification approaches typically ignore contextual information. We propose two approaches for utilizing contextual information when predicting argumentative revision purposes: developing contextual features for use in the classification paradigm of prior work, and transforming the classification problem to a sequence labeling task. Experimental results using two corpora of student essays demonstrate the utility of contextual information for predicting argumentative revision purposes.

1 Introduction

Incorporating natural language processing into systems that provide writing assistance beyond grammar is an area of increasing research and commercial interest (e.g., (Writelab, 2015; Roscoe et al., 2015)). As one example, the automatic recognition of the purpose of each of an author’s revisions allows writing assistance systems to provide better rewriting suggestions. In this paper, we propose context-based methods to improve the automatic identification of revision purposes in student argumentative writing. Argumentation plays an important role in analyzing many types of writing such as persuasive essays (Stab et al., 2014), scientific papers (Teufel, 2000) and law documents (Palau and Moens, 2009). In student papers, identifying revision purposes with respect to argument structure has been used to predict the grade improvement in the paper after revision (Zhang and Litman, 2015).

Existing works on the analysis of writing revisions (Adler et al., 2011; Bronner and Monz, 2012; Daxenberger and Gurevych, 2013; Zhang and Litman, 2015) typically compare two versions of a text to extract revisions, then classify the purpose of each revision in isolation. That is, while limited contextual features such as revision location have been utilized in prior work, such features are computed from the revision being classified but typically not its neighbors. In addition, ordinary classifiers rather than structured prediction models are typically used. To increase the role of context during prediction, in this paper we 1) introduce new contextual features (e.g., the impact of a revision on local text cohesion), and 2) transform revision purpose classification to a sequential labeling task to capture dependencies among revisions (as in Table 1). An experimental evaluation demonstrates the utility of our approach.

2 Related Work

There are multiple works on the classification of revisions (Adler et al., 2011; Javanmardi et al., 2011; Bronner and Monz, 2012; Daxenberger and Gurevych, 2013; Zhang and Litman, 2015). While different classification tasks were explored, similar approaches were taken by extracting features (location, text, meta-data, language) from the revised text to train a classification model (SVM, Random Forest, etc.) on the annotated data. One problem with prior works is that the contextual features used were typically shallow (location), while we cap-
Writer Richard Louv tells us to focus more on nature through his rhetorical questions, parallelism, and pathos. Louv’s rhetorical questions ask whether we value technology or nature over the other.

First Revision: 1->1, Type: Claim, Modify. Second Revision: 2->null, Type: Warrant, Delete

Table 1: Example dependency between Claim and Warrant revisions. Sentence 1 acts as the Claim (argument structure) of Draft 1 and sentence 2 acts as the Warrant for the Claim. Sentence 1 in Draft 1 is modified to sentence 1 (also acts as the Claim) of Draft 2. Sentence 2 in Draft 1 is deleted in Draft 2. The first revision is a Claim revision as it modifies the Claim of the paper by removing “rhetorical questions.” This leads to the second Warrant revision, which deletes the Warrant for “rhetorical questions.”

Table 2: Distribution of revisions in Corpus A, B.

3 Data Description

Revision purposes. To label our data, we adapt the schema defined in (Zhang and Litman, 2015) as it can be reliably annotated and is argument-oriented. Sentences across paper drafts are aligned manually based on semantic similarity and revision purpose categories are labeled on aligned sentences. The schema includes four categories (Claims/Ideas, Warrant/Reasoning/Backing, Rebuttal/Reservation and Evidence) based on Toulmin’s argumentation model (Toulmin, 2003), a General Content category for revisions that do not directly change the support/rebuttal of the claim (e.g. addition of introductory materials, conclusions, etc.), and three categories (Conventions, Clarity and Organization) based on the Surface categorizations in (Faigley and Witte, 1981). As we focus on argumentative changes, we merge all the Surface sub-categories into one Surface category. As Zhang and Litman (2015) reported that both Rebuttals and multiple labels for a single revision were rare, we merge Rebuttal and Warrant into one Warrant category and allow only a single (primary) label per revision.

