Contextualizing Citations for Scientific Summarization using Word Embeddings and Domain Knowledge

Arman Cohan
Information Retrieval Lab, Dept. of Computer Science
Georgetown University
arman@ir.cs.georgetown.edu

Nazli Goharian
Information Retrieval Lab, Dept. of Computer Science
Georgetown University
nazli@ir.cs.georgetown.edu

ABSTRACT

Citation texts are sometimes not very informative or in some cases inaccurate by themselves; they need the appropriate context from the referenced paper to reflect its exact contributions. To address this problem, we propose an unsupervised model that uses distributed representation of words as well as domain knowledge to extract the appropriate context from the reference paper. Evaluation results show the effectiveness of our model by significantly outperforming the state-of-the-art. We furthermore demonstrate how an effective contextualization method results in improving citation-based summarization of the scientific articles.

KEYWORDS

Text Summarization, Scientific Text, Information Retrieval

1 INTRODUCTION

In scientific literature, related work is often referenced along with a short textual description regarding that work which we call *citation text*. Citation texts usually highlight certain contributions of the referenced paper and a set of citation texts to a reference paper can provide useful information about that paper. Therefore, citation texts have been previously used to enhance many downstream tasks in IR/NLP such as search and summarization (e.g. [2, 15, 16]).

While useful, citation texts might lack the appropriate context from the reference article [4, 5, 18]. For example, details of the methods, assumptions or conditions for the obtained results are often not mentioned. Furthermore, in many cases the citing author might misunderstand or misquote the referenced paper and ascribe contributions to it that are not intended in that form. Hence, sometimes the citation text is not sufficiently informative or in other cases, even inaccurate [17]. This problem is more serious in life sciences where accurate dissemination of knowledge has direct impact on human lives.

We present an approach for addressing such concerns by adding the appropriate context from the reference article to the citation texts. Enriching the citation texts with relevant context from the reference paper helps the reader to better understand the context for the ideas, methods or findings stated in the citation text.

A challenge in citation contextualization is the discourse and terminology variations between the citing and the referenced authors. Hence, traditional IR models that rely on term matching for finding the relevant information are ineffective.

We propose to address this challenge by a model that utilizes word embeddings and domain specific knowledge. Specifically, our approach is a retrieval model for finding the appropriate context of citations, aimed at capturing terminology variations and paraphrasing between the citation text and its relevant reference context.

We perform two sets of experiments to evaluate the performance of our system. First, we evaluate the relevance of extracted contexts intrinsically. Then we evaluate the effect of citation contextualization on the application of scientific summarization. Experimental results on TAC 2014 benchmark show that our approach significantly outperforms several strong baselines in extracting the relevant contexts. We furthermore, demonstrate that our contextualization models can enhance summarizing scientific articles.

2 CONTEXTUALIZING CITATIONS

Given a citation text, our goal is to extract the most relevant context to it in the reference article. These contexts are essentially certain textual spans within the reference article. Throughout, colloquially, we refer to the citation text as query and reference spans in the reference article as documents. Our approach extends Language Models for IR (LM) by incorporating word embeddings and domain ontology to address shortcomings of LM for this research purpose. The goal in LM is to rank a document d according to the conditional probability $p(d|q) \propto p(q|d) = \prod_{q_i \in q} p(q_i|d)$ where q_i shows the tokens in the query q. Estimating $p(q_i|d)$ is often achieved by maximum likelihood estimate from term frequencies with some sort of smoothing. Using Dirichlet smoothing [21], we have:

$$p(q_i|d) = \frac{f(q_i, d) + \mu\, p(q_i|C)}{\sum_{w \in \mathcal{V}} f(w, d) + \mu} \tag{1}$$

where $f(q_i, d)$ shows the frequency of term q_i in document d, C is the entire collection, $\mathcal{V}$ is the vocabulary and μ the Dirichlet smoothing parameter. In the citation contextualization problem, *(i)* the target reference sentences are short documents and *(ii)* there exist terminology variations between the citing author and the referenced author. Hence, the citation terms usually do not appear in the documents and relying only on the frequencies of citation terms in the documents ($f(q_i, d)$) for estimating $p(q_i|d)$ yields an almost uniform smoothed distribution that is unable to decisively distinguish between the documents.

