On the Retrieval of Wikipedia Articles Containing Claims on Controversial Topics

Haggai Roitman, Shay Hummel,
Ella Rabinovich, Benjamin Sznajder
IBM Research
Haifa, Israel
{haggai,hummel,ellak,benjams}@il.ibm.com

Noam Slonim, Ehud Aharoni
IBM Research
Haifa, Israel
{noams,aehud}@il.ibm.com

ABSTRACT

This work presents a novel claim-oriented document retrieval task. For a given controversial topic, relevant articles containing claims that support or contest the topic are retrieved from a Wikipedia corpus. For that, a two-step retrieval approach is proposed. At the first step, an initial pool of articles that are relevant to the topic are retrieved using state-of-the-art retrieval methods. At the second step, articles in the initial pool are re-ranked according to their potential to contain as many relevant claims as possible using several claim discovery features. Hence, the second step aims at maximizing the overall claim recall of the retrieval system. Using a recently published claims benchmark, the proposed retrieval approach is demonstrated to provide more relevant claims compared to several other retrieval alternatives.

1. INTRODUCTION

Discussions about some **controversial topic** may take place in many real-life situations such as in politics, marketing, law and healthcare. During such a discussion, each participant who wishes to convince others about her position on the topic is expected to provide one or more good persuasive **arguments**. Each argument should be supported by one or more relevant **claims** [7, 18].

A plausible claim may be roughly defined as a **concise** (and general enough) **statement** that directly **supports** or **contests** the discussed topic [11]. A claim may range from a factual assertion to an opinionated statement [11]. Table 1 contains an example (borrowed from Levy et al. [11]) of a single argument on a given controversial topic, having two plausible claims (one factual and one opinionated) and one "invalid" claim (a rephrase of the original argument).

Providing sufficient plausible claims for supporting or contesting an argument during a discussion is extremely hard even for humans. Therefore, argumentation mining [14] has recently attracted the attention of many researchers, aiming at augmenting human argumentation capabilities with automatic methods. Such methods include, among others, methods for automatic detection of claims in text [3, 11].

Recently, Levy et al. [11] have suggested a method for automatic **detection** of context (topic) dependent claims in text. The input to Levy et al.'s method is a set of Wikipedia articles, assumed to contain relevant claims which need to be detected [11]. Wikipedia, the popular online encyclopedia, has been utilized so far as a knowledge source for many computational linguistic tasks [13]. To be useful for claim detection purposes, it would be desired that every input Wikipedia article contained as many relevant claims to the topic in context as possible. Higher claim coverage within such articles, in its turn, may allow to improve the performance of automatic claim detection methods by introducing less noise into the detection process [11].

While the potential of Wikipedia for claim detection is clear, existing claim detection methods strongly assume the availability of a (small enough) corpus which contains text documents (articles[1]) with relevant claims to a given topic [11]. Yet, automatic claim detection methods may not scale well in the presence of a large sized text corpus (such as Wikipedia), where an exhaustive analysis of the whole corpus for claims would be extremely inefficient. Therefore, a preliminary step of effective **document retrieval** over a large corpus that may result only in a small sub-set of "focused" documents with high potential to contain relevant claims would be desired [11]. While some Wikipedia articles are already manually annotated as "disputable", "controversial" or marked as "point-of-view" (POV) articles, discovering relevant claims solely by focusing on such articles may only provide a partial solution, as shall be demonstrated in this work. Many relevant claims may actually reside within articles that have no controversy issues.

Aiming to address this novel retrieval challenge, this work focuses on the retrieval of Wikipedia articles that contain claims relevant to a given controversial topic. The proposed articles retrieval scheme is based on a two-step approach. At the first step, an initial pool of articles relevant to the topic are retrieved using state-of-the-art retrieval methods. At the second step, articles in the initial pool are re-ranked according to their potential to contain as many relevant claims as possible using several claim discovery features. Hence, the second step aims at maximizing the overall claim recall of the retrieval system.

