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-

Proceedings of NAACL-HLT 2016, pages 1424–1430,
San Diego, California, June 12-17, 2016.

Draft 1	Draft 2
[1] Writer Richard Louv tells us to focus more on nature through his *rhetorical questions*, parallelism, and pathos. **[2]** *Louvs rhetorical questions as us whether we value technology or nature over the other.*	**[1]** Writer Richard Louv emphasises this expanding chasm between people and nature and tries to convince people to go back to nature through his parallelism and pathos.
[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."

ture additional contextual information as text cohesion/coherence changes and revision dependencies.

As our task focuses on identifying the argumentative purpose of writing revisions, work in argument mining is also relevant. In fact, many features for predicting argument structure (e.g., location, discourse connectives, punctuation) (Burstein and Marcu, 2003; Moens et al., 2007; Palau and Moens, 2009; Feng and Hirst, 2011) are also used in revision classification. In addition, Lawrence et al. (2014) use changes in topic to detect argumentation, which leads us to hypothesize that different types of argumentative revisions will have different impacts on text cohesion and coherence. Guo et al. (2011) and Park et al. (2015) both utilize Conditional Random Fields (CRFs) for identifying argumentative structures. While we focus on the different task of identifying revisions to argumentation, we similarly hypothesize that dependencies exist between revisions and thus utilize CRFs in our task. While our task is similar to argument mining, a key difference is that the revisions do not always appear near each other. For example, a 5-paragraph long essay might have only two or three revisions located at different paragraphs. Thus, the types of previous revisions cannot always be used as the contextual information. Moreover, the type of the revision is not necessarily the argument type of its revised sentence. For example, a revision on the evidence argument can be just a correction of spelling mistakes.

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* subcategories 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[1] and allow only a single (primary) label per revision.

Category	**# in A**	**# in B**
Total	1267	1044
Claims/Ideas	111	76
Warrant/Reasoning/Backing	390	327
Evidence	110	34
General Content	356	216
Surface	300	391

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

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-

[1]We also believe that differentiating *Warrant* and *Rebuttal* revisions requires sentiment analysis.

per (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 manor 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[2].

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

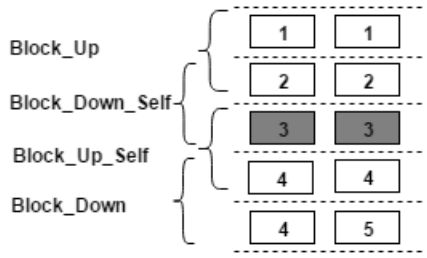

Figure 1: Example of cohesion blocks. A window of size 2 is created for both Draft 1 and Draft 2. Sequence of blocks were created by moving the window at the step of 1 (sentence).

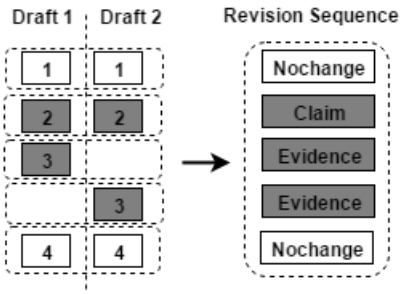

Figure 2: Example of revision sequence transformation. Each square corresponds to a sentence in the essay, the number of the square represents the index of the sentence in the essay. Dark squares are sentences that are changed. In the example, the 2nd sentence of Draft 1 is modified, the 3rd sentence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

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 ($Sim(Block\_Up, Block\_Up\_Self)$) and after ($Sim(Block\_Down, Block\_Down\_Self)$) are calculated as the cohesion/coherence scores Coh_Up and Coh_Down. The features are extracted separately for Draft 1 and Draft 2 sentences[3]. The second type describes the impact of sentence modification on cohesion and coherence[4]. Features Change_Up and Change_Down are extracted as the division of the cohesion/coherence scores of two drafts ($\frac{Coh\_Up(Draft2)}{Coh\_Up(Draft1)}$, $\frac{Coh\_Down(Draft2)}{Coh\_Down(Draft1)}$).