Embeddings. Distributed representation (embedding) of a word w in a field $\mathbb{F}$ is a mapping $w \rightarrow \mathbb{F}^n$ where words semantically similar to w will be ideally located close to it. Given a query q, we rank the documents d according to the following scoring function

SIGIR '17, Shinjuku, Tokyo, Japan
© 2017 ACM. 978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080740

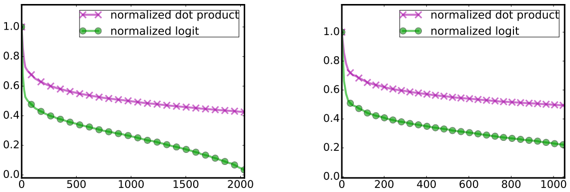

Figure 1: Dot product of embeddings and its logit for a sample word and its top most similar words (top 2000 and 1000).

which leverages this property:

$$p(q_i|d) = \frac{f_{sem}(q_i, d) + \mu \, p(q_i|C)}{\sum_{w \in V} f_{sem}(w, d) + \mu} \qquad (2)$$

where f_{sem} is a function that measures semantic relatedness of the query term q_i to the document d and is defined as: $f_{sem}(q_i, d) = \sum_{d_j \in d} s(q_i, d_j)$; where d_j's are document terms and $s(q_i, d_j)$ is the relatedness between the the query term and document term which is calculated by applying a similarity function to the distributed representations of q_i and d_j. We use a transformation (ϕ) of dot products between the unit vectors $e(q_i)$ and $e(d_j)$ corresponding to the embeddings of the terms q_i and d_j for the similarity function:

$$s(q_i, d_j) = \begin{cases} \phi(e(q_i).e(d_j)); & \text{if } e(q_i).e(d_j) > \tau \\ 0; & \text{otherwise} \end{cases}$$

We first explain the role of τ and then the reason for considering the function ϕ instead of raw dot product. τ is a parameter that controls the noise introduced by less similar words. Many unrelated word vectors have non-zero similarity scores and adding them up introduces noise to the model and reduces the performance. τ's function is to set the similarity between unrelated words to zero instead of a positive number. To identify an appropriate value for τ, we select a random set of words from the embedding model and calculate the average and standard deviation of pointwise absolute value of similarities between terms from these two samples. We then select the threshold τ to be two standard deviations larger than the average to only consider very high similarity values (this choice was empirically justified).

Examining term similarity values between words shows that there are many terms with high similarities associated with each term and these values are not highly discriminative. We apply a transfer function ϕ to the dot product $e(q_i).e(d_j)$ to dampen the effect of less similar words. In other words, we only want highly related words to have high similarity values and similarity should quickly drop as we move to less related words. We use the *logit* function for ϕ to achieve this dampening effect:

$$\phi(x) = \log(\frac{x}{1 - x})$$

Figure 1 shows this effect. The purple line is the normalized dot product of a sample word with the most similar words in the model. As illustrated, the similarity score differences among top words is not very discriminative. However, applying the logit function (green line) causes the less similar words to have lower similarity values to the target word.

Domain knowledge. Successful word embedding methods have previously shown to be effective in capturing syntactic and semantic relatedness between terms. These co-occurrence based models are data driven. On the other hand, domain ontologies and lexicons that are built by experts include some information that might not be captured by embedding methods [8]. Therefore, using domain knowledge can further help the embedding based retrieval model; we incorporate it in our model in the following ways:

1) *Retrofitting*: Faruqui et al. [6] proposed a model that uses the constraints on WordNet lexicon to modify the word vectors and pull synonymous words closer to each other. To inject the domain knowledge in the embeddings, we apply this model on two domain specific ontologies, namely, Mesh and Protein Ontologies (PO)[1]. We chose these two biomedical domain ontologies because they are in the same domain as the articles in the TAC dataset. Mesh is a broad ontology that consists of biomedical terms and PO is a more focused ontology related to biology of proteins and genes.