The contributions of this work can be summarized as follows:

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2891115 .

[1] The terms "document" and "article" are used interchangeably throughout this work.

Argument: “*The sale of violent video games to minors should be banned*”
Factual claim: “*Violent video games can increase childrenâĂŹs aggression*”
Opinionated claim: “*Video game publishers unethically train children in the use of weapons*”
Invalid claim: “*Violent video games should not be sold to children*”

Table 1: An example argument with two plausible claims and one invalid claim.

- A definition of a novel claim-oriented document retrieval task.
- The proposal of a retrieval solution over Wikipedia which aims at maximizing claim-recall.
- An evaluation of the solution using a recently published claims benchmark, demonstrating the ability of the retrieval solution to provide articles with more relevant claims compared to several other retrieval alternatives.

The rest of this paper is organized as follows. Section 2 reviews related works. Section 3 presents the novel claim-oriented document retrieval solution, which is then evaluated in Section 4. Finally, Section 5 concludes and suggests some directions for future work.

2. RELATED WORK

The task of claim-oriented document retrieval addressed in this work is strongly related to several computational linguistics tasks that depend on the quality of their set of input articles. More specifically, for a given text and possibly also an argument to satisfy, claim detection methods, which aim at extracting textual fragments (e.g., sentences) that may contain claims, are the most relevant to this work [3,11,14]. Among these works, only the recent work by Levy et al. [11] actually addresses the detection of claims related to a given topic, termed “*contex-dependent claim detection*” (CDCD) [11]. As a preliminary step, CDCD methods expect as input, a pre-retrieved (small) “focused” collection of relevant documents to a given topic [11]. Therefore, the claim retrieval solution described in this work can be viewed as a preceding and important step to any CDCD method.

The claim-oriented document retrieval task shares some relationship with two other retrieval tasks, namely opinion retrieval [12] and focused retrieval [8]. In an opinion retrieval setting, documents that are both relevant and opinionated about the query topic are required to be retrieved by the retrieval system [12]. Though, the task addressed in this work differs from the opinion retrieval task. As was mentioned in Section 1, a plausible claim is not only restricted to have an opinion “flavor” on a given topic, but may also have a factual “flavor” (see again the examples in Table 1). Therefore, compared to the opinion retrieval task, it is possible to return an article that contains only factual claims about a given topic as long as they support (or contest) the argument (topic).

Works on focused retrieval, such as those on question answering [9], passage retrieval [4] and element (XML) retrieval [2] have been suggested to provide a more holistic retrieval process, where answers to user queries are provided in a more fine granular form (e.g., a snippet, passage, etc). This in turn, is supposed to reduce the time a user needs to scan through answers within retrieved documents.

Among such works, the *Prove It!* task of the INEX Social Book Search Track [10] is the most relevant to this work. Given some **factual** statement, the *Prove It!* task evaluated various focused retrieval approaches for searching for book **pages** that may confirm or refute the statement [10].

While the *Prove It!* task shares some resemblance with the current retrieval task, it differs from the later in two important aspects. First, and most distinctive, the basic answer form in the *Prove It!* task is an **evidence** (a book page) about some fact, while in this work the basic answer form is a claim contained within a retrieved article. To remind, a claim needs to be as concise (a single short and general enough sentence [1, 11]) about the topic in mind as possible. Therefore, the definition of a fact by the *Prove It!* task is actually similar to the definition of a claim in this work [1, 7, 11, 18]. Hence, the two tasks only complement each other, where additional evidence on claims (facts) that are found by this work for a given topic (after applying some CDCD method on the retrieved claim-containing articles) may be further discovered using various evidence retrieval methods, such as the ones that were studied in the context of the *Prove It!* task.

Second, while the *Prove It!* task targets on a high **document-level precision**, the claim retrieval task targets on a high **claim-level recall**, strongly motivated by the CDCD methods in mind.

