Rhetorical Relations for Information Retrieval

Christina Lioma
Computer Science
University of Copenhagen
Denmark
c.lioma@diku.dk

Birger Larsen
Royal School of Library and
Information Science
Copenhagen Denmark
blar@iva.dk

Wei Lu
School of Information
Management
Wuhan University China
reedwhu@gmail.com

ABSTRACT

Typically, every part in most coherent text has some plausible reason for its presence, some function that it performs to the overall semantics of the text. Rhetorical relations, e.g. `contrast, cause, explanation`, describe how the parts of a text are linked to each other. Knowledge about this so-called discourse structure has been applied successfully to several natural language processing tasks. This work studies the use of rhetorical relations for Information Retrieval (IR): Is there a correlation between certain rhetorical relations and retrieval performance? Can knowledge about a document's rhetorical relations be useful to IR?

We present a language model modification that considers rhetorical relations when estimating the relevance of a document to a query. Empirical evaluation of different versions of our model on TREC settings shows that certain rhetorical relations can benefit retrieval effectiveness notably ($> 10\%$ in mean average precision over a state-of-the-art baseline).

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*linguistic processing*

Keywords

Rhetorical relations, discourse structure, retrieval model, probabilistic retrieval

1. INTRODUCTION

According to discourse analysis, every part in most coherent text tends to have some plausible reason for its presence, some function that it performs to the overall semantics of the text. Rhetorical relations, e.g. `contrast, explanation, condition`, are considered critical for text interpretation, because they signal how the parts of a text are linked to each other to form a coherent whole [23]. Unlike grammatical relations, which are generally explicitly manifest in language, rhetorical relations may be unstated. The goal of discourse analysis is therefore to infer rhetorical relations, and specifically to identify their span, constraints and function.

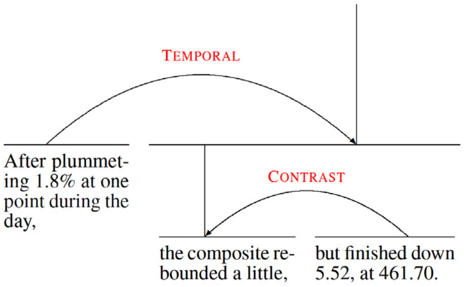
TEMPORAL
After plummeting 1.8% at one point during the day,
CONTRAST
the composite rebounded a little,
but finished down 5.52, at 461.70.

Figure 1: Rhetorical relations example (from [11]).

There is a large body of research on both descriptive and predictive models of rhetorical structure and discourse analysis in natural language text. For instance, annotation projects have taken significant steps towards developing semantic [12, 18] and discourse [5] annotated corpora. Some of these annotation efforts have already had a computational impact, making it possible to automatically induce semantic roles [15] and to automatically identify rhetorical relations [14], achieving near-human levels of performance on certain tasks [27]. In addition, applications of discourse analysis to automatic language processing tasks such as summarisation or classification (overviewed in section 2) indicate that rhetorical relations can enhance the performance of well-trained natural language processing systems.

Motivated by these advances, this work brings perspectives from discourse analysis into Information Retrieval (IR) with the aim of investigating if and how rhetorical relations can benefit retrieval effectiveness. Is there a correlation between certain rhetorical relations and retrieval performance? Can knowledge about a document's rhetorical relations be useful to IR? For example, consider the rhetorical relations of the text shown in Figure 1 (borrowed from [11]). Should some of the terms in this sentence be given extra weight by an IR system, according to their rhetorical relations? Can some rhetorical relations be considered more informative and hence more useful for IR ranking than others? These questions have been posed before (see discussion in section 2), however to our knowledge this is the first time that a principled integration of rhetorical relations into a probabilistic IR model improves precision by $> 10\%$.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
Copyright 2012 ACM 978-1-4503-1472-5/12/08 ...$15.00.

Reasoning about query - document relevance using the language modeling formalism [9], we present a model that conditions the probability of relevance between a query and a document on the rhetorical relations occurring in that document. We present an application of this model to an IR re-ranking task, where, given a list of documents initially retrieved for a query, the goal is to improve the ranking of the documents by refining their estimation of relevance to the query. Experimental evaluation of different versions of our model on TREC data and standard settings demonstrates that certain rhetorical relations can be beneficial to retrieval, with notable improvements to retrieval effectiveness (> 10% in mean average precision and other standard TREC evaluation measures over a state-of-the-art baseline).

2. RELATED WORK

Discourse analysis and rhetorical structures have been studied in the context of several automatic text processing applications. This has been partly enabled by the availability of discourse parsers - see [11, 14] for up-to-date overviews of discourse parsing technology. Studies of discourse analysis in relation to IR and its broader applications are briefly overviewed below. For a more general overview of discourse analysis approaches, see Wang et al. [33], section 2.

Sun & Chai [28] investigate the role of discourse processing and its implication on query expansion for a sequence of questions in scenario-based context question answering (QA). They consider a sequence of questions as a mini discourse. An empirical examination of three discourse theoretic models indicates that their discourse-based approach can significantly improve QA performance over a baseline of plain reference resolution.

In a different task, Wang et al. [33] parse Web user forum threads to determine the discourse dependencies between posts in order to improve information access over Web forum archives. They present three different methods for classifying the discourse relationships between posts, which are found to outperform an informed baseline.

Heerschop et al. [16] perform document sentiment analysis (partly) based on a document's discourse structure. They hypothesise that by splitting a text into important and less important text spans, and by subsequently making use of this information by weighting the sentiment conveyed by distinct text spans in accordance with their importance, they can improve the performance of a sentiment classifier. A document's discourse structure is obtained by applying rhetorical structure theory on a sentence level. They report a 4.5% improvement in sentiment classification accuracy when considering discourse, in comparison to a non-discourse based baseline. Similarly to this study, Somasundaran et al. [26] report improvements to opinion polarity classification when using discourse, and Morato et al. [24] report a positive dependence between classification performance and certain discourse variables. An overview of discourse analysis for opinion detection can be found in Zhou et al. [36].

In the area of text compression, Louis et al. [21] study the usefulness of rhetorical relations between sentences for summarisation. They find that most of the significant rhetorical relations are associated to non-discriminative sentences, i.e. sentences that are *not important* for summarisation. They report that rhetorical relations that may be intuitively perceived as highly salient do not provide strong indicators of informativeness; instead, the usefulness of rhetorical relations is in providing constraints for navigating through the text's structure. These findings are compatible with the study of Clarke & Lapata [7] into constraining text compression on the basis of rhetorical relations. For a more in-depth look into the impact of individual rhetorical relations to summarisation see Teufel & Moens [30].

