Ask the Crowd to Find Out What's Important

Sisay Fissaha Adafre
School of Computing, Dublin City University
sadafre@computing.dcu.ie

Maarten de Rijke
ISLA, University of Amsterdam
mdr@science.uva.nl

Abstract

We present a corpus-based method for estimating the importance of sentences. Our main contribution is two-fold. First, we introduce the idea of using the increasing amount of manually labeled category information (that is becoming available through collaborative knowledge creation efforts) to identify "typical information" for categories of entities. Second, we provide multiple types of empirical evidence for the usefulness of this notion of typical-information-for-a-category for estimating the importance of sentences.

1 Introduction

Estimating the importance of a sentence forms the core of a number of applications, including summarization, question answering, information retrieval, and topic detection and tracking. Existing approaches to importance estimation use different structural and content features of the textual unit, such as position of a sentence in a document, word overlap with section headings, and lexical features such as named entities or keywords that are characteristic of the document [7, 15, 16, 1]. We investigate a corpus-based method for estimating the importance of information for a given entity. Particularly, we propose to exploit the growing amount of manually labeled category information that is becoming available in the form of user-generated content (such as Wikipedia articles, or tags used to label blog posts) and/or through crowd-sourcing initiatives (such as the ESP game for labeling images, [19]). Given an entity of some category, we consider other entities of the same category and the properties that are *typically* described for them. That is, if a property is included in the descriptions of a significant portion of entities in the same category as our input entity, we assume it to be an important one.

To make things more concrete, let us look at an example. In Figure 1 we display a sample Wikipedia article, on *Rita Grande*, an Italian tennis player. The category information for this article is at the bottom of the article ("1975 births" and "Italian tennis players"). A few things are worth noting: compared to descriptions of other people (not just tennis players), the article has some typical biographical details (such as date and place of birth) but not all (e.g., there is nothing on education or marital status). On the other hand (as with other tennis players), the article contains a list of tournaments she has won in her career.

Rita Grande
Rita Grande (born March 23, 1975 in Naples, Italy) is a professional female tennis player from Italy.
WTA Tour titles (8)

- Singles (3)
  - 2001: Hobart, Australia
  - 2001: Bratislava, Slovakia
  - 2003: Casablanca, Morocco
- Singles finalist (1)
  - 1999: Hobart (lost to Chanda Rubin)
- Doubles (5)
  - 1999: 's-Hertogenbosch (with Silvia Farina Elia)
  - 2000: Hobart (with Emilie Loit)
  - 2000: Palermo (with Silvia Farina Elia)
  - 2001: Auckland (wth Alexandra Fusai)
  - 2002: Hobart (with Tathiana Garbin)

Categories: 1975 births — Italian tennis players

Figure 1. Sample Wikipedia article

In this paper we use Wikipedia, with its rich category information[1] as our starting point to explore the following proposition: information that is typically expressed in the descriptions of entities belonging to some category is information that is *important* for entities in that category. We test the viability of this proposition in a number of steps:

1. First, we determine whether there is a low divergence between descriptions of entities that belong to the same category—lower than beween descriptions of entities that belong to different categories.
2. Second, and building on a positive outcome for the

[1] We used the XML version of the English Wikipedia corpora made available by [4]. It contains 659,388 articles, and 2.28 categories per article, on average.

first step, we test whether typical-information-for-a-category coincides with *important* information for entities in a category. We examine this issue using Wikipedia itself, relating typical information to document structure and writing style.

3. Third, we test if typical-information-for-a-category coincides with *important* information for entities in another setting, viz. the TREC Question Answering track; here we use typical-information-for-a-category to distinguish between "vital or okay" sentences on the one hand, and "not okay" sentences on the other.
4. In our fourth and final step, we take the runs submitted for the "other" questions as part of the TREC 2005 Question Answering track, and use our "typical-information-is-important-information" strategy to filter out non-important snippets.