An additional line of works worth mentioning in the context of this work, are works on controversy detection, and more specifically works that detect controversial topics in Wikipedia [6, 19]. While such works may be used to enhance existing manual controversy annotation in Wikipedia, focusing the retrieval solely on controversial Wikipedia articles provides a much less effective solution to the task, as shall be demonstrated in this work.

To the best of our knowledge, this work is the first to address the **claim-oriented document retrieval** task.

3. SOLUTION

The goal of the claim-oriented document retrieval task is to retrieve documents (Wikipedia articles) that have the potential to contain as many relevant claims to a given controversial topic as possible. Therefore, a retrieved document's quality is judged relatively to the amount of relevant claims it contains.

3.1 Overall approach

A two-step document retrieval approach is now proposed for the claim-oriented document retrieval task. At the first step, an initial pool of Wikipedia articles that are “relevant” to the (controversial) topic in mind are retrieved using state-of-the-art retrieval methods (Section 3.2). The goal of this step is, therefore, to obtain documents that are as much focused on the topic as possible, assuming that such documents may have a better chance for containing claims about the topic. At the second step, articles in the initial pool are re-ranked according to their potential to contain as many relevant claims as possible (Section 3.3). Such claim-oriented document re-ranking is based on a combination of several claim features. Hence, the second step aims at maximizing the overall **claim recall** of the retrieval system.

3.2 Step 1: Topic-based document retrieval

The goal of the first step is to retrieve a pool of documents that are as much focused on a given topic as possible. Hence, the retrieval system takes a topic as an input query q and returns the k highest ranked ("most relevant") articles D_q in the (Wikipedia) corpus D according to q. Let $score_{topic}(d;q)$ denote the score of a document $d \in D$ given q.

As a preliminary step, each query q is "enhanced" using several methods, as follows. First, the query is expanded with bigram terms, obtained by considering every sequence of two adjacent query terms to be a bigram.

Next, (expanded) query q terms (unigrams and bigrams) are weighed using the *Search Result Overlap* (SROR) method of Song et al. [17]. According to this method, each query term $t \in q$ is weighed according to its predicted query drift, measured as the relative overlap between the query q's search results over Wikipedia corpus D when term t is once included in q and once excluded from q [17]. The smaller the overlap is, the more query drift is expected when term t's is absent from q; hence, term t is more important for encoding the topic expressed by query q [17]. Finally, the query is further expanded with lexical affinities using Carmel et al.'s [5] method, which was found to be quite effective for this topic-based retrieval step.

Given the result of query q enhancement, $score_{topic}(d;q)$ is calculated as follows:

$$score_{topic}(d;q) = \sum_{o \in \{T,B,FP,LA\}} \lambda_o score_o(d;q) \qquad (1)$$

$score_T(d;q)$, $score_B(d;q)$, $score_{FP}(d;q)$, $score_{LA}(d;q)$, are four score components that represent the document's score by using only its title, (main) body, first paragraph and the score of a lexical affinity only query over its (main) body, respectively. In this work, the Okapi-BM25 method [16] has provided the best scoring implementation of each component $score_o(d;q)$ ($o \in \{T,B,FP,LA\}$). λ_o ($o \in \{T,B,FP,LA\}$) further represents the relative importance (boost) of each component in the final topic-based document score.

3.3 Step 2: Claim-oriented document re-ranking

Recall that, the output of the first topic-based retrieval step is a list of k documents D_q that are presumed to be as much focused on the topic expressed by query q as possible. Let C_q denote the set of claims relevant to (topic) query q. The goal of the second step is to re-rank the documents in D_q (denoted D_q^{rerank} hereinafter) so to have documents with more relevant claims $c \in C_q$ ranked higher in the list. The logic behind such claim-oriented document re-ranking is to allow any CDCD method to consume fewer documents as an input (say $m = \sqrt{k}$), yet with a high chance to contain relevant claims (and in turn, obtain a better claim detection accuracy [11]).