In domain-specific IR, Yu et al. [34] focus on psychiatric document retrieval, which aims to assist users to locate documents relevant to their depressive problems. They propose the use of high-level discourse information extracted from queries and documents, such as negative life events, depressive symptoms and semantic relations between symptoms, to improve the precision of retrieval results. Their discourse-aware retrieval model achieves higher precision than the vector space and Okapi models.

Closer to our work, Wang et al. [31] extend an IR ranking model by adding a re-ranking strategy based on document discourse. Specifically, their re-ranking formula consists of the original retrieval status value computed with the BM11 model, which is then multiplied by a function that linearly combines inverse document frequency and term distance for each query term within a discourse unit. They focus on one discourse type only (`advantage-disadvantage`) which they identify manually in queries, and show that their approach improves retrieval performance for these queries. Our work differs on several points. We use an automatic (not manual) discourse parser to identify rhetorical relations in the documents to be retrieved (not queries). We consider 15 rhetorical relations (not 1) and we study their impact to retrieval performance using a modification of the IR language model.

Finally, Suwandaratna & Perera [29] also present a re-ranking approach for Web search that uses discourse structure. They report a heuristic algorithm for refining search results based on their rhetorical relations. Their implementation and evaluation is partly based on a series of ad-hoc choices, making it hard to compare with other approaches. They report a positive user-based evaluation of their system for ten test cases.

3. RANKING WITH RHETORICAL RELATIONS

There may be various ways of considering rhetorical relations in an IR setting. In this work, we view rhetorical relations as non-overlapping text spans, rather than a graph or a tree with structure and overlapping nodes [27]. We select a principled integration of rhetorical relation information into the retrieval model that ranks documents with respect to queries. The goal is to enable evidence about the rhetorical relations in a document to have a quantifiable impact upon the estimation of relevance of this document to a query, and to study that impact.

3.1 Model Derivation

Let q be a query, d a document, D a collection of documents, and ψ_g a rhetorical relation in the collection (so that $\sum_{\psi_g} p(\psi_g|d) = 1$). In probabilistic IR, each d in D can be ranked by its probability $p(d|q)$ of being relevant to q. Using Bayes' law:

$$p(d|q) = \frac{p(q|d)p(d)}{p(q)} \overset{rank}{=} p(q|d) \qquad (1)$$

where the right-hand side of Equation 1 is derived as follows: $p(q)$ is dropped because it is fixed for all documents, and $p(d)$ can be dropped on the assumption that it is uniform in the absence of any prior knowledge about any document. Using the language modeling approach to IR [9], $p(q|d)$ can be interpreted as the probability of generating the terms in q from a model induced by d, or more simply how likely it is that the document is about the same topic as the query. $p(q|d)$ can be estimated in different ways, for instance using Dirichlet, Jelinek-Mercer, or two-stage smoothing [35].

We introduce into Equation 1 the probability of generating the query terms from a model induced by d and by its rhetorical relations $\psi_g \in d$ as follows:

$$p(q|d) = \sum_{\psi_g} p(q|d, \psi_g) p(\psi_g|d) \qquad (2)$$

We now explain the two components in Equation 2. The first component, $p(q|d, \psi_g)$, can be interpreted as the probability of generating the query terms from a model induced by d and ψ_g. We estimate $p(q|d, \psi_g)$ as a simple mixture of the probabilities of generating q from d and ψ_g:

$$p(q|d, \psi_g) = (1 - \kappa) \cdot p(q|d) + \kappa \cdot p(q|\psi_g) \qquad (3)$$

where $p(q|d)$ is the (baseline) probability of relevance between q and d mentioned in the beginning of this section, κ is a free parameter, and $p(q|\psi_g)$ can be interpreted as the probability of generating q from a model induced by the rhetorical relation ψ_g, or more simply, the 'likelihood of relevance' between the terms in the query and the terms in the rhetorical relation.

The second component of Equation 2, $p(\psi_g|d)$, is the probability of the rhetorical relation given the document. Similarly to above, this can be interpreted as the probability of generating the terms in ψ_g from a model induced by d, or more simply the likelihood of relevance between the terms in the rhetorical relation and the terms in the document.

3.2 Model Induction

To make Equations 2-3 operational we need to compute $p(q|\psi_g)$ and $p(\psi_g|d)$. One simple way of doing so is using the respective maximum likelihood estimations:

$$\log p(q|\psi_g) = \sum_{i=1}^{|q|} \frac{f(q_i, \psi_g)}{|\psi_g|} \qquad (4)$$

where $f(q_i, \psi_g)$ is the frequency of the query term q_i in ψ_g, and $|\psi_g|$ is the number of terms in ψ_g.

$$\log p(\psi_g|d) = \sum_{j=1}^{|\psi_g|} \frac{f(\psi_{gj}, d)}{|d|} \qquad (5)$$

where $f(\psi_{gj}, d)$ is the frequency of the rhetorical relation term ψ_{gj} in d, and $|d|$ is the number of terms in d. In this work, we use the above equations and, to compensate for zero-frequency cases, we apply add-one smoothing.

Alternative principled estimations of Equations 4-5 are possible (e.g. Dirichlet, Good-Turing) and could potentially improve the performance reported in this work. For instance, one could discount the frequencies in Equations 4-5 by a respective collection model using Dirichlet smoothing: $\log p_s(q|\psi_g) = \sum_{i=1}^{|q|} \frac{f(q_i,\psi_g)+\mu \cdot p(q_i|\Psi)}{|\psi_g|+\mu}$ where μ would be the smoothing parameter and Ψ would be the collection of all rhetorical relations in D. A similarly Dirichlet smoothed alternative estimation of Equation 5 would be: $\log p_s(\psi_g|d) = \sum_{j=1}^{|\psi_g|} \frac{f(\psi_{gj},d)+\mu \cdot p(\psi_{gj}|D)}{|d|+\mu}$. We choose to use maximum likelihood instead of Dirichlet to avoid introducing the extra Dirichlet smoothing parameter μ when investigating the effect of rhetorical relations upon retrieval.