2) *Interpolating in the LM*: We also directly incorporate the domain knowledge in the retrieval model; we modify the LM into the following interpolated LM with parameter λ:

$$p(q_i|d) = \lambda p_1(q_i|d) + (1 - \lambda) p_2(q_i|d)$$

where p_1 is estimated using Eq. 2 and p_2 is similar to p_1 except that we replace f_{sem} with the function f_{ont} which considers domain ontology in calculating similarities:

$$f_{ont}(q_i, d) = \sum_{d_j \in d} s_2(q_i, d_j); \; s_2(q_i, d_j) = \begin{cases} 1, & \text{if } q_i = d_j \\ \gamma, & \text{if } q_i \approx d_j \\ 0, & \text{o.w.} \end{cases}$$

where $\gamma \in [0, 1]$ is a parameter and $q_i \approx d_j$ shows that there is an *is-synonym* relation in ontology between q_i and d_j[2].

3 EXPERIMENTS

Data. We use the TAC 2014 Biomedical Summarization benchmark[3]. This dataset contains 220 scientific biomedical journal articles and 313 total citation texts where the relevant contexts for each citation text are annotated by 4 experts.

Baselines. To our knowledge, the only published results on TAC 2014 is [4], where the authors utilized query reformulation (QR) based on UMLS ontology. In addition to [4], we also implement several other strong baselines to better evaluate the effectiveness of our model: 1) *BM25*; 2) *VSM*: Vector Space Model that was used in [4]; 3) *DESM*: Dual Embedding Space Model which is a recent embedding based retrieval model [12]; and 4) *LMD-LDA*: Language modeling with LDA smoothing which is a recent extension of the LMD to also account for the latent topics [10]. All the baseline parameters are tuned for the best performance, and the same preprocessing is applied to all the baselines and our methods.

Our methods. We first report results based on training the embeddings on Wikipedia (WE_{Wiki}). Since TAC dataset is in biomedical domain, many of the biomedical terms might be either out-of-vocabulary or not captured in the correct context using general embeddings, therefore we also train biomedical embeddings (WE_{Bio})[4]. In addition, we report results for biomedical embeddings with retrofitting (WE_{Bio}+rtrft), as well as interpolating domain knowledge (WE_{Bio}+dmn)

3.1 Intrinsic Evaluation

First, we analyze the effectiveness of our proposed approach for contextualization intrinsically. That is, we evaluate the quality of the

[1] https://www.nlm.nih.gov/mesh/; http://pir.georgetown.edu/pro/

[2] The values of the parameters γ and λ were selected empirically by grid search

[3] http://www.nist.gov/tac/2014/BiomedSumm/

[4] We train biomedical embeddings on TREC Genomics 2004 and 2006 collections (both Wikipedia and Genomics embeddings were trained using gensim implementation of Word2Vec, negative sampling, window size of 5 and 300 dimensions.

Table 1: Results on TAC 2014 dataset. c-P, c-R, c-F: character offset Precision, Recall and F-1 scores; Rg: Rouge; c-P@K: character offset precision at K. †shows statistical significant improvement over the best baseline performance (two-tailed t-test, $p<0.05$). Values are percentages.

Method	c-P	c-R	c-F	nDCG	Rg-1	Rg-2	Rg-3	c-P@1	c-P@5
VSM [4]	20.5	24.7	21.2	48.1	49.5	26.4	20.0	31.9	26.1
BM25	19.5	18.6	17.8	38.1	43.6	23.2	16.3	25.5	24.2
DESM [12]	20.3	23.8	22.3	45.6	50.3	26.2	20.6	32.5	26.5
LMD-LDA [10]	22.6	24.8	22.3	46.0	48.3	26.4	20.1	31.4	27.7
QR [4]	22.2	29.4	23.8	49.8	50.6	27.2	21.8	37.7	28.1
WE_{Wiki}	21.8	28.5	23.2	†52.8	50.0	26.9	20.9	36.5	29.9
WE_{Bio}	23.9	†31.2	†25.5	†57.1	51.9	†29.2	†23.1	†46.2	†34.1
WE_{Bio}+rtrft	†24.8	†**33.6**	†26.4	†58.3	52.4	†**30.7**	†24.0	†55.5	†34.9
WE_{Bio}+dmn	†**25.4**	†33.0	†**27.0**	†**59.8**	†**53.0**	†30.6	†**24.4**	†**56.1**	†**37.1**

Table 2: Top relevant words to the word "expression" according to embeddings trained on Wikipedia vs. Genomics.