The proposed claim-oriented document re-ranking is based on a combination of several different document scorers $score_f(d;q,D_q)$, each scorer assigns a score to every document $d \in D_q$ based on some claim-discovery feature f. The set of features examined in this work is based on a preliminary domain study that was conducted using a random sample of documents that contained relevant (and irrelevant) claims to a given seed of example controversial topics (details about such topics are described later on in Section 4.1). The followings are, therefore, some basic observations about the nature of (relevant) claim features which guided the design of the automatic claim-oriented document re-ranking approach in this work.

Observation 1 (O1) Usage of the existing manual annotation of Wikipedia articles such as "*disputable*", "*controversial*" or "*POV*" should be carefully done. While relevant claims to a given topic may indeed be contained in "controversial" articles, in some cases, preferring such articles may actually turn up to be a pitfall for the retrieval system. Relevant claims about a disputable or controversial topic may, in many cases, be contained in articles that were not manually annotated as having controversial issues.

Observation 2 (O2) Articles that contain claims tend to include "controversy" related terms (e.g., "*dispute*", "*criticism*", etc) that may hint about the existence of controversial issues with the topic in mind.

Observation 3 (O3) Sentences that contain claims tend to include an "evidence" in the form of one or more citations to internal (i.e., `[[...]]` outlinks to other Wikipedia articles) or external sources (Wikipedia's `<ref>...</ref>` references) related to the topic in mind.

Observation 4 (O4) Sentences that contain claims tend to include a "conjugated-that" expression (e.g., "*...claim that..*", "*...argue that...*", etc) **prior** to any mention of topic related terms.

Focusing on the above four observations as guidelines for our claim-discovery feature engineering, Table 2 describes the complete list of features on which documents in the initial retrieved list D_q are scored for claim-oriented document re-ranking.

Feature f1 holds the topic-based score of each document in D_q that was obtained in the first retrieval step. To address **O1**, feature f2 was designed, indicating whether an article $d \in D_q$ is annotated in Wikipedia as "controversial" or not.

To address **O2**, features f3-7 were designed to capture various topic-independent (general) and topic-dependent (context)"controversy" aspects of Wikipedia articles in D_q. For that, a manually crafted "Controversy Lexicon" CL was created (during the preliminary domain study) using a seed list with several controversy related terms (e.g., "*criticism*", "*dispute*", etc). This seed list was further expanded with synonyms using a thesaurus. Table 3 contains the complete list of "controversy" terms that were used in this work.

Given an article $d \in D_q$, its general "controversy relevance" score is obtained by measuring the TF-IDF similarity between the CL terms to three different textual parts of the article, namely, its title, (main) body, and headers[2]. The contextual (topic) "controversy relevance" score is further obtained by measuring the relative proximity of the CL terms to the topic q terms in each article. The proximity score is calculated proportionally to the distance (up to some maximum "window") between a give pair of (CL,topic) terms[3].

[2] Having terms like "*Criticism*", "*Controversy*" in an article's header has a high potential for having relevant claims in the succeeding section.

[3] The actual proximity scores in this paper are based on Apache Lucene's `SpanQuery` API.

	Feature	Description
f1	`Article_Topic_Relevance`	Articles's topic score: $score_{topic}(d; q)$
f2	`Article_Has_Cntrv_Annotation`	1 if article includes some "disputable", "controversial" or "POV" annotation, otherwise 0.
f3	`Content_Cntrv_General_Sim`	Article's content general similarity to "Controversy Lexicon"
f4	`Title_Cntrv_General_Sim`	Article's title general similarity to "Controversy Lexicon"
f5	`Headers_Cntrv_General_Sim`	Article's headers general similarity to "Controversy Lexicon"
f6	`Content_Cntrv_Context_Sim`	Article's content contextual (topic-based) similarity to "Controversy Lexicon"
f7	`Title_Cntrv_Context_Sim`	Article's title contextual (topic-based) similarity to "Controversy Lexicon"
f8	`Reference_Proximity_To_Topic_Terms`	Proximity of topic terms to external Wikipedia references
f9	`Wikilink_Proximity_To_Topic_Terms`	Proximity of topic terms to internal Wikipedia outlinks
f10	`CThat_Proximity_To_Topic_Terms`	Proximity of topic terms to "conjugated-that" expressions