Another alternative would be to use Good-Turing smoothing, however doing so would scale down the maximum likelihood estimations in Equations 4-5 by a factor of $1 - \frac{E(1)}{|\psi_g|}$ and $1 - \frac{E(1)}{|d|}$ respectively, where $\frac{E(1)}{|\psi_g|}$ (resp. $\frac{E(1)}{|d|}$) is the estimate of how many items in the numerator of Equation 4 (resp. Equation 5) have occurred once in the sample of the denominator (see Gale & Sampson [13] for more on Good-Turing smoothing). In effect, for Equation 4 this scaling down would reduce the probability of the query terms that we have seen in ψ_g, making room for query terms that we have not seen. For our setting this would not be necessary, because in practice most queries and most rhetorical relations correspond to rather short text spans. Good-Turing smoothing might be better suited for larger samples [13].

Overall, the model presented in this section can be seen as a 'basic model' for ranking documents (partly) according to their rhetorical relations. Different variations on this basic model are certainly possible, however we choose to use the simple maximum likelihood version of this model for this exploratory investigation into the potential benefits of using rhetorical relations for IR.

4. EVALUATION

4.1 Experimental Setup

We evaluate our model on the task of re-ranking an initial list of documents, which has been retrieved in response to a query. Re-ranking is a well-known IR practice that can enhance retrieval performance notably [19]. The baseline of our experiments consists of the top 1000 documents retrieved for each query using a state-of-the-art retrieval model (language model with Dirichlet smoothing[1] [9]). Our approach reranks these documents using Equation 2.

4.1.1 Dataset and Pre-processing

We experiment with the TREC datasets of the Web 2009 (queries 1-50) and Web 2010 (queries 51-100) tracks, that contain collectively 100 queries and their relevance assessments on the Clueweb09 cat. B dataset[2] (50,220,423 web pages in English crawled between January and February 2009). We choose these datasets because they are used widely in the community, allowing comparisons with state-of-the-art. We remove spam using the spam rankings of Cormack et al. [8] with the recommended setting of percentile-score < 70 indicating spam[3].

We consider a subset of this collection, consisting of the top 1000 documents that have been retrieved in response to each query by the baseline retrieval model on tuned settings

[1] We also experimented with Jelinek-Mercer and two-stage smoothing for the baseline retrieval model. Dirichlet and two-stage gave higher scores. We chose Dirichlet over two-stage because it includes one less parameter to tune.

[2] http://lemurproject.org/clueweb09.php/

[3] Note that removing spam from Clueweb09 cat B. is known to give overall lower retrieval scores than keeping spam [3].

Table 1: Examples of the 15 rhetorical relations (in bold italics) of our dataset, identified by the SPADE discourse parser [27]

Rhetorical relation	Example sentences with rhetorical relations italicised and bold
attribution	... the islands now known as the Gilbert Islands were settled ***by Austronesian-speaking people*** ...
background	... many whites had left the country ***when Kenyatta divided their land among blacks*** ...
cause-result	... ***I plugged "wives" into the search box and came up with the following results*** ...
comparison	... so for humans, ***it is stronger than coloured*** to frustrate these unexpected numbers ...
condition	... Conditional money ***based upon care for the pet*** ...
consequence	... voltage drop with the cruise control switch ***could cause erratic cruise control operation*** ...
contrast	... ***Although it started out as a research project***, the ARPANET quickly developed into ...
elaboration	... order accutane ***no prescription required*** ...
enablement	... The project will also ***offer exercise programs and make eye care services accessible*** ...
evaluation	... such advances will be reflected in an ever-***greater proportion of grade A recommendations*** ...
explanation	... the concept ***called as "evolutionary developmental biology" or shortly "evo-devo"*** ...
manner-means	... Fill current path ***using even-odd rule, then paint the path*** ...
summary	... Safety Last, Girl Shy, Hot Water, The Kid Brother, Speedy ***(all with lively orchestral scores)*** ...
temporal	... Take time out ***before you start writing*** ...
topic-comment	... ***Director Mark Smith expressed support for greyhound adoption*** ...

(described in section 4.1.2) using the Indri IR system[4] for indexing and retrieval. For this subset, we strip HTML annotation using our in-house WHU-REAPER crawling and web parsing toolkit[5]. Rhetorical relations are identified using the freely available SPADE discourse parser [27]. Table 1 shows the 15 types of rhetorical relations identified by this process, with examples taken from the re-ranking dataset.

4.1.2 Parameter Tuning

Two parameters are involved in these experiments: the Dirichlet smoothing parameter μ of the retrieval model (used by both the baseline and our approach) and the mixture parameter κ of our model. Both parameters are tuned using 5-fold cross validation for each query set separately; results reported are the average over the five test sets. μ is tuned across {100, 500, 800, 1000, 2000, 3000, 4000, 5000, 8000, 10000} (using the range of Zhai & Lafferty [35]) and κ is tuned across {0.1, 0.3, 0.5, 0.7, 0.9}.

Performance is reported and tuned separately for Mean Average Precision (MAP), Binary Preference (BPREF), and Normalised Discounted Cumulated Gain (NDCG). These measures contribute different aspects to the overall evaluation: BPREF measures the average precision of a ranked list; it differs from MAP in that it does not treat non-assessed documents as explicitly non-relevant (whereas MAP does) [4]. This is a useful insight, especially for a collection as large as Clueweb09 cat. B where the chances of retrieving non-assessed documents are higher. NDCG measures the gain of a document based on its position in the result list. The gain is accumulated from the top of the ranked list to the bottom, with the gain of each document discounted at lower ranks. This gain is relative to the ideal based on a known recall base of relevance assessments [17]. Finally, we test the statistical significance of our results using the t-test at 95% and 99% confidence levels [25].

4.2 Findings

Figure 2 shows the distribution of the rhetorical relations in our re-ranking dataset as a percentage of the total number of rhetorical relations. `Elaboration, attribution` and `background` are the most frequent rhetorical relations, whereas `topic-comment` is the most infrequent. This happens because quite often in text a topic forms the nucleus of the discourse, which is then linked by a number of different rhetorical relations, for instance about its background, elaborating on an aspect, or attributing parts of it to some entity. As a result, several types of other rhetorical relations can correspond to a single `topic-comment`. Note that the distribution of rhetorical relations reported here is in agreement with the literature, e.g. Teufel & Moens [30] also report a 5% occurrence of `contrast`, albeit in the domain of scientific articles.

4.2.1 Retrieval-Enhancing Rhetorical Relations

Table 2 shows the performance of our model against the baseline, for each rhetorical relation and evaluation measure. The baseline performance is among the highest reported in the literature for these setings; for instance Bendersky et al. [3] report MAP=0.1605 for a tuned language model baseline with the Web 2009 track queries on Clueweb cat. B without spam.