Corpora. Our experiments use two corpora consisting of Drafts 1 and 2 of papers written by high school students taking AP-English courses; papers were revised after receiving and generating peer feedback. Corpus A was collected in our earlier pa-
paper (Zhang and Litman, 2015), although the original annotations were modified as described above. It contains 47 paper draft pairs about placing contemporaries in Dante’s Inferno. Corpus B was collected in the same manner as A with agreement Kappa 0.69. It contains 63 paper draft pairs explaining the rhetorical strategies used by the speaker/author of a previously read lecture/essay. Both corpora were double coded and gold standard labels were created upon agreement of two annotators. Two example annotated revisions from Corpus B are shown in Table 1, while the distribution of annotated revision purposes for both corpora are shown in Table 2.

4 Utilizing Context

4.1 Adding contextual features

Our previous work (Zhang and Litman, 2015) used three types of features primarily from prior work (Adler et al., 2011; Bronner and Monz, 2012; Daxenberger and Gurevych, 2013) for argumentative revision classification. Location features encode the location of the sentence in the paragraph and the location of the sentence’s paragraph in the essay. Textual features encode revision operation, sentence length, edit distance between aligned sentences and the difference in sentence length and punctuation numbers. Language features encode part of speech (POS) unigrams and difference in POS tag counts.

We implement this feature set as the baseline as our tasks are similar, then propose two new types of contextual features. The first type (Ext) extends prior work by extracting the baseline features from not only the aligned sentence pair representing the revision in question, but also for the sentence pairs before and after the revision. The second type (Coh) measures the cohesion and coherence changes in a 2-sentence block around the revision.

Utilizing the cohesion and coherence difference. Inspired by (Lee et al., 2015; Vaughan and McDonald, 1986), we hypothesize that different revisions can have different impacts on the cohesion and coherence of the essay. We propose to extract features for both impact on cohesion (lexical) and impact on coherence (semantic). Inspired by (Hearst, 1997), sequences of blocks are created for sentences in both Draft 1 and Draft 2 as demonstrated in Figure 1. Two types of features are extracted. The first type describes the cohesion and coherence between the revised sentence and its adjacent sentences. The similarity (lexical/semantic) between the revised sentence block and the sentence block before and after the revision are calculated as the cohesion/coherence scores Coh_Up and Coh_Down. The features are extracted separately for Draft 1 and Draft 2 sentences. The second type describes the impact of sentence modification on cohesion and coherence. Features Change_Up and Change_Down are extracted as the division of the cohesion/coherence scores of two drafts ($\frac{\text{Coh}_\text{Up}(\text{Draft 2})}{\text{Coh}_\text{Up}(\text{Draft 1})}$ and $\frac{\text{Coh}_\text{Down}(\text{Draft 2})}{\text{Coh}_\text{Down}(\text{Draft 1})}$).

A bag-of-word representation is generated for

---

Figure 1: Example of cohesion blocks. A window of size 2 is created for both Draft 1 and Draft 2. Sequence of blocks were created by moving the window at the step of 1 (sentence).

Figure 2: Example of revision sequence transformation. Each square corresponds to a sentence in the essay, the number of the square represents the index of the sentence in the essay. Dark squares are sentences that are changed. In the example, the 2nd sentence of Draft 1 is modified, the 3rd sentence is deleted and a new sentence is added in Draft 2.

---

2In this paper we consider the most adjacent sentence only.

3For the added and deleted sentences, features of the empty sentence in the other draft are set to 0.

4The feature values of sentence additions/deletions are 0.
Table 3: The average of 10-fold (student) cross-validation results on Corpora A and B. Unweighted precision (P), Unweighted recall (R) and Unweighted F-measure (F) are reported. Results of CRFs on paragraph-level segments are reported (there is no significant difference between essay level and paragraph level). * indicates significantly better than the baseline. Bold indicates significantly better than all other results (Paired T-test, p < 0.05).

We conducted labeling on both essay-level and paragraph-level sequences. The essay-level treats the whole essay as a sequence segment while the paragraph-level treats each paragraph as a segment. After labeling, the label of each changed sentence pair is marked as the purpose of the revision.

5 Experiments and Results

Our prior work (Zhang and Litman, 2014) proposed an approach for the alignment of sentences. The approach achieves 92% accuracy on both corpora. In this paper we focus on the prediction task and assume we have gold-standard sentence alignments.