Table 2: List of claim-discovery features used for scoring documents returned by the first step for claim-oriented document re-ranking.

dispute, disputable, disagreement, debate, polemic, feud, question, schism, wrangle, controversy, dispeace, dissension, criticism, argue, disagree, argument, claim, conflict, opposition, adversary, antagonism, oppose, object, loggerheads, quarrel, fuss, moot, hassle, altercate, case, evidence, clash, issue, problem, emphasize, recommend, suggest, assert, defend, maintain, reject, support, challenge, doubt, refute, confirm, prove, validate, establish, substantiate, verify, against, resist, support, agree, consent, concur, accept, refuse, plead, right, justify, justification

Table 3: "Controversy Lexicon" (CL) terms.

To address **O3**, features f8-f9 (see Table 2) were designed, where the proximity between the topic q terms to article references[4] (or outlinks to other Wikipedia articles) is measured (in a similar manner to the contextual features of **O2**). Therefore, an article that has more terms in q that appear as close as possible to references (or outlinks) is scored higher.

Finally, to address **O4**, feature f10 was designed, where the proximity between the topic q terms to "conjugated-that" expressions (e.g., "*claim that*", "*argue that*") is measured (in a similar manner as before with an additional constraint that the "conjugated-that" expression should precede any topic q term). The "conjugated-that" terms lexicon CTL was obtained in a similar way to the Controversy Lexicon. Table 4 contains the complete list of "conjugated-that" terms that was used in this work. Again, an article that has more terms in q that appear as close as possible to "conjugated-that" expressions is scored higher.

The re-ranking score of each article $d \in D_q$ is obtained by combining the various feature scores $score_f(d; q, D_q)$. In this work, a weighted version of the CombMNZ fusion method was adopted [20] and was found to provide the best score combination strategy. Overall, each document $d \in D_q$ is assigned with 10 different (feature) scores. The various feature weights $\{w_f\}_{f=1}^{10}$ are learned using linear regression [20].

As a final step, the retrieval system returns the top-m ($\ll k$) articles in D_q^{rerank} with the highest re-ranking score $score_{rerank}(d; q, D_q)$. Hence, an effective re-ranking would be such whose claim-recall based on the top-m articles in D_q^{rerank} is higher than the one based on the top-m articles originally retrieved in D_q.

4. EVALUATION

4.1 Datasets

Two tightly coupled datasets were used for the evaluation. The first dataset is a Wikipedia corpus, used to retrieve articles relevant to a given controversial topic. This corpus is based on the (English) Wikipedia dump dated from April 2012, containing a total of 3,931,373 (indexed) articles.

The second dataset is based on a benchmark for content dependent claim detection (CDCD), recently made publicly available by Aharoni et al. [1]. This dataset was used in this work for evaluating the quality of articles retrieved from the Wikipedia corpus. The CDCD dataset includes several debate motions on "controversial" topics that were randomly selected from the `http://idebate.org` database [1]. For each debate motion, the list of relevant claims that appear in Wikipedia articles was manually identified [1]. Articles in this dataset were obtained from the **same** Wikipedia dump [1]. Overall, the CDCD dataset includes 44 debate motions[5], 1,739 confirmed claims contained across 626 Wikipedia articles (with 39.52±33.13 claims on average per debate motion). A debate motion is given in a declarative form (usually starts with "*This House...*"), expressing an argument used to support or contest some controversial topic. Table 5 includes several examples of debate motions in the CDCD dataset [1].

