Recognizing Reputation Defence Strategies in Critical Political Exchanges

Nona Naderi
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
nona@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
gh@cs.toronto.edu

Abstract

We propose a new task of automatically detecting reputation defence strategies in the field of computational argumentation. We cast the problem as relation classification, where given a pair of reputation threat and reputation defence, we determine the reputation defence strategy. We annotate a dataset of parliamentary questions and answers with reputation defence strategies. We then propose a model based on supervised learning to address the detection of these strategies, and report promising experimental results.

1 Introduction

Reputation management and defence is important in personal and professional relations. Every day, individuals, companies, and governments are faced with allegations or threats to their reputation, and they use reputation defence strategies to minimize the damage. One example in recent years was the case of Airbus Helicopters, which faced bribery allegations in a Greek NH-90 helicopter deal. In a statement, it defended its reputation using a denial strategy: *These allegations are groundless and damage the reputation of Airbus Helicopters.*[1] Maintaining good reputation is especially important in political rhetoric, and is considered as one of its primary goals. When faced with criticism, politicians use various strategies to react to it and defend themselves to others—both to their critic and to their audience. These strategies are a component of political argumentation. Recent years have seen a surge of studies that computationally analyze various aspects of arguments, such as identification of arguments (Moens et al., 2007) and analysis of argument structures (Mochales and Moens, 2008; Peldszus and Stede, 2015; Stab and Gurevych, 2014a), and identification of argumentation schemes (Feng and Hirst, 2011). Current approaches, however, have mostly ignored the interaction between the parties involved in the argumentation process, where one party is critical of the other and the other party needs to overcome the doubts.

Consider the question-and-answer sessions in Westminister-style parliamentary debates, where the government of the day is held accountable by the opposition. Opposition members ask confrontational questions, and the government ministers respond. In the face of criticism, they may use various reputation defence strategies to try to maintain a positive image.

In this paper, we propose a novel task of identifying reputation defence strategies in given dialogical argumentation. No annotated data is available for this task, so we examine whether and how reputation defence strategies are used in parliamentary debates to respond to the opposition, and create a new corpus of Canadian parliamentary debates annotated with reputation defence strategies. We focus on the most agreed-upon strategies, namely *denial, excuse, justification,* and *concession* (Benoit, 1995). For example, politicians may deny having caused a bad situation (denial) or try to evade responsibility (excuse), or promise to fix the situation (concession). Table 1(a) presents an example from the Canadian parliament, where the government minister makes an excuse for a situation, and Table 1(b) presents an example of a concession.

We then investigate what features are good predictors of the reputation defence strategies used in each case. The present work is a step towards a

[1] *Airbus Helicopters rejects bribery allegations in Greek NH-90 deal*, Reuters, 2015-03-23

Proceedings of Recent Advances in Natural Language Processing, pages 527–535,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_069

Excuse	Concession
Q. Mr. Speaker, contrary to the Conservatives' claims, we are still short 30,000 jobs to get back to the level we were at before the crisis. For example, the Quebec forestry industry, which has lost 18,000 jobs since 2005, is struggling to get out of this difficult situation. Will the government understand that the crisis is far from over in the forestry industry and that it needs a comprehensive policy to support and modernize the industry, as was the case with the auto industry in Ontario? **A.** Mr. Speaker, all of the forestry experts in the country agree that it is a matter of markets. Unfortunately, the only ones who do not get it are the members opposite. They are playing politics with these people's jobs. The markets are difficult. Our workers are among the best in the world and we will continue to support them. Billions of dollars have been put into improving green practices through the community adjustment fund, and we will continue to support the forestry industry with research and development.	**Q.** Mr. Speaker, on December 9, just a few days from now, the École de médecine vétérinaire de Saint-Hyacinthe will have to report to the American Veterinary Association on the major investments required for its full accreditation to be restored. Does the Prime Minister grasp the urgency of the situation and does he not realize that the Government of Quebec has already put $41 million into the school and that it is now time he and his government did their share? It is urgent, a matter of days. **A.** Mr. Speaker, as has been said many times, this side and the government recognize the importance of the veterinary colleges, not only the one in Quebec but in the other three provinces in this country. We will do all we can to ensure that they maintain and continue their accreditation.