We observe that different rhetorical relations perform differently across evaluation measures and query sets. The four rhetorical relations that improve performance over the baseline consistently for all evaluation measures and query sets (shaded rows in Table 2) are: `background, cause-result, condition` and `topic-comment`. `Topic-comment` is one of the overall best-performing rhetorical relations, which in simple terms means that boosting the weight of the topical part of a document improves its estimation of relevance.

A closer look at which rhetorical relations decrease performance presents a more uneven picture as no relations consistently underperform for all measures and query sets. Some relations, such as `explanation` and `enablement` for Web 2009, and `summary` and `evaluation` for Web 2010, are among the lowest performing, but are not under the baseline across all measures and both query sets. This implies that separating rhetorical relations into those that generally can enhance retrieval performance and those that cannot may not be straight-forward. Even though exploring the family likeness between useful relations and ones that give no mileage is an interesting discussion, in the rest of the pa-

[4] http://www.lemurproject.org/

[5] Freely available by emailing the third author.

Table 2: Retrieval performance with rhetorical relations and without (baseline). * (**) marks stat. significance at 95% (99%) using the t-test. Bold means > baseline. % shows the difference from the baseline. Shaded rows indicate consistent improvements over the baseline at all times.

rhetorical relation	Web 2009 (queries 1-50)						Web 2010 (queries 51-100)					
	MAP		BPREF		NDCG		MAP		BPREF		NDCG	
none (baseline)	0.1625		0.3230		0.3893		0.0986		0.2240		0.2920	
attribution	**0.1654***	+1.8%	**0.3275****	+1.4%	**0.3927****	+0.9%	0.0924	-6.2%	**0.2549****	+13.8%	**0.3008****	+3.0%
background	**0.1646**	+1.3%	**0.3291****	+1.9%	**0.3910**	+0.4%	**0.1086***	+10.2%	**0.2623****	+17.1%	**0.3070****	+5.1%
cause-result	**0.1626**	+0.1%	**0.3255****	+0.8%	**0.3900**	+0.2%	**0.1015**	+2.9%	**0.2491***	+11.2%	**0.3079**	+5.4%
comparison	0.1610	-0.9%	**0.3251***	+0.6%	0.3877	-0.4%	**0.1017**	+3.1%	**0.2282**	+1.9%	**0.3040****	+4.1%
condition	**0.1632**	+0.5%	**0.3258****	+0.9%	**0.3903**	+0.3%	**0.0999**	+1.3%	**0.2470****	+10.3%	**0.2936**	+0.5%
consequence	0.1602	-1.4%	**0.3250**	+0.6%	0.3874	-0.5%	0.0945	-4.1%	**0.2377***	+6.1%	**0.2840****	-2.7%
contrast	0.1549*	-4.6%	**0.3269****	+1.2%	**0.3897**	+0.1%	**0.1103***	+11.8%	**0.2531****	+13.0%	**0.3069****	+5.1%
elaboration	0.1556*	-4.2%	**0.3292****	+1.9%	0.3866	-0.7%	0.0951	-3.5%	**0.2598****	+16.0%	**0.3005****	+2.9%
enablement	0.1601	-1.4%	**0.3240**	+0.3%	0.3869*	-0.6%	**0.1010**	+2.4%	**0.2316***	+3.4%	**0.2992***	+2.5%
evaluation	**0.1632**	+0.5%	**0.3242**	+0.4%	0.3886	-0.2%	0.0814**	-17.4%	**0.2313***	+3.3%	**0.2902**	-0.6%
explanation	0.1546	-4.9%	**0.3259***	+0.9%	0.3813	-2.1%	**0.1034**	+4.9%	**0.2645****	+18.1%	**0.3069****	+5.1%
manner-means	0.1623	-0.1%	**0.3253***	+0.7%	0.3884	-0.2%	0.0986	-	**0.2324***	+3.7%	**0.2897**	-0.8%
summary	**0.1626**	+0.1%	**0.3241**	+0.3%	0.3879	-0.4%	0.0862	-12.6%	0.2220*	-0.9%	**0.2928**	+0.3%
temporal	0.1615	-0.6%	**0.3262****	+1.0%	0.3887	-0.2%	0.0921	-6.6%	**0.2546****	+13.7%	**0.3052**	+4.5%
topic-comment	**0.1673**	+3.0%	**0.3375**	+4.5%	**0.3976***	+2.1%	**0.1090***	+10.5%	**0.2476***	+10.5%	**0.3009**	+3.1%

Table 3: Effect of the rhetorical relation to the retrieval model as indicated by parameter κ (see Equation 3), for the tuned runs of Table 2. Shaded rows indicate rhetorical relations that consistently improve performance over the baseline at all times.

rhetorical relation	Web 2009 (queries 1-50)			Web 2010 (queries 51-100)		
	MAP	BPREF	NDCG	MAP	BPREF	NDCG
attribution	0.1	0.5	0.1	0.3	0.5	0.3
background	0.2	0.6	0.2	0.3	0.7	0.3
cause-result	0.3	0.7	0.3	0.5	0.7	0.5
comparison	0.4	0.7	0.4	0.3	0.5	0.3
condition	0.3	0.7	0.3	0.3	0.5	0.3
consequence	0.5	0.7	0.5	0.5	0.7	0.5
contrast	0.3	0.7	0.3	0.3	0.5	0.3
elaboration	0.1	0.5	0.1	0.3	0.5	0.3
enablement	0.1	0.9	0.1	0.3	0.5	0.3
evaluation	0.5	0.7	0.5	0.5	0.7	0.5
explanation	0.5	0.7	0.5	0.5	0.7	0.5
manner-means	0.5	0.7	0.5	0.5	0.7	0.5
summary	0.5	0.7	0.5	0.3	0.7	0.3
temporal	0.1	0.7	0.1	0.3	0.5	0.3
topic-comment	0.5	0.5	0.5	0.5	0.7	0.5

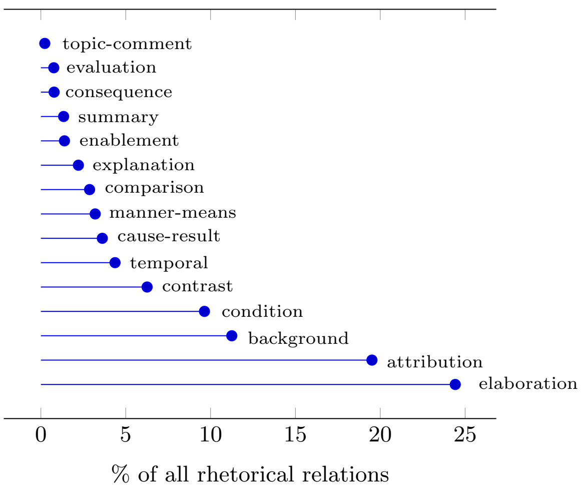

Figure 2: % distribution of rhetorical relations in our dataset.

per we focus on those rhetorical relations that consistently improve retrieval performance (for these datasets).