Claims in this dataset span from factual statements (evidenced for example by some cited study or expert) to opinions (even anecdotal) [1].

4.2 Setup

The Apache Lucene[6] search library (version 4.9) was used for the solution implementation. The various debate motions in the CDCD dataset were used as (topic) queries

[4] For this purpose, a preliminary data processing step was applied which replaced such references with a special token (e.g., "`$REF$`").

[5] The CDCD dataset described in [1] includes only 32 motions; a more up-to-date dataset was obtained by courtesy of [1].

[6] http://lucene.apache.org

Neutral meaning terms: said, say, state
"Wishful thinking" terms: recognise, believe, assume, consider, hypothesize, think
Position terms: argue, claim, emphasize, recommend, suggest, assert, defend, maintain,reject, support, challenge, doubt, put, forward, refute,
Proof terms: confirm, prove, validate, establish, substantiate, verify
Action terms: analyze, estimate, examine, investigate, study, apply, evaluate, find, observe

Table 4: "Conjugated-that" Terms Lexicon (CTL).

"*This House believes that the sale of violent video games to minors should be banned*"
"*This House supports the one-child policy of the republic of China*"
"*This House would ban gambling*"
"*This House would introduce year round schooling*"

Table 5: Example of several debate motions in the CDCD dataset [1]

(the prefix "*This House*" was removed from each query). Wikipedia articles and queries were processed using Lucene's default analysis, i.e., tokenization, (English) stemming, stopwords removal (excluding the term "*that*" to support the "conjugated-that" proximity search).

Recall that, with the assumption of a CDCD method in mind, the goal of this work is to retrieve as many relevant claims as possible, contained in articles that are returned as a response to a (debate) topic query. For that, two versions of the *generalized Recall* measure, previously suggested for evaluating focused retrieval tasks [15], were adapted to the current task.

The first recall measure (denoted $gR@m$) captures the number of documents with relevance to a given query q retrieved up to a document-rank m, divided by the total number of relevant documents in the corpus [15]. Given a query q, a document $d \in D$ is determined to be relevant to q if it contains at least one claim from C_q. For a given query q, relevant claims C_q and document d, let $C_q(d)$ further denote the subset of claims in C_q that are contained in d[7]. For a given query q and the top-m retrieved documents, this recall measure is calculated as follows [15]:

$$gR@m = \frac{\sum_{j=1}^{m} rel(d_j)}{Nrel}, \tag{2}$$

where $rel(d) = 1$ iff $|C_q(d)| > 0$ (else 0) and $Nrel$ is the total number of documents in corpus D that are relevant to query q.

The second recall measure (denoted $gR'@m$) further assumes that the relevance of documents in D to a given query q is directly proportional to the number of relevant claims contained in each document [15]. Therefore, the more relevant claims are contained in a retrieved document, the better. For a given query q and the top-m retrieved documents, this recall measure is calculated as follows [15]:

$$gR'@m = \frac{\sum_{j=1}^{m} rsize(d_j)}{Trel}, \tag{3}$$

where $rsize(d) = |C_q(d)|$ and $Trel = |C_q|$. Therefore, this measure better captures the claim-level recall of the retrieval solution, whose maximization is the main goal of this work.

Fixing $m = 20$, for each detabe motion query q in the CDCD dataset, the quality of the top-20 retrieved Wikipedia articles returned by the proposed solution was measured using the two recall measures. The intermediate document pool size k of the first step in the solution was further set so to satisfy $m = \sqrt{k}$ (i.e., $k = 400$).