Table 1: Question and answer pairs from Canadian parliamentary proceedings annotated with reputation defence strategies: (a) 2011-02-01, Robert Bouchard (Q) and Denis Lebel (A); (b) 2002-12-03, Lyle Vanclief, (Q) and Yvan Loubier (A).

deeper understanding and evaluation of (political) arguments. Natural arguments are generally enthymematic, which means some of their elements are left implicit. Identifying these implicit argument elements is a very difficult task. Knowing what strategy is used in defence arguments may help in reconstruction of these missing elements. Furthermore, extracting defence strategies can facilitate identifying contradictory and inconsistent arguments.

2 Related Work

While the task of automatically identifying reputation defence strategies has not been addressed previously, some researchers have focused on classifying the relations between argumentative components (Stab and Gurevych, 2014b; Nguyen and Litman, 2016). Others focused on classifying online discussions as agreement and disagreement with respect to a side of the debate on an issue (Abbott et al., 2011; Wang and Cardie, 2014; Rosenthal and McKeown, 2015). They employed various features, such as thread structure features, lexical (e.g., n-grams, number of words), and syntactic features (e.g., POS tags, dependency relations). Mukherjee and Liu (2013) proposed a semi-supervised generative model to extract agreement and disagreement expression types from discussion forums. Cabrio and Villata (2012) used a textual entailment approach to find pro and con arguments in a set of forum debates selected from Debatepedia.

Rosenthal and McKeown (2015) employed a supervised approach to classify forum discussions as agreement and disagreement and found that similar lexical and syntactic structures in a pair of posts were important for the classification task. Biran and Rambow (2011) used discourse markers to classify single sentences as a justification of a claim or not. Peldszus (2014) focused on identifying attack and support relations in microtexts. However, none of these looked at the interactions between arguments of two parties. Here, we aim to analyze these interactions, particularly when one party criticized the other, and the other addresses the criticism.

3 Data

For our analysis, we focus on pairs of questions and answers extracted from Oral Question period from Canadian parliamentary proceedings. The purpose of questions asked in Oral Question period is to hold the government accountable for its actions[2]. While both government backbenchers and opposition members ask questions during this period, the questions asked by opposition members are more confrontational than the questions asked by the backbenchers. The questions asked by government backbenchers tend to be more clarification questions; therefore, we extracted the pairs where the questions were asked by opposition members.

[2] http://www.ourcommons.ca/About/Compendium/Questions/c_g_questions-e.htm

Reputation defence strategies
Denial: 1. The government denies that the situation in question occurred. 2. The government denies causing the situation in question.
Excuse (evading responsibility): 1. The situation in question occurred in response to some other situations. 2. The situation in question occurred because of lack of information or control over important factors. 3. Some accidents caused the situation. 4. The motives or intentions of the government were good.
Justification (reducing offensiveness): 1. The government tries to increase positive feeling towards it (for example by mentioning positive actions the government performed in the past). 2. The government tries to convince the audience that the situation is not as bad they say. 3. The government tries to distinguish the situation in question from similar but less desirable situations. 4. The government tries to place the situation in a different or broader context. 5. The government attacks the opposition or questions their credibility. 6. The government offers compensation for the situation.
Concession (corrective actions): 1. The government promises to restore the situation to what it was before. 2. The government promises to make changes (for example to prevent the recurrence of the situation).
None of these strategies

Table 2: Conditions for each reputation defence strategy.

Q. Mr. Speaker, I would like the Minister of Public Safety and Emergency Preparedness to tell that to over 23,000 women who in 2003 were sexually assaulted or raped, and whose lives will never be the same again. Even more, I would like the minister to explain to these women why our prison libraries include pornographic magazines. Will the minister explain why our prison libraries feel it is necessary to provide pornographic material to violent sex offenders?
A. Mr. Speaker, as I just said, and maybe the hon. member did not hear me, I want to assure her that strict controls are in place to restrict access to any material that could be considered demeaning, could jeopardize the safety of any individual or the institution, is sexually violent or involves children or could be detrimental to the offender's treatment. We take the safety of our correctional institutions very seriously.

Table 3: Disagreement among three annotators, annotated variously as *denial, justification,* and *concession*; 2005-05-30, Lynne Yelich (Q) and Anne McLellan (A).