Our main contribution is two-fold. We introduce the idea of using the increasing amount of manually labeled category information to identify "typical information" for categories of entities. And we provide multiple types of empirical evidence for the usefulness of typical-information-for-a-category for estimating the importance of sentences.

The remainder of the paper is organized as follows. In the next section, we provide background information on importance estimation and Wikipedia related work in language technology. Section 3 provides empirical results of the within category similarity experiments (Step 1). Then, in Section 4 we relate "typical" information to "important" information within the setting of Wikipedia itself (Step 2). In Section 5 we do the same thing, but in the setting of the TREC QA track (Step 3), and in Section 6 we detail our reranking experiments (Step 4). We conclude in Section 7.

2 Background

Wikipedia has attracted much interest from researchers in various disciplines. While some study different aspects of Wikipedia itself, including information quality, motivation of users, patterns of collaboration, network structures, underlying technology, e.g., [22]. Others are interested in applying its content to solve research problems in different domains, e.g., question answering and other types of information retrieval [11, 13, 9]. Wikipedia has also been used for computing semantic word relations, named-entity disambiguation, text classification and other related activities, and as a document collection for assessing solutions to various retrieval and knowledge representation tasks, including INEX, and WiQA (CLEF 2006) Contest [3, 17, 4, 8].

Our interest in this paper is very specific: the task of estimating the importance of a sentence. Typically, sentence importance is modelled in terms of the importance of the constituting words. A commonly used measure to assess the importance of words in a sentence is the *inverse document frequency*. More advanced techniques define importance in a *centroid-based* manner: the most central sentences in a document (or cluster of documents) are the ones that contain more words from the centroid of the cluster [14]. More recently, *centrality-based* methods have been used, whose estimation relies on the concept of centrality in a cluster of documents viewed as a network of sentences [5].

We assume that information that is shared by multiple entities of the same category is likely to be important to entities of that category. This assumption is an extension of the idea that languages used in a specific domain are more constrained than the general language [6]. These contraints are realized in the syntactic structure and word co-occurence patterns in these structures. Furthermore, texts from the same domain have also been shown to exhibit a certain degree of content overlap [21]. Concretely, the descriptions of entities of a given category in Wikipedia typically consist of idiosyncratic information peculiar to each individual entity, and instances of a general class of properties applicable to all entities in the category—the properties we exploit. In order to test the validity of our assumption under different similarity measures, we apply different algorithms (KL-divergence and cosine similarity).

3 Is There Any Typical Information?

Before we can make use of information that is "typical" for entities of some category, we need to establish the fact that such typical information exists. Working on the English Wikipedia (see footnote 1), we proceed as follows: we compare the content of Wikipedia articles of entities within a single given category against Wikipedia articles outside this given category. We take a set of categories and compare their word distributions with word distributions of a corpus constructed from a set of randomly selected Wikipedia articles. Our aim, then, is to find out whether articles within a category look more like each other than like articles outside their category. To this end, we take the following steps:

- Select a random sample of C Wikipedia categories.
- From each category C_i, take a random sample of N articles, which we call the *category list*.
- From each category C_i, take a random sample of n ($n < N$) articles, called the *sample list* (S_i); we remove these articles from the category list.
- Take a random sample R of articles from the whole Wikipedia; this yields the *random list*.
- Construct language models using C_is, S_is, and R.
- For each category, compute the KL-divergences between C_is and S_is (the "within-category" KL-divergence), and between the C_is and R (the "outside-category" KL-divergence).
- Plot the resulting values.