The first four columns of Table 3 show the performance of baseline features with and without our new contextual features using an SVM prediction model. The last four columns show the performance of CRFs. All experiments are conducted using 10-fold (student) cross-validation with 300 features selected using learning gain ratio.

For the SVM approach, we observe that the SVM model implemented with Weka (Hall et al., 2009) and SVM yielded the best performance. The approach achieves 92% accuracy on both corpora. In this paper we focus on the prediction task and assume we have gold-standard sentence alignments.

5 Experiments and Results

Our prior work (Zhang and Litman, 2014) proposed an approach for the alignment of sentences. The approach achieves 92% accuracy on both corpora. In this paper we focus on the prediction task and assume we have gold-standard sentence alignments.

The first four columns of Table 3 show the performance of baseline features with and without our new contextual features using an SVM prediction model. The last four columns show the performance of CRFs. All experiments are conducted using 10-fold (student) cross-validation with 300 features selected using learning gain ratio.

For the SVM approach, we observe that the SVM model implemented with Weka (Hall et al., 2009) and SVM yielded the best performance. The approach achieves 92% accuracy on both corpora. In this paper we focus on the prediction task and assume we have gold-standard sentence alignments.

5 Experiments and Results

Our prior work (Zhang and Litman, 2014) proposed an approach for the alignment of sentences. The approach achieves 92% accuracy on both corpora. In this paper we focus on the prediction task and assume we have gold-standard sentence alignments.

The first four columns of Table 3 show the performance of baseline features with and without our new contextual features using an SVM prediction model. The last four columns show the performance of CRFs. All experiments are conducted using 10-fold (student) cross-validation with 300 features selected using learning gain ratio.

For the SVM approach, we observe that the SVM model implemented with Weka (Hall et al., 2009) and SVM yielded the best performance. The approach achieves 92% accuracy on both corpora. In this paper we focus on the prediction task and assume we have gold-standard sentence alignments.
indeed improve the prediction of argumentative revision types. The Ext feature set - which computes features for not only the revision but also its immediately adjacent sentences - also yields a slight (although not significant) improvement. However, adding the two feature sets together does not further improve the performance using the SVM model. The CRF approach almost always yields the best results for both corpora, with all such CRF results better than all other results. This indicates that dependencies exist among argumentative revisions that cannot be identified with traditional classification approaches.

6 Error Analysis

To have a better understanding of how the sequence labeling approach improves the classification performance, we counted the errors of the cross-validation results on Corpus A (where the revisions are more evenly distributed). Figure 3 demonstrates the comparison of errors made by SVM and CRFs.\(^{12}\)

We notice that the CRF approach makes less errors than the SVM approach in recognizing Claim changes (General-Claim, Evidence-Claim, Warrant-Claim, Surface-Claim). This matches our intuition that there exists dependency between revisions on supporting materials and revisions on Claim. We also observe that same problems exist in both approaches. The biggest difficulty is the differentiation between General and Warrant revisions, which counts 37.6% of the SVM errors and 40.1% of CRFs errors. It is also common that Claim and Evidence revisions are classified as Warrant revisions. Approaches need to be designed for such cases to further improve the classification performance.

7 Conclusion

In this paper we proposed different methods for utilizing contextual information when predicting the argumentative purpose of revisions in student writing. Adding features that captured changes in text cohesion and coherence, as well as using sequence modeling to capture revision dependencies, both significantly improved predictive performance in an experimental evaluation.

In the future, we plan to investigate whether performance can be further improved when more sentences in the context are included. Also, we plan to investigate whether revision dependencies exist in other types of corpora such as Wikipedia revisions. While the corpora used in this study cannot be published because of the lack of required IRB, we are starting a user study project (Zhang et al., 2016) on the application of our proposed techniques and will publish the data collected from this project.

Acknowledgments

We would like to thank our annotators, especially Jiaoyang Li, who contributed significantly to the building of our corpus. We also want to thank the members of the SWoRD and ITSPoke groups for their helpful feedback and all the anonymous reviewers for their suggestions. This research is funded by the Learning Research and Development Center of the University of Pittsburgh.

---

\(^{12}\)Both use models with all the features.
References


Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. Frontiers and Connections between Argumentation Theory and Natural Language Processing, Bertinoro, Italy.