To study whether reputation defence strategies are used in the parliamentary debates, we first ran a pilot study and asked three expert annotators to annotate 100 random pairs of the extracted questions and answers with one of the reputation strategies or none of the strategies. We prepared detailed guidelines to describe the conditions that need to be satisfied for choosing each reputation defence strategy. Table 2 presents the conditions provided to the annotators (all are adapted from Benoit (1995)).

We further conducted a larger annotation study with 1500 random pairs of the extracted questions and answers on the crowd-sourcing platform (CrowdFlower[3]). Contributors were shown a question and answer pair from the parliamentary debates on various issues, and were asked to choose which strategies (based on the conditions presented in Table 2) had been used by the government in response to criticism. We asked for at least three annotations per pair from the English-speaking countries. To maintain the annotation quality, we allowed only the highest-quality contributors to participate, and also included some test pairs. On each page, each participant was presented with one test pair and three other pairs, and had to maintain 70% accuracy throughout the job. In total, we included 56 test questions for 1500 pairs. Each response was paid \$0.04. Only 10% of the question and answer pairs were annotated with *none of the strategies* by the annotators, which shows that these strategies can represent the data reasonably well. Almost 70% of the pairs were agreed upon by two or more annotators, but in order to obtain a more reliable corpus, we accepted the pairs for which at least three annotators agreed on a single answer, and discarded the pairs where fewer than three annotators agreed. For the expert annotations, three annotators achieved full agreement on a single answer for 32 pairs. In total, the

[3]https://www.crowdflower.com/

Verb type	Examples
Concealment	conceal
Psych	amuse, admire
Desire	want, long
Judgment	judge, approve
Assessment	estimate
Searching	investigate
Social interaction	correspond, meet
Communication	inquire, advise
Existence	exist, survive
Aspectual	begin, continue
Allow	allow, permit
Admit	admit
Succeed	succeed

Table 4: VerbNet classes that we used.

LIWC category	Examples
Analytic	–
Negations	no, not
Interrogatives	how, what
Affective processes	happy
Positive emotions	nice
Negative emotions	hurt
Cognitive processes	cause
Insight	think
Causation	because
Tentative	perhaps
Certainty	always
Perceptual processes	heard
Achievement	success
Power	superior
Past focus	talked
Present focus	is
Future focus	will
Assent	agree

Table 5: LIWC features that we used.

reliable crowd and expert annotations resulted in a set of 493 pairs, of which 170 were annotated as *denial,* 36 as *excuse*, 173 as *justification,* 95 as *concession,* and 19 as *none of these strategies.* The average number of tokens in each pair is 171, with the longest pair being 356 words. These pairs of questions and answers are on different topics.

We further examined the discarded pairs of questions that were not agreed upon by at least three annotators to investigate the source of disagreements. Disagreements between the annotators were generally due to the use of multiple strategies or vague answers that do not contribute to the goal of the dialogue; they simply look like relevant answers, but they do not really address the questions. Table 3 shows an example of disagreement between three annotators.

Q. Mr. Speaker, contrary to what the Prime Minister says, Canada's actions so far lead us to conclude that it is siding with the United States by supporting, through its silence, comments made by U.S. Secretary of Defense, Donald Rumsfeld, who wants to ignore NATO and the UN if it suits his purposes. Is the Prime Minister aware that his silence is contributing to undermining international institutions and that this complacent attitude breaks with Canada's tradition of respecting major international institutions? **A. Mr. Speaker, I firmly reject the suggestion that the Prime Minister has been silent.** Our position is clear. We have always encouraged and supported an approach that goes through the United Nations and through the Security Council. We have gotten here, in some measure, thanks to the efforts of the Prime Minister. He has never been silent, he has been active on the international scene and we are very proud of what he has done.

Table 6: An example *Comparison* relation between two parts of question and answer, specified in bold; 2003-02-12, Francine Lalonde (Q) and Bill Graham (A).

4 Approach

We formulate the task as a classification task. Given a question and answer pair, we identify which of the four reputation defence strategies, *denial*, *justification*, *excuse*, and *concession* is used in the answer. In order to capture the characteristics of each strategy, we explore two classes of features: features that are based solely on the answers, and features that describe the relation between the question and the answer.