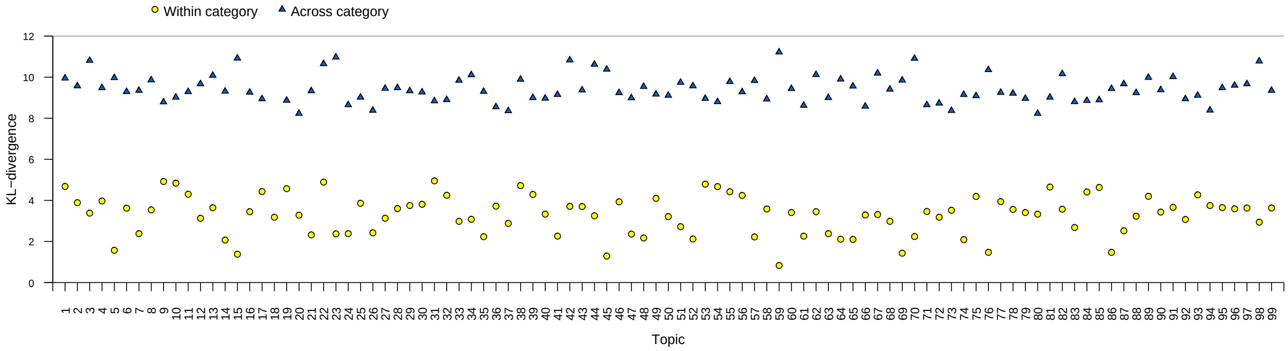

Figure 2. KL-divergence: within categories vs. across categories.

For our case study, we took a random sample of 100 Wikipedia categories. From each category, we took a random sample of $N = 100$ Wikipedia articles. We split these samples into two sets: a sample list consisting of $n = 30$ Wikipedia articles and a list consisting of the remaining 70 Wikipedia articles. As our random sample R we took 50 articles from Wikipedia. We induced three language models using the three types of lists. We used (smoothed) unigram language models of the following forms:

$$\begin{aligned} P_{C_i}(w) &= \lambda \cdot p_{ml}(w|C_i) + (1-\lambda) \cdot p(w|W) \\ P_{S_i}(w) &= \lambda \cdot p_{ml}(w|S_i) + (1-\lambda) \cdot p(w|W) \\ P_R(w) &= \lambda \cdot p_{ml}(w|R) + (1-\lambda) \cdot p(w|W). \end{aligned}$$

where $i = 1, \ldots, 100$, and $p_{ml}(w|C_i)$, $p_{ml}(w|S_i)$, $p_{ml}(w|R)$ are the maximum likelihood estimates of w given the lists C_i, S_i, and R, respectively, and $p(w|W)$ is the background likelihood based on the entire Wikipedia corpus [10]. In all cases, we set λ to be 0.9. We then computed the KL-divergence between $P_{S_i}(\cdot)$ and $P_{C_i}(\cdot)$, and between $P_R(\cdot)$ and $P_{C_i}(\cdot)$ using the formula given below, replacing X by P_{S_i} and P_R in each computation, respectively:

$$KL(X||P_{C_i}) = \sum_w X(w) \log\left(\frac{X(w)}{P_{C_i}(w)}\right)$$

The results are shown in Figure 2, where the lower graph shows the KL-divergence between the $P_{S_i}(\cdot)$ and $P_{C_i}(\cdot)$, whereas the upper graph shows the KL-divergence between $P_R(\cdot)$ and $P_{C_i}(\cdot)$. Notice that for each of the 100 categories that we sampled, the within-category KL-divergence is smaller than the outside KL-divergence: for our sample, the average within-category KL-divergence is 3.3, while the average outside-category KL-divergence is 9.43. We interpret this as saying that Wikipedia articles within a category look more like each other than like articles outside their category. That is, each Wikipedia category does indeed contain information that is typical for that category.

4 Typical Implies Important 1

Now that we have found evidence to support our claim that Wikipedia articles from a given category contain information that is typical for that category, we turn to the second of the steps outlined in the introduction: typical information is important. More formally, we check whether the likelihood of certain information within a category has a correlation with the importance of the information for articles/entities within that category. In this section, we use data we generated from Wikipedia for this purpose, based on Wikipedia's layout structure and authoring conventions.