Improvements over the baseline are generally higher for Web 2010 than Web 2009, possibly because the former baseline is weaker, with potentially more room for improvement. An interesting trend is that more rhetorical relations improve performance according to BPREF than according to MAP and NDCG. As BPREF is the only of these evaluation measures that does not consider non-assessed documents as non-relevant, this indicates the presence of non-assessed documents in the ranking.

The scores shown in Table 2 are averaged over tens of queries, meaning that they can be affected by outliers. Figure 3 presents a detailed per-query overview of the performance of each query in relation to the baseline for each of the 15 rhetorical relations[6]. The plotted points represent the difference in MAP between our approach and the baseline. Positive points indicate that our approach outperforms the baseline. The points are sorted.

We observe that although the overall performance of the Web 2010 query set is lower than that of the Web 2009 query set, the improvements over the baseline of the 2010 set are consistently larger. Only in one case, `topic-comment`, do the plotted points clearly cross. Overall both query sets show similar plots with outliers at both ends of the scale. However, the 2009 query set tends to have a somewhat larger proportion of negative outliers, which goes some way towards explaining the lower improvements over the baseline observed for Web 2009. The Web 2010 set shows improvements over the baseline for most of the rhetorical relations and for the majority of the queries.

4.2.2 *Quantifying the Contribution of Rhetorical Relations to the Ranking*

Exactly how much impact each rhetorical relation has on the ranking can be seen in Table 3. The table lists the κ values for the best performing tuned runs from Table 2, where high κ values mean that the rhetorical relations are given more weight in the ranking (see Equation 3). We see that none of the values are above 0.5 for MAP and NDCG, indicating that too much emphasis on the rhetorical rela-

[6] Similar trends are observed in the corresponding figures for BPREF and NDCG, which are not included here for brevity.

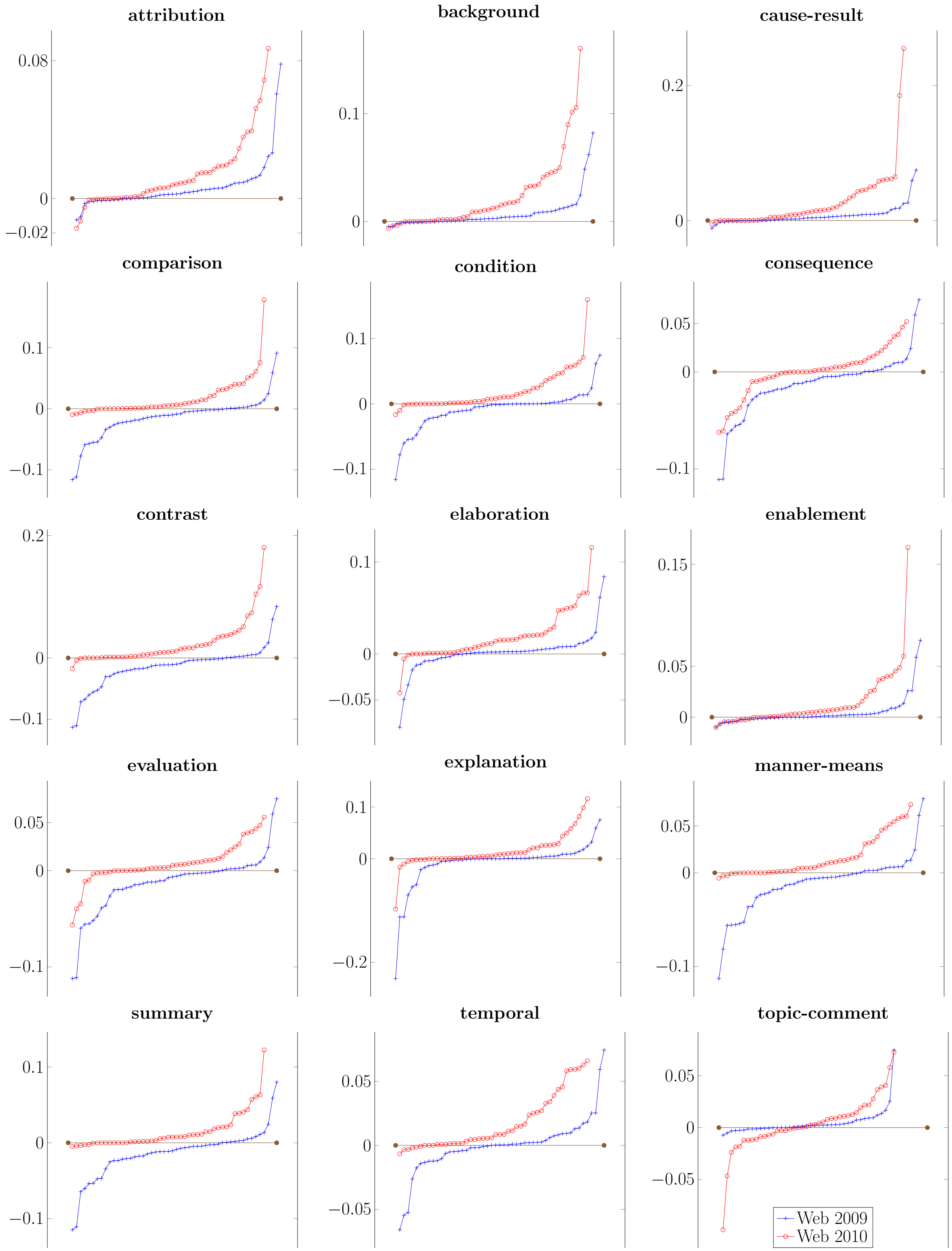

Figure 3: Sorted per-query difference in MAP between the baseline and our model (y-axis), for each rhetorical relation. The horizontal line marks the baseline. + and *o* mark the 2009 and 2010 query sets.

tions may not be beneficial to performance. Consistent with Table 2, BPREF follows a different trend than MAP and NDCG, which could be due to the fact that it is a different type of evaluation measure as discussed above in section 4.1.2. With BPREF, non-assessed documents are not explicitly penalised in the evaluation (as in MAP and NDCG) - resulting in overall higher κ values for best performing runs, typically of around 0.5-0.7.