4.1 Features from Answers

VerbNet Classes Certain verb classes can indicate defence strategies; for example, *assure* is often used in *justification* or *concession* strategies, e.g., *I want to assure the House that we are taking measures.* To this end, we use the VerbNet lexicon (Schuler, 2005), which groups verbs by their shared semantic meaning and syntactic behavior. Table 4 shows the verb classes that we use. We use the count of verb class occurrences as features.

Positive and Negative Sentiments and Emotions Motivated by the conditions for the *justification* strategy (Table 2), we examined the positive and negative sentiments and emotions expressed in the answers. Emotions are extracted using Linguistic Inquiry and Word Count (Tausczik and Pennebaker, 2010), and sentiments are extracted using OpinionFinder (Wilson et al., 2005).

Features	Acc.(%)	F_1 (%)
Majority Class (justification)	36.50	–
Production rules	49.78	46.31
Unigrams (q + a) (tf-idf)	52.53	49.54
Unigrams (a) (tf-idf)	53.35	51.32
Unigrams (a) (tf-idf) + LIWC	53.57	53.07
Unigrams (a) + VerbNet v class	53.78	51.62
Unigrams (a) + VerbNet v class + Sentiments	56.11	54.02
Unigrams (a) + VerbNet v class + Sentiments + Negation	56.33	55.55
Unigrams (a) + Discourse + Similarity	55.26	53.04
Unigrams (a) + VerbNet v class + Sentiments + Negation + Discourse	56.96	56.33
Unigrams (a) + VerbNet v class + Sentiments + Negation + Discourse + Similarity (best model)	**57.59**	**56.92**

Table 7: The performance of different models for classification of four reputation defence strategies (five-fold cross-validation).

Features	Denial	Excuse	Justification	Concession
Production rules	59.4	0.0	51.8	30.8
Unigrams (q + a) (tf-idf)	62.6	10.0	55.6	28.2
Unigrams (a) (tf-idf)	62.4	13.6	55.6	36.4
Unigrams (a) (tf-idf) + LIWC	64.0	**19.4**	54.2	41.0
Best model	**65.0**	18.0	**59.8**	**48.0**

Table 8: Average F_1 of different models for classification of four reputation defence strategies (five-fold cross-validation).

Past and Future Focus Verb tense can reveal the difference between strategies; for example, in *denial*, the focus is more likely to be on the past, e.g., *as I said in French, I never gave advice about the privatization of the Toronto airport*, whereas in *concession,* the focus tends to be on the future, e.g., *I promise the hon. member and all members of the special forces that I will work with them to ensure they are justly and properly treated*.

Negation *Denials* tend to be expressed using *never*, *not*, *no*, *nobody*, and *none*, e.g., *I never solicited funds*.

Insight and Achievement These categories are mostly associated with *justification* strategies, e.g., *I think when we can help farmers in Canada, it is our duty to do so*, and *We will continue to invest in this fashion. It is a proven success.* To compute these features, we use Linguistic Inquiry and Word Count (LIWC), a tool that counts occurrences of words by their psychological categories. We used 18 LIWC categories, presented in Table 5.

4.2 Features Describing Relations between a Question and Answer Pair

Discourse Relations Discourse relations have been shown to be effective in identifying support and attack relations in persuasive essays (Nguyen and Litman, 2016). While Nguyen and Litman (2016)'s work focused on only the attack and support relations between argumentative components in a paragraph, nonetheless, we believe that discourse relations can be informative features for identifying reputation defence strategies. Here we use shallow discourse relations (*Class level*), including *Comparison*, *Contingency*, and *Expansion* between the question and answer pairs (extracted using End-to-End PDTB-Styled Discourse Parser (Lin et al., 2014)).[4] For example, consider the question and answer pair in Table 6, where the discourse relation (parts in bold) between the question and answer is *Comparison* and indicates the *denial* strategy. While fine-grained discourse relations (*type* level) can be informative for identifying reputation strategies, for our analysis, we focused on only major classes of discourse relations because discourse parsers usually yield less reliable results for fine-grained relations.