A bag-of-word representation is generated for

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

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

[4]The feature values of sentence additions/deletions are 0.

		SVM				CRFs			
		Base(B)	**B+Ext**	**B+Coh**	**All**	**B**	**B+Ext**	**B+Coh**	**All**
A	**P**	0.666	0.689	0.673	0.684	0.682	**0.703**∗	0.686	0.701∗
	R	0.620	0.632	0.630	0.630	0.633	**0.642**∗	0.635	0.642∗
	F	0.615	0.630	0.619	0.626	0.632	**0.644**∗	0.633	0.643∗
B	**P**	0.530	0.543	0.559 ∗	0.553∗	0.598∗	0.615 ∗	0.639∗	**0.655**∗
	R	0.516	0.525	0.534	0.532	0.518	0.524	0.532	0.532
	F	0.502	0.510	0.524 ∗	0.520∗	0.550∗	0.559∗	0.573∗	**0.584**∗

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, $p < 0.05$).

each sentence block after stop-word filtering and stemming. Jaccard similarity is used for the calculation of lexical similarity between sentence blocks. Word embedding vectors (Mikolov et al., 2013) are used for the calculation of semantic similarity. A vector is calculated for each sentence block by summing up the embedding vectors of words that are not stop-words[5]. Afterwards the similarity is calculated as the cosine similarity between the block vectors. This approach has been taken by multiple groups in the SemEval-2015 semantic similarity task (SemEval-2015 Task 1)(Xu et al., 2015).

4.2 Transforming to sequence labeling

To capture dependencies among predicted revisions, we transform the revisions to a consecutive sequence and label it with Conditional Random Fields (CRFs) as demonstrated in Figure 2. For both drafts, sentences are sorted according to their order of occurrence in the essay. Aligned sentences are put into the same row and each aligned pair of sentences is treated as a unit of revision. The "cross-aligned" pairs of sentences[6] (which does not often occur) are broken into deleted and added sentences (i.e, the cross-aligned sentences in Draft 1 are treated as deleted and the sentences in Draft 2 are treated as added.). After generating the sequence, each revision unit in the sequence is assigned the revision purpose label according to the annotations, with unchanged sentence pairs labeled as *Nochange*.

[5]We also tried the average of embedding vectors but observed no significant difference between the two approaches.

[6]Sentences in Draft 1 switched their positions in Draft 2, the cross-aligned sentences cannot be both in the same row and following their order of occurrence at the same time.

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[7].

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[8]. 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[9]. The last four columns show the performance of CRFs[10]. All experiments are conducted using 10-fold (student) cross-validation with 300 features selected using learning gain ratio[11].

For the SVM approach, we observe that the **Coh** features yield a significant improvement over the baseline features in Corpus B, and a non-significant improvement in Corpus A. This indicates that changes in text cohesion and coherence can in-

[7]Revisions on cross-aligned pairs are marked as *Surface*.

[8]Similar to settings in (Daxenberger and Gurevych, 2013)

[9]We compared three models used in discourse analysis and revision classification (C4.5 Decision Tree, SVM and Random Forests) (Burstein et al., 2003; Bronner and Monz, 2012; Stab and Gurevych, 2014) and SVM yielded the best performance.

[10]SVM model implemented with Weka (Hall et al., 2009) and CRF model implemented with CRFSuite (Okazaki, 2007)

[11]We tested with parameters 100, 200, 300, 500 on a development dataset disjoint from Corpora A and B and chose 300 which yielded the best performance.