Let us explain. It is well-known that the most important information in a news article appears first [18]. We argue that a similar phenomenon occurs in Wikipedia. According to Wikipedia's authoring guidelines,[2] articles need to be structured in such a way that the *lead* (which contains the leading one or more paragraphs) provides essential facts about the title, essentially serving as a summary of the article. Since the leads are created by humans we assume that there is a strong correlation between position (especially at the top positions) and importance in Wikipedia articles. For our quantitative analysis of typical vs important information, we exploit this assumption. Specifically, we take the first sentences of a Wikipedia article to be important, and check whether these sentences receive a higher likelihood within the corpus generated from the category information for that page than the remaining sentences on the page.

4.1 Ranking Sentences within an Article

We define an algorithm for ranking sentences from a given Wikipedia article that incorporates the typical information found in a category. Let j be a Wikipedia article, C_{kj} one of the categories assigned to j, and s_{ij} sentence i from article j. From each category C_{kj} assigned to j, we take a random sample of at most M articles. We then combine these samples and create a *category corpus*, CAT_j. We rank sentences based on the score they receive accord-

[2] See `http://en.wikipedia.org/wiki/Wikipedia:Guide_to_layout`. Even more detailed instructions have been provided for biography articles, specifying what type of essential facts should go into the lead, etc. See `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style_\%28biographies\%29`.

ing to the following language modeling function $\mu(\cdot,\cdot)$ due to [10], which compares the probability of a word within the category corpus against its a priori probability:

$$\mu(s_{ij}, CAT_j) = \sum_{w \in s_{ij}} \Big(P(w|s_{ij}) \cdot \log P(w|CAT_j) \quad (1)$$
$$- P(w|s_{ij}) \cdot \log P(w|W) \Big),$$

where w's are words in sentence s_{ij}, $P(w|s_{ij})$ is the maximum likelihood estimate of w in sentence s_{ij}, $P(w|CAT_j)$ is the likelihood in the category, and $P(w|W)$ is the background likelihood (estimated from all Wikipedia articles).

4.2 Experimental Setup

To test our hypothesis that typical information correlates with important information we consider the ranking produced by the scoring formula (1) and see to which extent it ranks important sentences before non-important sentences. We select a random sample of M ($M = 50$) Wikipedia articles from a category C. These articles constitute our *test sample* for which we rank the corresponding sentences. For an article j in the *test sample*, we take all the categories assigned to it and generate the *category corpus* as described above. We then perform the following steps:

- Select a Wikipedia article j from the *test sample*.
- Get all the sentences of j;
- Generate a unigram language model based on the *category corpus*;
- Score each sentence using the language model (based on the scoring function given in Eq. 1);
- Sort the sentences in descending order of their scores and take the top k sentences for further analysis.

We use R-precision as our evaluation measure: i.e., precision at rank R, where R is the number of important sentences for a given article [2]. Based on our earlier discussion, we simply take the first n sentences as our gold standard of important sentences within a Wikipedia article.

4.3 Results

We summarize the results of our experiments in Table 1. Looking at R = 5 we see that 42% of the top ranked sentences are among the first 5 sentences in Wikipedia. Given the fact that each article contains, on average, more than 20 sentences, this result provides evidence of a correlation between importance and likelihood within a category. Furthermore, a manual inspection of the results suggests that although the sentences ranked higher by Eq. 1 are mostly important, they may not necessarily appear at the beginning of the Wikipedia article.

R	1	2	3	4	5
R-precision	0.026	0.167	0.256	0.340	0.421

Table 1. R-precision scores for sentence ranking based on Eq. 1.

Overall, then, category-based ranking of sentences provides a clear bias that correlates with the importance of a sentence for the topic of the article.