Syntactic Production Rules Stab and Gurevych (2014b) used production rules to classify support and non-support argument relations in persuasive essays, and found them to be effective features. Their work also focused on

[4] *Temporal* relations have not been effective in our classification task, which is also in line with expectations (Biran and Rambow, 2011; Stab and Gurevych, 2014b).

	Features	Denial		Justification		Concession	
		Acc(%)	F_1(%)	Acc(%)	F_1(%)	Acc(%)	F_1(%)
Justification	Best model	**74.35**	**74.74**			70.51	69.14
	BOW + LIWC	72.59	72.49			66.39	64.51
	BOW + VerbNet	70.85	70.79			**70.87**	**69.28**
	BOW + VerbNet + Sent + Neg	72.89	72.72			69.39	68.01
	BOW + Discourse + Similarity	73.18	73.04			67.15	65.48
	Production rules	67.95	67.80			65.32	63.43
	Majority	50.44	–			64.55	–
Concession	Best model	76.23	76.40	70.51	69.14		
	BOW + LIWC	**77.36**	**76.72**	66.39	64.51		
	BOW + VerbNet	75.09	74.52	**70.87**	**69.28**		
	BOW + VerbNet + Sent + Neg	**76.98**	**76.91**	69.39	68.01		
	BOW + Discourse + Similarity	75.85	75.16	67.15	65.48		
	Production rules	76.98	75.90	65.32	63.43		
	Majority	64.15	–	64.55	–		
Excuse	Best model	83.02	81.69	82.31	78.16	66.35	64.74
	BOW + LIWC	84.43	80.28	83.74	79.15	71.68	67.93
	BOW + VerbNet	82.98	78.89	83.28	78.72	68.60	66.15
	BOW + VerbNet + Sent + Neg	81.57	80.25	81.84	77.85	68.60	66.81
	BOW + Discourse + Similarity	84.43	79.91	83.26	78.28	71.71	66.88
	Production rules	82.00	75.01	83.29	76.98	71.71	64.80
	Majority	82.52	–	82.78	–	72.51	–

Table 9: The performance of the models for pairwise classification (five-fold cross-validation). Best model includes discourse relations, cosine similarity, unigrams, verb classes, negations, and positive and negative sentiments in the answers.

the relations in a paragraph. Here, we explore the impact of the production rules in capturing the syntactic characteristics of reputation management strategies. We consider binary features for production rules (e.g., VP → VBZ NP SBAR, VP → VB NP PP) that appear only in the answer, and both in the question and the answer (Lin et al. (2009) and Feng and Hirst (2012) used these features for identifying shallow discourse relations and RST discourse relations, respectively). We used the Stanford parser (Klein and Manning, 2003) to perform the pre-processing.

Similarity Measures Simple lexical similarity methods have been shown to be robust in recognizing textual entailment, which can help capture strategies such as *denial* and *concession*. We compute the average semantic similarity between the question and the answer sentences from the cosine similarity between their vectors. To represent the questions and answers, we sum their word2vec embeddings (Mikolov et al., 2013).

5 Results

The classification is performed using a class-weighted Support Vector Machine model with a linear kernel[5]. The classifiers were trained and tested with the crowd-sourced data described in section 3 using five-fold cross validation. The baselines that we use are the majority class, where all instances are classified as *justification*, and the bag-of-words representations (weighted using *tf-idf*) of the question and answer pairs and the bag-of-words representations of answers. The bag-of-words representation of answers is the strongest baseline on our dataset and yields an accuracy of 53.35%. To determine the efficacy of the features, we train individual classifiers on the feature classes. The results are reported in terms of accuracy and average F_1-measure.

Multi-class Classification Table 7 reports the results for multi-class classification. The best performance was 57.59% accuracy, which was achieved by using discourse relations and cosine similarity between the question and answer, and verb classes, positive and negative sentiments (extracted using OpinionFinder), negations, and the unigrams from the answers. This model yields a 20-point improvement over the majority baseline and at least a 4-point improvement over bag-of-words baselines. Our ablation studies to measure the contributions of different components show that all features are helpful, with verb classes, sentiments, negations, and unigrams (from answers)

[5]LibSVM implementation (Pedregosa et al., 2011).