The effectiveness of the proposed claim-oriented document retrieval approach (denoted `Rerank` hereinafter) was compared to four other baselines. Three baselines were based on a "pure" **topic-only** retrieval approach as follows. The first, `LucBM25`, solely used Lucene's BM25 scoring; i.e., no additional query enhancements (described Section 3.2) were applied. The second, `Topic`, implemented a topic-only retrieval approach according to Eq. 1; i.e., no additional reranking step was applied. The third, `TopicCntrv`, refines the `Topic` approach by only considering articles that are (manually) annotated as "controversial" in Wikipedia. Finally, the last baseline, `Unified`, "directly" retrieved the top-m documents in D according to a single unified topic/claim scoring (which combined all topic and claim feature scores together using a linear regression).

A 10-fold cross-validation was performed in order to evaluate the proposed approach and the various baselines. On each fold, the train set (i.e., consisting of 90% of the debate motions and their corresponding labeled articles/claims in the CDCD dataset) was used for learning the free parameters (i.e., λ_o, w_f and maximum proximity window sizes). The quality of the various retrieval methods was recorded using the test set (consisting of the 10% remaining debate motions). The average performance across all folds was measured and is reported next.

Table 6 depicts the evaluation results. The best recall numbers obtained by any of the methods are further highlighted in bold. All values reported in this table were verified for statistical significance using a paired t-test (p-value < 0.05).

Comparing `Topic` and `LucBM25`, it is apparent that the additional query enhancements (i.e., term weighting and expansion steps) that were described in Section 3.2 provide further boost to a topic-only document retrieval (up to 29.4% and 17.8% improvement in document and claim recall, respectively). This serves as an empirical proof to the base assumption (that was made in Section 3.1) that the more "topic-focused" documents are retrieved[8]; the more claims may be contained in such documents.

Next, comparing `TopicCntrv` and `Topic`, it is clear that discovering relevant claims by solely focusing on articles that were manually annotated as "controversial" in Wikipedia only provides a partial solution. Many relevant claims ac-

[7] $C_q(d)$ of various documents in the CDCD dataset are mutually exclusive.

[8] This was easily verified by recording (document) Precision@20, with 0.12 and 0.15 (i.e., +25% improvement) obtained by the `LucBM25` and `Topic` baselines, respectively.

	Documents Recall			**Claims Recall**		
Methods	gR@5	gR@10	gR@20	gR'@5	gR'@10	gR'@20
`LucBM25`	0.13	0.25	0.34	0.19	0.36	0.45
`Topic`	0.27	0.37	0.44	0.38	0.47	0.53
`TopicCntrv`	0.05	0.06	0.06	0.07	0.08	0.08
`ClaimUnified`	**0.28**	0.37	0.49	**0.41**	0.50	0.58
`ClaimRerank`	$\mathbf{0.28}^{lt}_{c}$	$\mathbf{0.41}^{lt}_{uc}$ (+10.8%)	$\mathbf{0.51}^{lt}_{uc}$ (+4.1%)	$\mathbf{0.41}^{lt}_{c}$	$\mathbf{0.55}^{lt}_{uc}$ (+10%)	$\mathbf{0.64}^{lt}_{uc}$ (+10.3%)

Table 6: Evaluation results. The letters *l*, *t*, *c* and *u* mark a statistically significant difference with the `LucBM25`, `Topic`, `TopicCntrv` and `ClaimUnified` baselines.

tually reside within articles that have no controversy issues. Next, comparing `Unified` and `Rerank` together side by side against `Topic`, it is further apparent that a topic-only strategy for a claim-oriented document retrieval is an inferior approach. The consideration of the various claim discovery features further boosts the document and claim recall by 15.9% and 20.8%, respectively.

Finally, `Rerank`, which is the retrieval approach proposed by this work, is superior to `Unified`, with a significant improvement in both document and claim recall for $m > 5$ (up to 10.8% and 10.3% improvement for document and claim recall, respectively). Using sequential forward selection, the following (claim discovery) features were found to be the most influential (ordered by their marginal contribution to each other on top of the `Topic` baseline): **f6** (+9.4%), **f3** (+1.7%), **f8-9** (+3.4%), **f10** (+1.6%), **f7** (+3.2%).