5 Typical Implies Important 2

We present further empirical investigations of the correlation between typicality and importance of information within a category of objects. As in the previous section, we try to determine whether information that is typical for a category of entities (i.e., Wikipedia articles) is indeed important for that category, but instead of using "important" sentences from Wikipedia articles, we use "important" snippets identified by the TREC assessors as part of their assessments of the results submitted for the so-called "other" questions for the TREC 2005 QA track [20]. We take the outputs of the submitted runs for the TREC 2005 QA track for "other" questions and split them into two classes; "at least okay" sentences and "not okay" sentences. Applying our method on the TREC runs shows the relative merits of our approach, since these runs come from different systems that implement different approaches.

Using this data, we conducted two experiments to explore whether our notion of typical information can be used as a useful heuristic for identifying important snippets. In this section, we present an experiment that relates typical information to "at least okay" information, and in the next we build on this relation to filter out non-important snippets.

5.1 Scoring Sentences

We work on the user submitted runs for TREC 2005 QA track for "other" questions. We take those topics that have a corresponding Wikipedia article, and for each such topic we take all categories and generate a *reference corpus* (capturing typical information) as described below. We then compute the similarity between the set of "at least okay" sentences and the sentences in the reference corpus. We repeat the same steps for "not okay" sentences and then compare the results of these two steps.

We use the following aggregative method of computing the similarity between sentences of the target topic, and sentences in the reference corpus [12]. Let t be a topic, $s_1, \ldots, s_{m_t}$ the assessed sentences for t (labeled "at least okay" or "not okay"), and $r_1, \ldots, r_N$ the sentences in the reference corpus. For each assessed sentence s_i, we compute its cosine similarity with all sentences in the reference corpus, and take the mean of the scores:

$$corpus\_score(s_i) = N^{-1} \cdot \sum_{j=1}^{N} cosine(s_i, r_j) \quad (2)$$

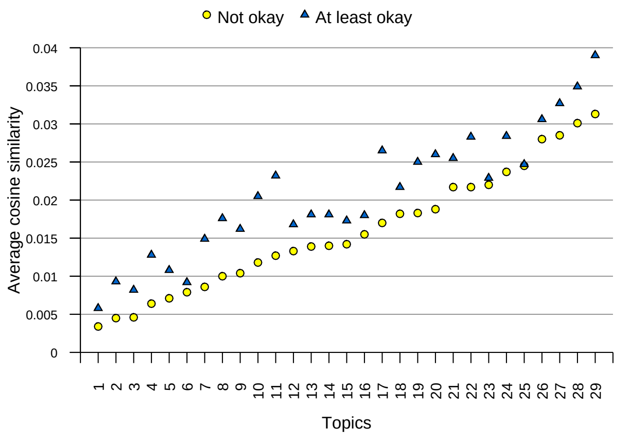

Figure 3. Average similarity scores for "at least okay" and "not okay" sentences on the one hand and reference sentences on the other; topics ordered by avg. score on "not okay" sentences.

Then, we aggregate these scores in two groups: one for the "at least okay" sentences and one for the "not okay" sentences by taking the average of the scores per group.

5.2 Experimental Setup

Our experiment is meant to answer the following question: Can the typical information captured in the reference corpus distinguish between "at least okay" sentences on the one hand, and "not okay" on the other hand? That is, are the sentences in the reference corpus more similar to the former, the latter, or is there no observable difference?

We took 29 topics from TREC 2005 QA track that have corresponding Wikipedia articles. For each topic, we generated a *reference corpus* capturing its associated typical information in a very straightforward manner: we take all Wikipedia categories assigned to the topic and from each category, we select a random sample of k_i Wikipedia pages, and fixed the total number of selected pages to be 50.

5.3 Results

Fig. 3 plots the scores, with two data points per topic, one for the "not okay" sentences, and one for the "at least okay" sentences. Clearly,"at least okay" sentences are strictly more similar to sentences from the reference corpus than "not okay" sentences for almost all the topics. Put differently, the sentences in the reference corpus allow to distinguish between the two, using a simple similarity measure.