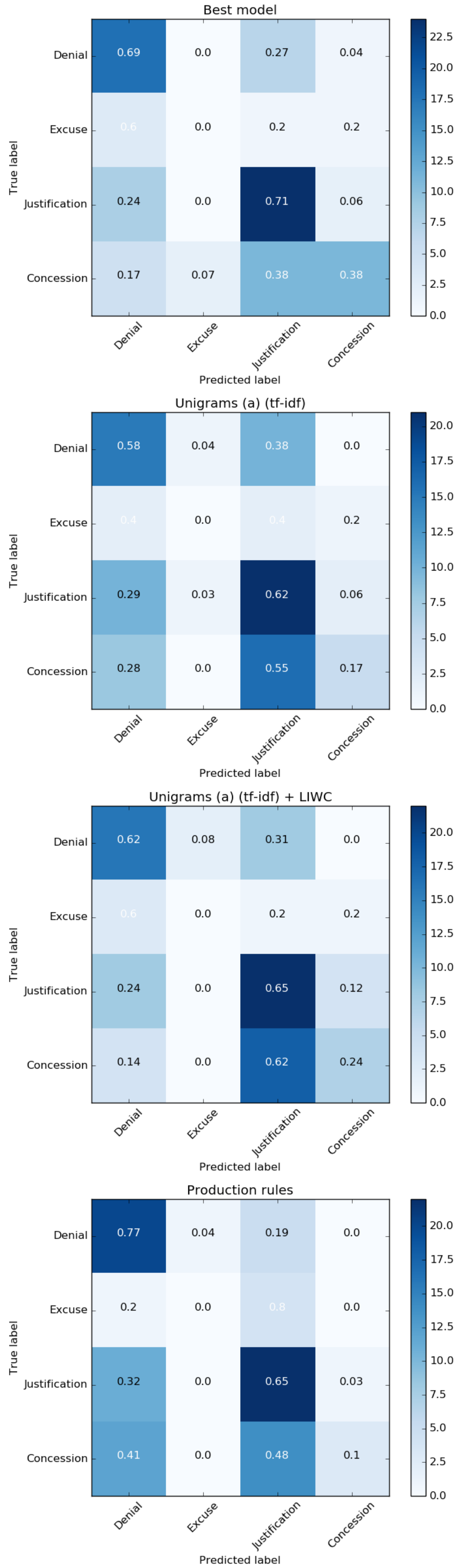

Figure 1: Normalized confusion matrices for reputation defence classification.

being the most helpful for distinguishing between strategies. Using the LIWC features also improves the performance over all the baselines. While production rules are informative features, the performance of this classifier is lower than the bag-of-words baseline.

Table 8 reports the average F_1-measure of five-fold cross validation for each reputation defence strategy in multi-class classification. The best performance for identifying *denial, justification,* and *concession* is achieved by the best model. LIWC features are most informative for identifying *excuse* strategy.

Pairwise Classification We further experimented with *pairwise* classification (one-versus-one) for the six possible pairings of the four strategies to find the most informative features for each strategy (Table 9). For each of the six classifiers, we considered the data for the two strategies against each other. In *pairwise* classification, almost all models improve over the majority baseline, except for *excuse*, for which the training data is very small. In distinguishing between *denial* and *justification*, the combination of verb classes, sentiments, negations, discourse relations, cosine similarity, and unigrams from the answers yields the best performance. The most informative features in distinguishing *concessions* and *justifications* are VerbNet classes. In distinguishing between *denial* and *concession,* the features extracted from the answers contribute the most.

Reputation Defence Errors Figure 1 shows confusion matrices for the best model, the baseline unigram (a) model, LIWC model, and production rule model for the first fold of cross-validation. The most common confusion is misclassifying the *concession* strategy as the *justification* strategy. The best model makes this error less often. Production rules often misclassify the *concession* strategy as the *denial* strategy as well.