Further we observe that the rhetorical relations that consistently improve performance over the baseline, as indicated in Table 2, differ in κ values for their best performing runs. For example, $\kappa = 0.2$ - 0.3 for `background` and $\kappa = 0.5$ for `topic-comment`. This implies that, to use rhetorical relations successfully for IR, it is not sufficient to know which rhetorical relations should be considered in the ranking and which not; also knowledge about how much emphasis to put on each rhetorical relation is needed for optimal IR performance.

Finally, note that the frequency of rhetorical relations does not affect their impact to retrieval. For instance, the three best performing rhetorical relations, `topic-comment`, `background` and `cause-result` constitute respectively approximately ∼1%, 11% and 5% of all rhetorical relations, as shown in Figure 2.

5. OPTIMISED RANKING WITH RHETORICAL RELATIONS

5.1 Rhetorical Relation Selection

The findings in section 4.2 show that some rhetorical relations can be more beneficial to retrieval performance than others. An ideal solution would not consider the lexical statistics of all rhetorical relations in a document, but rather it would select to include in the ranking only those rhetorical relations that have a higher likelihood of enhancing retrieval performance. This can be formulated as finding the optimal rhetorical relation $\hat{\psi}_g$ that maximises the expected retrieval scores according to an evaluation measure (e.g. MAP) for a query-document pair:

$$\hat{\psi}_g = \arg\max_{\psi_g \in \Psi} E[y|q,d] \tag{6}$$

where E denotes the expectation and y the retrieval score (rest of notation as defined in section 3).

Bayesian decision theory allows to reason about this type of expectation, for instance see [32]. In this work, we treat this as a problem of Bayesian posterior inference, where the goal is to estimate the retrieval performance associated with a rhetorical relation, given the observed retrieval scores it fetches on a number of queries. Then, we can consider the rhetorical relation associated with the highest retrieval performance as optimal. For this estimation, we split our dataset into different parts so that we use the observations from one to make inferences about the other (see section 5.2 for details).

Let $n = 15$ be the rhetorical relations shown in Table 2, and x_j be the number of queries for which retrieval with the j^{th} rhetorical relation gets a retrieval score y_j. For now we assume that all rhetorical relations may be expected to have similar retrieval performance, with the j^{th} rhetorical relation having an average performance ratio per query λ_j (estimated as $\frac{y_j}{x_j}$). Various densities can be used to fit similar data [22], one of which is the Poisson distribution. Let us assume that, conditional on λ_j, the retrieval scores y_j have independent Poisson distributions with means $\lambda_j x_j$. Let us further assume that the λ_j are independent realisations of a gamma variable with parameters α and β, and that β itself has a prior gamma distribution with parameters ν and ϕ. Thus

$$f(y|\lambda) = \prod_{j=1}^{n} \frac{(x_j\lambda_j)^{y_j}}{y_j!} e^{-x_j\lambda_j}$$

$$\pi(\lambda|\beta) = \prod_{j=1}^{n} \frac{\beta^{\alpha}\lambda_j^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\lambda_j}$$

$$\pi(\beta) = \frac{\phi^{\nu}\beta^{\nu-1}}{\Gamma(\nu)} e^{-\phi\beta}$$

so that the joint probability density of the retrieval scores y, the average performance ratios λ, and β is

$$f(y|\lambda)f(\lambda|\beta)\pi(\beta) = c\prod_{j=1}^{n}\{\lambda_j^{y_j+\alpha-1}e^{-\lambda_j(x_j+\beta)}\}\cdot\beta^{n\alpha+\nu-1}e^{-\phi\beta} \tag{7}$$

where c is a constant of proportionality.

The conditional density of β can be computed by various numerical approximations, one of which is the Laplace method [2], which we use here. To find the conditional density of β we integrate over the λ_j to obtain

$$f(y,\beta) = c\prod_{j=1}^{n}\{(x_j+\beta)^{-(y_j+\alpha)}\Gamma(y_j+\alpha)\}\cdot\beta^{n\alpha+\nu-1}e^{-\phi\beta} \tag{8}$$

from which the marginal density of y is obtained by further integration to give

$$f(y) = c\prod_{j=1}^{n}\Gamma(y_j+\alpha)\cdot\int_0^{\infty} e^{-h(\beta)}d\beta \tag{9}$$

where $h(\beta) = \phi\beta - (n\alpha+\nu-1)log\beta + \sum(y_j+\alpha)log(x_j+\beta)$. Let I denote the integral in this expression. In this work, we take an uninformative prior for β, with $\nu = 0.1$ and $\phi = 1$ and use $\alpha = 1.8$[7]. We then apply Laplace's method to I, resulting in the approximate posterior density for β, $\tilde{\pi}(\beta|y) = \tilde{I}^{-1}e^{-h(\beta)}$.

To calculate approximate posterior densities for λ_j we integrate Equation 7 over λ_i, $i \neq j$ and then we apply Laplace's method to the numerator and denominator integrals of

$$\pi(\lambda_j|y) = \frac{\lambda_j^{y_j+\alpha-1}e^{-\lambda_j x_j}\int_0^{\infty}e^{-h_j(\beta)}d\lambda}{\Gamma(y_j+\alpha)\int_0^{\infty}e^{-h(\beta)}d\beta}$$

where

$$h_j(\beta) = (\phi+\lambda_j)\beta - (n\alpha+\nu-1)log\beta + \sum_{i\neq j}(y_i+\alpha)log(x_i+\beta)$$

The resulting denominator is again $\tilde{I}_1$, while the numerator must be recalculated at each of a range of values for λ_j. The output is the (posterior) expected retrieval performance associated with each rhetorical relation.

5.2 Experiments

[7] These values are not tuned; they are the default values of this approach as illustrated in [10], chapter 11.3, pages 603-604.