6 Filtering Snippets

In this section, we turn to our fourth and final experiment: we take the runs submitted for "other" questions as part of the TREC 2005 QA track, and for each run we filter out the non-important snippets using the "typical-for-a-category" idea introduced earlier.

The filtering method works as follows. Let $Run_1, \ldots, Run_c$ be the submitted runs ($c = 58$), $t_1, \ldots, t_k$ the TREC topics that have Wikipedia entries ($k = 29$), $s_1^{tr}, \ldots, s_m^{tr}$ the sentences returned for topic t as part of run Run_r, and $r_1, \ldots, r_N$ are the same as before. Given a submitted snippet s_l^{ij} (for topic t_j from run Run_i), we define its $corpus\_score(s_l^{ij})$ as in Eq. 2:

$$corpus\_score(s_l^{ij}) = N^{-1} \cdot \sum_{n=1}^{N} cosine(s_l^{ij}, r_n) \qquad (3)$$

Once we have computed the scores for each snippet, we sort the snippets for a topic of a given run in decreasing order of their scores and retain only those snippets s for which $corpus\_score(s)$ exceeds a certain threshold (we used 0.02 for the experiments below).

Following the TREC QA track procedure for scoring responses to "other" questions, we score the runs in terms of F-score (based on the "vital" snippets in the ground truth made available by TREC), only using the topics that have a Wikipedia entry. Fig. 4 lists the results of the filtering experiment. Filtering based on similarity to typical information for a category (as formalized in Eq. 3) nearly always improves the F-score (producing improvements in 54 out of 58 runs). The cases where our filtering leads to a reduction in F-scores were runs that returned short snippets or named entities as important facts for the topic—this is no surprise as our method depends on lexical overlap and assumes longer snippets with richer lexical information. In sum, then, similarity to typical information is a simple but effective way of filtering out non-important information.

7 Conclusion

We introduced the idea of using information that is typically provided about entities within a category for estimating the importance of sentences about entities within the same category. We presented a step-by-step analysis of the idea and of ways of exploiting it. First, we used manually labeled category information to identify "typical information." Second, we showed that there exists a positive correlation between "typical information" within a category and its importance for entities of that category. Finally, we developed a method of ranking snippets based on these ideas.

Our method exploits the rich category information that is found in Wikipedia. It is sufficiently generic, however, to be applicable to other document collections with similar properties: i.e., with multiple documents about entities that are grouped in categories, either explicitly or implicitly through, for instance, user provided tags. Our methods assume that the category of a given entity is known or there is system that accurately predicts a category information. We are currently working on methods for mapping an entity to the most appropriate Wikipedia category.