6 Discussion

The results show that the features proposed above are successful in distinguishing *denial* and *justification* strategies, but the small training set for *excuse* and *concession* strategies did not allow the model to effectively detect these strategies. While the performance of the model can benefit from more training data, the limited performance could

Q. Mr. Speaker, Canadians are being prevented from obtaining their passports under the guise of increased national security. In the last six months my constituency office has been inundated by hundreds of angry constituents. Some have even been forced to cancel trips, costing them thousands of dollars, due to the incompetence of the government. I have repeatedly raised their concerns with the passport department of foreign affairs to no avail. When the advertised processing time is 45 working days, why are my constituents waiting months for their passports?
A. Mr. Speaker, the hon. member was good enough in the introduction to his question to point out there is a problem in terms of new security measures and there is a great deal of increased flow of demands for passports. The passport office is making a serious and concerted effort to respond to these requests. I regret any inconvenience to the hon. member or to Canadian citizens. I want to assure the House that we are taking measures. We have brought in people this weekend and we will be working around the clock to reduce and eliminate the backlog of requests. We have put in measures to enable people to get their passports more quickly and to deal with it more efficiently. I will be circulating to the hon. member, and all members, statements as to how the department is responding to this.

Table 10: An example of the *justification* strategy used together with the *concession* strategy; 2003-02-12, Andy Burton (Q) and Bill Graham (A).

be also due to the labeling task. By limiting the crowd annotators to choose the most prominent strategy, we attempted to study the characteristics of each strategy in isolation, but the results of the annotation process and classification task show that some defence strategies can be employed in combination with each other. Table 10 shows an example from our corpus that was misclassified by the model as the *concession* strategy, and when we examined the pair[6], we observed that although the main strategy in the defence is *justification* to reduce the offensiveness, corrective actions are further offered (the *concession* strategy).

Moreover, some questions express multiple reputation threats, which may require multiple defence strategies to address the threats. These cases require further analysis of the reputation threats and allegations. We chose parliamentary debates to study reputation defence strategies because reputation threat and defence arguments are more numerous in this data, and the data is easily accessible.

[6]Three annotators marked this relation as *justification* and one annotator marked it as *concession*, we considered agreement by three annotators as gold.

7 Conclusion

We have addressed a new task of automatically identifying reputation defence strategies. While reputation defence strategies are used in various social settings and managing reputations against attacks is vital for any individual, in parliamentary settings, they impact decision making as well. Thus, we computationally analyzed reputation defence strategies in parliamentary speeches. We also created a corpus for analysis of reputation strategies. We explored various features for classifying four reputation defence strategies. Our results show that while the models benefit most from the features extracted from the defence, they can be improved using the features that capture the relation between a threat and defence pair. Our promising results suggest a new research direction and allow for a better understanding of political exchanges and large-scale analysis of participant behaviors.

Acknowledgments

This research is financially supported by the Natural Sciences and Engineering Research Council of Canada. We thank Patricia Araujo Thaine and Simon Emond for their contributions to the pilot study of annotations. We thank Gerald Penn, Suzanne Stevenson, and Frank Rudzicz for their valuable feedback, as well as the anonymous reviewers for their suggestions.

References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 2–11.

William L Benoit. 1995. *Accounts, Excuses, and Apologies: A Theory of Image Restoration Strategies*. State University of New York Press, Albany.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*. IEEE Computer Society, Washington, DC, USA, pages 162–168.

Elena Cabrio and Serena Villata. 2012. Natural language arguments: A combined approach. In *Proceedings of 20th European Conference on Artificial Intelligence*. IOS Press, Amsterdam, The Netherlands, pages 205–210.

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*. Association for Computational Linguistics, pages 987–996.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jeju Island, Korea, pages 60–68.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 423–430.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 343–351.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering* 20(2):151–184.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Raquel Mochales and Marie-Francine Moens. 2008. Study on the structure of argumentation in case law. In *Proceedings of the 2008 Conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*. IOS Press, Amsterdam, The Netherlands, pages 11–20.

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*. ACM, New York, USA, pages 225–230.

Arjun Mukherjee and Bing Liu. 2013. Discovering user interactions in ideological discussions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, (Volume 1: Long Papers)*. pages 671–681.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1127–1137.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Andreas Peldszus. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 938–948.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, pages 168–177.

Karin Kipper Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Christian Stab and Iryna Gurevych. 2014a. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Doha, Qatar, pages 46–56.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Doha, Qatar, pages 46–56.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1):24–54.

Lu Wang and Claire Cardie. 2014. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Baltimore, Maryland, pages 97–106.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 347–354.