Table 4: Retrieval performance with optimal rhetorical relations (inferred, observed) and without rhetorical relations (baseline). (1)-(5) refers to the five randomised samplings used to infer the optimal rhetorical relations. Bold marks better than baseline.

rhetorical relation	Web 2009 (queries 1-50)						Web 2010 (queries 51-100)					
	MAP		BPREF		NDCG		MAP		BPREF		NDCG	
none (baseline)	0.1625		0.3230		0.3894		0.0967		0.2198		0.2890	
optimal$_{inferred}$ (1)	**0.1879**	+15.6%	**0.3503**	+8.5%	**0.4224**	+8.5%	**0.1355**	+40.1%	**0.2859**	+30.1%	**0.3347**	+15.8%
optimal$_{inferred}$ (2)	**0.1948**	+19.9%	**0.3585**	+11.0%	**0.4202**	+7.9%	**0.1285**	+32.9%	**0.2841**	+29.3%	**0.3394**	+17.4%
optimal$_{inferred}$ (3)	**0.1984**	+22.1%	**0.3532**	+9.3%	**0.4169**	+7.1%	**0.1358**	+40.0%	**0.2906**	+32.2%	**0.3388**	+17.2%
optimal$_{inferred}$ (4)	**0.1952**	+20.1%	**0.3479**	+7.7%	**0.4282**	+10.0%	**0.1360**	+40.6%	**0.2874**	+30.8%	**0.3336**	+15.4%
optimal$_{inferred}$ (5)	**0.1950**	+20.0%	**0.3528**	+9.2%	**0.4287**	+10.1%	**0.1340**	+38.6%	**0.2865**	+30.3%	**0.3322**	+14.9%
optimal$_{observed}$	**0.2157**	+32.7%	**0.3660**	+13.3%	**0.4412**	+13.3%	**0.1474**	+52.4%	**0.2978**	+35.5%	**0.3569**	+23.5%

5.2.1 Setup

The observations required to make the above inference are triples of *rhetorical relation - query number - retrieval score*. To avoid overfitting, we pool randomly 50% of the observations from the 2009 Web query scores and 50% of the observations from the 2010 Web query scores. We use this pool to infer the expected retrieval performance of each rhetorical relation. We repeat this randomised pooling five times, each time randomly pertrubing the data, producing five different sets of observations. We then use each set to infer the expected best performing rhetorical relation per query, in accordance to Equation 6. Following this, we use the model introduced in section 3, Equation 2, to rank documents with respect to queries only for optimal (as inferred) rhetorical relations. We evaluate the above method using the same experimental settings described in section 4.1.

5.2.2 Findings

Table 4 shows the runs corresponding to the five different inferences of the best rhetorical relation that use our model (optimal$_{inferred}$ (1)-(5) respectively). We also report the optimal retrieval performance actually observed in the dataset when using the best rhetorical relation per query (optimal$_{observed}$). Optimal here means with respect to the choice of rhetorical relation, not with respect to the Dirichlet μ parameter of the baseline retrieval model.

Table 4 shows that our optimised ranking model for rhetorical relations is better than the baseline for any of the five random inferences on all three evaluation measures. The probability of getting such a positive result by chance is $\frac{1}{2^5} < 0.05$, and thus the improvements are statistically significant. The improvements over the baseline are considerable, a very promising finding given the relatively low number of observations used for optimising the choice of rhetorical relations. Experiments involving larger query sets can be reasonably expected to perform on a par with state-of-the-art performance.

More generally, the improvements in Table 4 signal that rhetorical relations (derived automatically as shown in this work) could potentially be useful features for 'linguistically-uninformed' learning-to-rank approaches.

6. DISCUSSION

6.1 Rhetorical Relation Distribution

The distribution of the 15 rhetorical relations we identified in our dataset is not the same for all rhetorical relations (see Figure 2). Some types, e.g. `topic-comment`, tend to be very sparse, whereas relations such as `elaboration` prevail. This has no impact on the model presented in section 3, but it can bias the optimised inference of the model presented in section 5. The lower the occurrence of a rhetorical relation in the dataset, the fewer the observations of retrieval performance associated with it, and hence the weaker the predictions we can infer about whether it is optimal or not. A fairer setting would be to have the same number of 'query - retrieval performance' observations for all rhetorical relations - however that would imply fiddling with the document distribution of our dataset significantly, potentially harming its quality as a test collection.

6.2 Limitations

A general limitation of discourse analysis is that not all types of text are susceptible to it. For instance, legal text, contracts, or item lists often lack rhetorical structure. In this work, we made no effort to identify and exempt such types of text from the discourse parsing. We reasoned that, as the SPADE parser includes a first-step grammatical parsing, the initial grammatical parsing of these types of text would flag out ill-formed parts (e.g. missing a verb, or consisting of extremely long sentences), which would then be skipped by the discourse analysis. This was indeed the case, however at a certain efficiency cost. Overall processing speed for SPADE was approximately 19 seconds per document (including the initial grammatical parsing), on a machine of 9 GB RAM, 8 core processor at 2.27GHz. One way of improving this performance would be to update the first-step grammatical parsing. Currently this depends on the well-known Charniak parser [6], which is one of the best performing grammatical parsers, however no longer supported. Other state-of-the-art faster grammatical parsers, e.g. the Stanford parser[8], could be adapted and plugged into SPADE instead.

The choice of applying our model for re-ranking as opposed to ranking all documents was closely related to the efficiency concerns discussed above. Our model is not specific to re-ranking only, however, using SPADE on approximately 50 million documents was too expensive at this point. Improving the discourse parser's efficiency is something we are currently working on, with the aim to apply our model for full ranking and see if the conclusions drawn from this work hold.

Finally, the accuracy of the discourse parser was not considered in this work, apart from indications in the literature that SPADE is a generally well-performing parser [27]. Given that the default version of the parser we used is trained on news articles, one may reason that its accuracy could improve if we train it on the retrieval collection, or on doc-

[8] http://nlp.stanford.edu/software/lex-parser.shtml

uments of the same domain. Note that, parsing accuracy aside, rhetorical relations assignment is not an entirely unambiguous process, even to humans [23]. For the purposes of this work, this type of fine-grained ambiguity may however not be important to retrieval performance.

6.3 Future Extensions

Future extensions include primarily making SPADE scalable on large collections of documents as discussed above, as well as using more than one rhetorical relation per document. For instance, the posterior probabilities estimated in section 5.1 could be used to weight the text in each rhetorical relation. If those posteriors are too flat, an exponent could make them peakier. As the exponent goes to infinity, the maximum relation model presented in section 5.2 would be recovered. In addition, we intend to refine the discourse analysis by considering the nucleus (i.e. central) versus satellite (i.e. peripheral) rhetorical relations for IR, as well as to improve the effectiveness of the discourse parser by training it on data of the same domain. As discussed in section 3.2, we will also investigate alternative estimations of Equations 2-3.

An interesting future research direction is the potential relation between rhetorical relations and user context: for instance, in a search session including several query reformulations, is there a correlation between the progression of the information need of the user and the rhetorical relations that the retrieval system should boost in a document (e.g. `elaboration`), as indicated by Sun & Chai [28]? Another interesting future extension of this work is in relation to evaluation measures of graded relevance measures on an inter-document level, as investigated in XML retrieval [20] for instance. If parts of a document can be regarded as more or less relevant, this may be reflected to their discourse structure. This might be especially useful for multi-threaded documents, such as multiple-user reviews and opinions, where the discourse relations tend to shift markedly. Finally, the current operationalisation of our model is simplistic in the sense that the term 'rhetorical relation' is coerced into meaning 'non-overlapping text fragment' and the actual relation between bits of text is discarded in the process. In future work we could apply fielded XML retrieval models in order to investigate nested structuring among rhetorical relations.