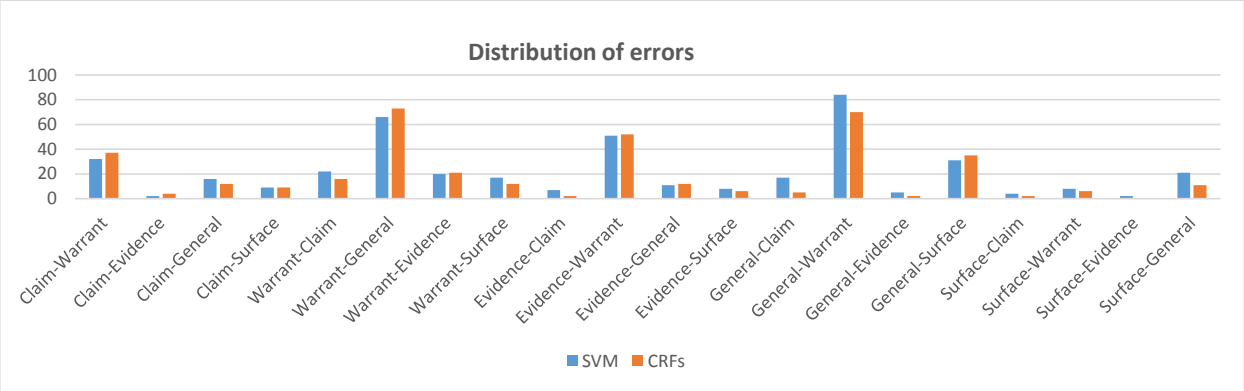

Figure 3: The number of classification errors on Corpus A, "Warrant-General" represents classifying *Warrant* as *General*.

deed 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.

[12]Both use models with all the features.

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.

References

B. Thomas Adler, Luca De Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'11, pages 277–288, Berlin, Heidelberg. Springer-Verlag.

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 356–366. Association for Computational Linguistics.

Jill Burstein and Daniel Marcu. 2003. A machine learning approach for identification thesis and conclusion statements in student essays. *Computers and the Humanities*, 37(4):455–467.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the write stuff: Automatic identification of discourse structure in student essays. *Intelligent Systems, IEEE*, 18(1):32–39.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, Washington, USA, October. Association for Computational Linguistics.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication*, pages 400–414.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996. Association for Computational Linguistics.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 273–283. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Marti A Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Sara Javanmardi, David W McDonald, and Cristina V Lopes. 2011. Vandalism detection in wikipedia: a high-performing, feature-rich model and its reduction through lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 82–90. ACM.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, Baltimore, Maryland, June. Association for Computational Linguistics.

John Lee, Chak Yan Yeung, Amir Zeldes, Marc Reznicek, Anke Lüdeling, and Jonathan Webster. 2015. Cityu corpus of essay drafts of english language learners: a corpus of textual revision in second language writing. *Language Resources and Evaluation*, pages 1–25.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM.

Joonsuk Park, Arzoo Katiyar, and Bishan Yang. 2015. Conditional random fields for identifying appropriate types of support for propositions in online user comments. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 39–44, Denver, CO, June. Association for Computational Linguistics.

Rod D Roscoe, Erica L Snow, Laura K Allen, and Danielle S McNamara. 2015. Automated detection of essay revising patterns: applications for intelligent feedback in a writing tutor. *Technology, Instruction, Cognition, and Learning*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pages 46–56.

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*.

Simone Teufel. 2000. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge University Press.

Marie M Vaughan and David D McDonald. 1986. A model of revision in natural language generation. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, pages 90–96. Association for Computational Linguistics.

Writelab. 2015. WriteLab. http://home.writelab.com. [Online; accessed 10-03-2015].

Wei Xu, Chris Callison-Burch, and William B Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*.

Fan Zhang and Diane Litman. 2014. Sentence-level rewriting detection. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 149–154, Baltimore, Maryland, June. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143, Denver, Colorado, June. Association for Computational Linguistics.

Fan Zhang, Rebecca Hwa, Diane Litman, and Huma Hashemi. 2016. Argrewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, San Diego, California, June. Association for Computational Linguistics.