5. CONCLUSION AND FUTURE WORK

This work focused on a novel **claim-oriented document retrieval** task. Given a (controversial) topic, Wikipedia articles that contain as many relevant claims as possible to the topic were retrieved. A two-step retrieval approach was proposed, where a pool of articles that are focused on the topic are first retrieved and then re-ranked according to several claim discovery features. An evaluation of the proposed approach, using a recently published claims benchmark, has demonstrated its ability to provide more relevant claims compared to several other retrieval alternatives.

As a future work, additional claim discovery features may be explored while existing ones may be further improved. For example, the "Controversy Lexicon" which was manually "crafted" in this work could be automatically generated. One possibility would be to utilize Wikipedia's manually annotated controversial articles to automatically learn the lexicon. As another example, new features may be extracted by analyzing the user discussions and edit histories that accompany Wikipedia articles; e.g., the length of an edit history (or the frequency of its edits) of some Wikipedia article may sometimes be a good controversy indicator [19].

Acknowledgement

The authors would like to thank Iris Eiron for her preliminary contribution to this work.

6. REFERENCES

[1] E. Aharoni, A. Polnarov, T. Lavee, D. Hershcovich, R. Levy, R. Rinott, D. Gutfreund, and N. Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, ACL '14, 2014.

[2] P. Bellot, A. Doucet, S. Geva, S. Gurajada, J. Kamps, G. Kazai, M. Koolen, A. Mishra, V. Moriceau, J. Mothe, et al. Overview of inex 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 269–281. Springer, 2013.

[3] E. Cabrio and S. Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 208–212, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[4] J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 302–310, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[5] D. Carmel, E. Farchi, Y. Petruschka, and A. Soffer. Automatic query refinement using lexical affinities with maximal information gain. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 283–290, New York, NY, USA, 2002. ACM.

[6] S. Dori-Hacohen and J. Allan. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 1845–1848, New York, NY, USA, 2013. ACM.

[7] A. Freeley and D. Steinberg. *Argumentation and debate.* Cengage Learning, 2013.

[8] S. Geva, J. Kamps, M. Lethonen, R. Schenkel, J. A. Thom, and A. Trotman. Overview of the inex 2009 ad hoc track. focused retrieval and evaluation. In *Focused retrieval and evaluation*, pages 4–25. Springer, 2010.

[9] O. Kolomiyets and M.-F. Moens. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434, 2011.

[10] M. Koolen, G. Kazai, M. Preminger, and A. Doucet. Overview of the inex 2013 social book search track. In *In CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, 2013.

[11] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim. Context dependent claim detection. In *Proceedings of COLIG '14*, 2014.

[12] B. Liu and L. Zhang. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer, 2012.

[13] O. Medelyan, D. Milne, C. Legg, and I. H. Witten. Mining meaning from wikipedia. *Int. J. Hum.-Comput. Stud.*, 67(9):716–754, Sept. 2009.

[14] R. M. Palau and M.-F. Moens. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA, 2009. ACM.

[15] J. Pehcevski, J. A. Thom, et al. Evaluating focused retrieval tasks. In *SIGIR 2007 Workshop on Focused Retrieval*, 2007.

[16] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. pages 109–126, 1996.

[17] W. Song, Y. Zhang, Y. Xie, T. Liu, and S. Li. Query term ranking based on search results overlap. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1253–1254, New York, NY, USA, 2011. ACM.

[18] S. Toulmin. *The Uses of Argument.* Cambridge University Press, 1958.

[19] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, H. W. Lauw, and K. Chang. On ranking controversies in wikipedia: Models and evaluation. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 171–182, New York, NY, USA, 2008. ACM.

[20] S. Wu. *Data Fusion in Information Retrieval.* Springer Publishing Company, Incorporated, 2012.