General (Wikipedia)	Biomedical (Genomics)
interpretation	upregulation
sense	mrna
emotion	induction
function	protein
intension	abundance
manifestation	gene
expressive	downregulation

Table 3: Breakdown of our best model's character F-score (c-F) by quartiles of human performance measured by c-P.

Quartiles (c-P)	Q1	Q2	Q3	Q4
c-F of our model (mean ± stdev.)	16.14 ±20.20	25.41 ±7.78	33.72 ±5.81	37.50 ±5.93

extracted citation contexts using our contextualization methods in terms of how accurate they are with respect to human annotations.

Evaluation. We consider the following evaluation metrics for assessing the quality of the retrieved contexts for each citation from multiple aspects: (*i*) Character offset overlaps of the retrieved contexts with human annotations in terms of precision (c-P), recall (c-R) and F-score (c-F). These are the recommended metrics for the task per TAC[5]. (*ii*) nDCG: we treat any partial overlaps with the gold standard as a correct context and then calculate the nDCG scores. (*iii*) Rouge-N scores: To also consider the content similarity of the retrieved contexts with the gold standard, we calculate the Rouge scores between them. (*iv*) Character precision at K (c-P@K): Since we are usually interested in the top retrieved spans, we consider character offset precision only for the top K spans and we denote it with "c-P@K".

Results. The results of intrinsic evaluation of contextualization are presented in Table 1. Our models (last 4 rows of table 1) achieve significant improvements over the baselines consistently across most of the metrics. This shows the effectiveness of our models viewed from different aspects in comparison with the baselines. The best baseline performance is the query reformulation (QR) method by [4] which improves over other baselines.

We observe that using general domain embeddings does not provide much advantage in comparison with the best baseline (compare WE_{wiki} and QR in the Table). However, using the domain specific embeddings (WE_{Bio}) results in 10% c-F improvement over the best baseline. This is expected since word relations in the biomedical context are better captured with biomedical embeddings. In Table 2 an illustrative word "expression" gives better intuition why is that the case. As shown, using general embeddings (left column in the table), the most similar words to "expression" are those related to the general meaning of it. However, many of these related words are not relevant in the biomedical context. In the biomedical context, "expression" refers to "the appearance in a phenotype attributed to a particular gene". As shown on the right column, the domain specific embeddings (Bio) trained on genomics data are able to capture this meaning. This further confirms the inferior performance of the out-of-domain word embeddings in capturing correct word-level semantics [13]. Last two rows in Table 1 show incorporating the domain knowledge in the model which results in significant improvement over the best baseline in terms of most metrics (e.g. 14% and 16% c-F improvements). This shows that domain ontologies provide additional information that the domain trained embeddings may not contain. While both our methods of incorporating domain ontologies prove to be effective, interpolating domain knowledge directly (WE_{Bio}+dmn) has the edge over retrofitting (WE_{Bio}+rtrft). This is likely due to the direct effect of ontology on the interpolated language model, whereas in retrofitting, the ontology first affects the embeddings and then the context extraction model.

To analyze the performance of our system more closely, we took the context identified by 1 annotator as the candidate and the other 3 as gold standard and evaluated the precision to obtain an estimate of human performance on each citation. We then divided the citations based on human performance to 4 groups by quartiles. Table 3 shows our system's performance on each of these groups. We observe that, when human precision is higher (upper quartiles in the table), our system also performs better and with more confidence (lower std). Therefore, the system errors correlate well with human disagreement on the correct context for the citations. Averaged over the 4 annotators for each citation, the mean precision was 56.7% (note that this translates to our c-P@1 metric). In Table 1, we observe that our best method (c-P@1 of 56.1%) is comparable with average human precision score (c-P@1 of 56.7%) which further demonstrates the effectiveness of our model.