7. CONCLUSIONS

Rhetorical relations, e.g. `contrast, explanation, condition`, indicate the different ways in which the parts of a text are linked to each other to form a coherent whole. This work studied two questions: Is there a correlation between certain rhetorical relations and retrieval performance? Can knowledge about a document's rhetorical relations be useful to IR? To address these, we presented a retrieval model that conditions the probability of relevance between a query and a document on the rhetorical relations occurring in that document. We applied that model to an IR re-ranking scenario for Web search. Experimental evaluation of different versions of our model on TREC data and standard settings demonstrated that certain rhetorical relations can be beneficial to retrieval, with >10% improvements to retrieval precision. Furthermore, we showed that these improvements over the baseline can improve significantly, when the optimal rhetorical relation per document is selected for retrieval.

Overall, three rhetorical relations were found to benefit retrieval performance notably and consistently for different evaluation measures and query sets: `background, cause-result` and `topic-comment`. In retrospect, this is perhaps not surprising, since these are among the most salient discourse relations on an intuitive basis: the main topic or theme of a text, its background, causes and results [21]. Future extensions and research directions of this work include considering more than one rhetorical relation per document, applying our model for ranking all documents (as opposed to re-ranking only) and experimenting with alternative estimations of its components.

8. ACKNOWLEDGMENTS

We thank Kasper Hornbæk, Jakob Grue Simonsen, Raf Guns, Qikai Cheng and the anonymous reviewers for helping improve this paper. Work partially funded by the Danish International Development Agency DANIDA (grant no. 10-087721) and the National Natural Science Foundation of China (grant no. 71173164).

9. REFERENCES

[1] *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL.* ACL, 2011.

[2] A. Azevedo-Filho and R. D. Shachter. Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In R. L. de Mántaras and D. Poole, editors, *UAI*, pages 28–36. Morgan Kaufmann, 1994.

[3] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In I. King, W. Nejdl, and H. Li, editors, *WSDM*, pages 95–104. ACM, 2011.

[4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *SIGIR*, pages 25–32. ACM, 2004.

[5] L. Carlson, D. Marcu, and M. E. Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer Academic Publishers, 2003.

[6] E. Charniak. A maximum-entropy-inspired parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 132–139, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[7] J. Clarke and M. Lapata. Discourse constraints for document compression. *Computational Linguistics*, 36(3):411–441, 2010.

[8] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010.

[9] W. B. Croft and J. Lafferty. *Language Modeling for Information Retrieval.* Kluwer Academic Publishers, Norwell, MA, USA, 2003.

[10] A. C. Davison. *Statistical Models.* Cambridge University Press, New York, 2009.

[11] D. A. duVerle and H. Prendinger. A novel discourse parser based on support vector machine classification.

In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 665–673, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[12] C. J. Fillmore, C. F. Baker, and S. Hiroaki. The framenet database and software tools. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 1157–1160, 2002.

[13] W. A. Gale and G. Sampson. Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

[14] S. Ghosh, R. Johansson, G. Riccardi, and S. Tonelli. Shallow discourse parsing with conditional random fields. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1071–1079, Chiang Mai, Thailand, 2011.

[15] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. In *ACL*. ACL, 2000.

[16] B. Heerschop, F. Goossen, A. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong. Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1061–1070, New York, NY, USA, 2011. ACM.

[17] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[18] P. Kingsbury and M. Palmer. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages x–x, 2002.

[19] E. Krikon and O. Kurland. A study of the integration of passage-, document-, and cluster-based information for re-ranking search results. *Inf. Retr.*, 14(6):593–616, 2011.

[20] M. Lalmas. *XML Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[21] A. Louis, A. K. Joshi, and A. Nenkova. Discourse indicators for content selection in summarization. In R. Fernández, Y. Katagiri, K. Komatani, O. Lemon, and M. Nakano, editors, *SIGDIAL Conference*, pages 147–156. The Association for Computer Linguistics, 2010.

[22] R. Manmatha, T. M. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *SIGIR*, pages 267–275. ACM, 2001.

[23] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281, 1988.

[24] J. Morato, J. Llorens, G. Genova, and J. A. Moreiro. Experiments in discourse analysis impact on information classification and retrieval algorithms. *Inf. Process. Manage.*, 39:825–851, November 2003.

[25] M. D. Smucker, J. Allan, and B. Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *SIGIR*, pages 630–631. ACM, 2009.

[26] S. Somasundaran, G. Namata, J. Wiebe, and L. Getoor. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *EMNLP*, pages 170–179. ACL, 2009.

[27] R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information. In *HLT-NAACL*, 2003.

[28] M. Sun and J. Y. Chai. Discourse processing for context question answering based on linguistic knowledge. *Know.-Based Syst.*, 20:511–526, August 2007.

[29] N. Suwandaratna and U. Perera. Discourse marker based topic identification and search results refining. In *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on*, pages 119–125, 2010.

[30] S. Teufel and M. Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[31] D. Y. Wang, R. W. P. Luk, K.-F. Wong, and K. L. Kwok. An information retrieval approach based on discourse type. In C. Kop, G. Fliedl, H. C. Mayr, and E. Métais, editors, *NLDB*, volume 3999 of *Lecture Notes in Computer Science*, pages 197–202. Springer, 2006.

[32] J. Wang and J. Zhu. On statistical analysis and optimization of information retrieval effectiveness metrics. In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, and J. Savoy, editors, *SIGIR*, pages 226–233. ACM, 2010.

[33] L. Wang, M. Lui, S. N. Kim, J. Nivre, and T. Baldwin. Predicting thread discourse structure over technical web forums. In *EMNLP* [1], pages 13–25.

[34] L.-C. Yu, C.-H. Wu, and F.-L. Jang. Psychiatric document retrieval using a discourse-aware model. *Artif. Intell.*, 173:817–829, May 2009.

[35] C. Zhai and J. D. Lafferty. Two-stage language models for information retrieval. In *SIGIR*, pages 49–56. ACM, 2002.

[36] L. Zhou, B. Li, W. Gao, Z. Wei, and K.-F. Wong. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *EMNLP* [1], pages 162–171.