3.2 External evaluation

Citation-based summarization can effectively capture various contributions and aspects of the paper by utilizing citation texts [15]. However; as argued in section 1, citation texts do not always accurately reflect the original paper. We show how adding context from the original paper can address this concern, while keeping the benefits of citation-based summarization. Specifically, we compare how using no contextualization, versus various proposed contextualization approaches affect the quality of summarization. We apply the following well-known summarization algorithms on the set of citation texts, and the retrieved citation-contexts: LexRank, LSA-based, SumBasic, and KL-Divergence (For space constraints, we will not explain these approaches here; refer to [14] for details). We then compare the effect of our proposed contextualization methods using the standard Rouge-N summarization evaluation metrics.

Results. The results of external evaluation are illustrated in Table 4. The first row ("No context") shows the performance of each

[5] https://tac.nist.gov/2014/BiomedSumm/guidelines.html

Table 4: Effect of contextualization on summarization. Columns are summarization algorithms and rows show citation contextualization approaches. *No Context* uses only citations without any contextualization. Evaluation metrics are Rouge (Rg) scores. (†) shows statistically significant improvement over the best baseline performance (p<0.05).

	KLSUM		LexRank		LSA		SumBasic	
Method	Rg1	Rg2	Rg1	Rg2	Rg1	Rg2	Rg1	Rg2
No Context	36.0	8.3	41.3	10.8	34.7	6.5	38.7	8.7
VSM [4]	35.3	7.9	40.0	9.9	33.5	6.2	39.5	9.4
BM25	35.5	8.0	39.8	9.9	33.7	6.0	38.9	8.6
DESM [12]	36.3	8.7	40.2	10.4	32.6	6.5	38.3	7.9
LMD-LDA [10]	38.4	9.1	43.1	11.0	37.8	7.6	40.1	8.9
QR [4]	39.9	10.2	43.8	11.7	38.9	8.0	40.1	8.6
WE$_{\text{Wiki}}$	39.7	10.2	42.7	11.8	38.0	8.0	40.2	9.2
WE$_{\text{Bio}}$	†41.7	†11.7	†45.6	†**13.8**	†40.3	†9.1	†42.4	†**12.6**
WE$_{\text{Bio}}$+rtrft	†42.9	†12.2	†46.2	11.6	†40.0	8.9	†41.3	9.7
WE$_{\text{Bio}}$+dmn	†**44.0**	†**13.4**	†**47.3**	†13.6	†**42.3**	†**10.4**	†**44.0**	†11.7

summarization approach solely on the citations without any contextualization. The next 5 rows show the baselines and last 4 rows are our proposed contextualization methods. As shown, effective contextualization positively impacts the generated summaries. For example, our best method is "WE$_{Bio}$ + dmn" which significantly improves the quality of generated summaries in terms of Rouge over the ones without any context. We observe that two low-performing baseline methods for contextualization according to Table 1 ("VSM" and "BM25") also do not result in any improvements for summarization. Therefore, the intrinsic quality of citation contextualization has direct impact on the quality of generated summaries. These results further demonstrate that effective contextualization is helpful for scientific citation-based summarization.

4 RELATED WORK

Related work has mostly focused on extracting the citation text in the citing article (e.g. [1]). In this work, given the citation texts, we focus on extracting its relevant context from the reference paper. Related work have also shown that citation texts can be used in different applications such as summarization [2, 3, 9, 11, 15, 20]. Our proposed model utilizes word embeddings and the domain knowledge. Embeddings have been recently used in general information retrieval models. Vulić and Moens [19] proposed an architecture for learning word embeddings in multilingual settings and used them in document and query representation. Mitra et al. [12] proposed dual embedded space model that predicts document aboutness by comparing the centroid of word vectors to query terms. Ganguly et al. [7] used embeddings to transform term weights in a translation model for retrieval. Their model uses embeddings to expand documents and use co-occurrences for estimation. Unlike these works, we directly use embeddings in estimating the likelihood of query given documents; we furthermore incorporate ways to utilize domain specific knowledge in our model. The most relevant prior work to ours is [4] where the authors approached the problem using a vector space model similarity ranking and query reformulations.

5 CONCLUSIONS

Citation texts are textual spans in a citing article that explain certain contributions of a reference paper. We presented an effective model for contextualizing citation texts (associating them with the appropriate context from the reference paper). We obtained statistically significant improvements in multiple evaluation metrics over several strong baseline, and we matched the human annotators precision. We showed that incorporating embeddings and domain knowledge in the language modeling based retrieval is effective for situations where there are high terminology variations between the source and the target (such as citations and their reference context). Citation contextualization not only can help the readers to better understand the citation texts but also as we demonstrated, they can improve other downstream applications such as scientific document summarization. Overall, our results show that citation contextualization enables us to take advantage of the benefits of citation texts, while ensuring accurate dissemination of the claims, ideas and findings of the original referenced paper.

ACKNOWLEDGEMENTS

We thank the three anonymous reviewers for their helpful comments and suggestions. This work was partially supported by National Science Foundation (NSF) through grant CNS-1204347.

REFERENCES

[1] Amjad Abu-Jbara and Dragomir Radev. 2012. Reference scope identification in citing sentences. In *NAACL-HLT*. ACL, 80–90.

[2] Arman Cohan and Nazli Goharian. 2015. Scientific Article Summarization Using Citation-Context and Article's Discourse Structure. In *EMNLP*. 390–400.

[3] Arman Cohan and Nazli Goharian. 2017. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries* (2017).

[4] Arman Cohan, Luca Soldaini, and Nazli Goharian. 2015. Matching Citation Text and Cited Spans in Biomedical Literature: a Search-Oriented Approach. In *NAACL-HLT*. 1042–1048.

[5] Anita de Waard and Henk Pander Maat. 2012. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Workshop on Detecting Structure in Scholarly Discourse*. ACL, 47–55.

[6] Manaal Faruqui, Jesse Dodge, Kumar Sujay Jauhar, Chris Dyer, Eduard Hovy, and A. Noah Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *NAACL-HLT*. Association for Computational Linguistics, 1606–1615.

[7] Debasis Ganguly, Dwaipayan Roy, Mandar Mitra, and Gareth J.F. Jones. 2015. Word Embedding Based Generalized Language Model for Information Retrieval. In *SIGIR*. ACM, 795–798.

[8] Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with similarity estimation. *Computational Linguistics* (2015).

[9] Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 2016. Overview of the CL-SciSumm 2016 Shared Task.. In *BIRNDL@ JCDL*.

[10] Fanghong Jian, Jimmy Xiangji Huang, Jiashu Zhao, Tingting He, and Po Hu. 2016. A simple enhancement for ad-hoc information retrieval via topic modelling. In *SIGIR*. ACM, 733–736.

[11] Qiaozhu Mei and ChengXiang Zhai. 2008. Generating Impact-Based Summaries for Scientific Literature.. In *ACL*, Vol. 8. 816–824.

[12] Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. 2016. A dual embedding space model for document ranking. *CoRR arXiv:1602.01137* (2016).

[13] Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016. Classify or Select: Neural Architectures for Extractive Document Summarization. *arXiv:1611.04244* (2016).

[14] Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining text data*. Springer, 43–76.

[15] Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific Paper Summarization Using Citation Summary Networks *(COLING '08)*. 689–696.

[16] Anna Ritchie, Stephen Robertson, and Simone Teufel. 2008. Comparing Citation Contexts for Information Retrieval. In *CIKM*. ACM, 213–222.

[17] Ágnes Sándor and Anita De Waard. 2012. Identifying claimed knowledge updates in biomedical research articles. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*. ACL, 10–17.

[18] Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics* 28 (2002).

[19] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *SIGIR*.

[20] Stephen Wan, Cécile Paris, and Robert Dale. 2009. Whetting the appetite of scientists: Producing summaries tailored to the citation context. In *JCDL*. 59–68.

[21] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *TOIS* 22, 2 (2004), 179–214.