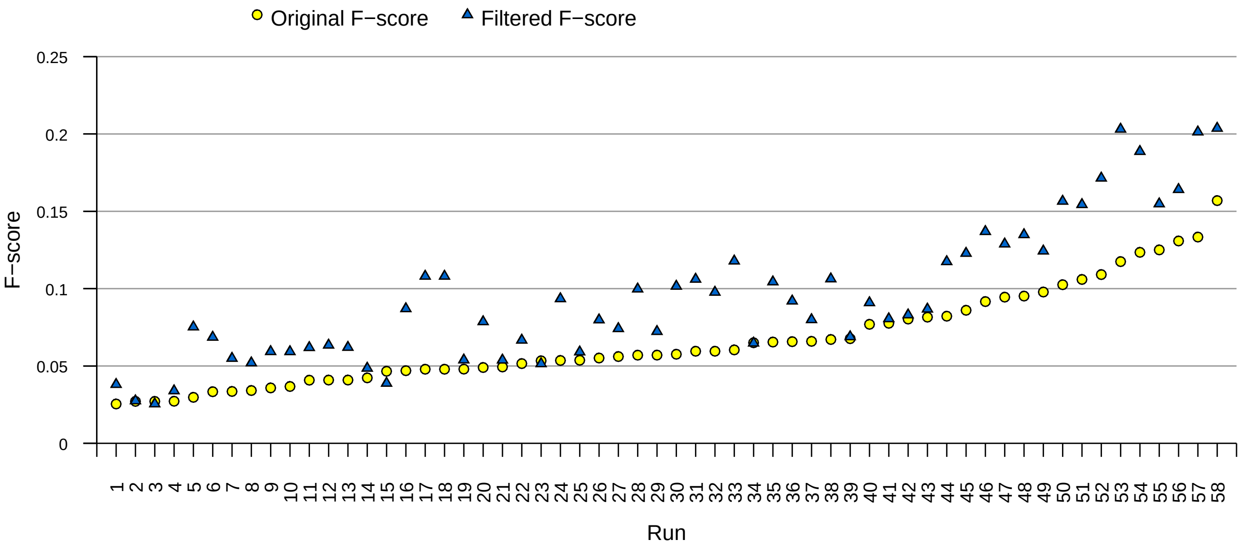

Figure 4. Improvement in F-score obtained by filtering the original runs submitted for the TREC 2005 QA track; scored using "vital" snippets only, and only taking topics with a Wikipedia entry into account. Runs ordered by performance without filtering, i.e., original TREC scores.

8 Acknowledgements

This research was supported by the Netherlands Organisation for Scientific Research (NWO) by a grant under project number 220-80-001.

References

[1] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *SIGIR 2003*, pages 314–321, 2003.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] R. Bunescu and M. Pasça. Using encyclopedic knowledge for named entity disambiguation. In *EACL-06*, pages 9–16, 2006.

[4] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[5] G. Erkan and D. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479, 2004.

[6] Z. Harris. *A Theory of Language and Information: A Mathematical Approach*. Clarendon Press, 1991.

[7] W. Hildebrandt, B. Katz, and J. Lin. Answering definition questions with multiple knowledge sources. In *HLT-NAACL*, pages 49–56, 2004.

[8] V. Jijkoun and M. de Rijke. Overview of WiQA 2006. In A. Nardi, C. Peters, and J. Vicedo, editors, *Working Notes CLEF 2006*, September 2006.

[9] B. Katz, G. Marton, G. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller, O. Uzuner, and A. Wilcox. External knowledge sources for question answering. In *TREC 2005*, 2005.

[10] W. Kraaij and M. Spitters. Language models for topic tracking. In B. Croft and J. Lafferty, editors, *Language Models for Information Retrieval*. Kluwer Academic Publishers, 2003.

[11] L. Lita, W. Hunt, and E. Nyberg. Resource analysis for question answering. In *Companion Volume to ACL 2004*, pages 162–165, 2004.

[12] D. Metzler, Y. Bernstein, W. Croft, A. Moffat, and J. Zobel. Similarity measures for tracking information flow. In *CIKM 2005*, pages 517–524, 2005.

[13] G. Mishne, M. de Rijke, and V. Jijkoun. Using a reference corpus as a user model for focused information retrieval. *J. Digital Information Management*, 3(1):47–52, 2005.

[14] D. Radev, H. Jing, M. Sty, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938, 2004.

[15] B. Schiffman, I. Mani, and K. Concepcion. Producing biographical summaries: combining linguistic knowledge with corpus statistics. In *ACL 2001*, pages 458–465, 2001.

[16] I. Soboroff and D. Harman. Novelty Detection: The TREC Experience. In *HLT/EMNLP 2005*, pages 105–112, 2005.

[17] M. Strube and S. Ponzetto. WikiRelate! Computing semantic relatedness using wikipedia. In *AAAI '06*, July 2006.

[18] T. van Dijk. *News as Discourse*. Lawrence Erlbaum Associates, Inc., 1988.

[19] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04*, pages 319–326, 2004.

[20] E. Voorhees and H. Dang. Overview of the TREC 2005 Question Answering Track. In *TREC 2005*, 2006.

[21] A. Wray. *Formulaic Language and the Lexicon.* Cambridge University Press, 2002.

[22] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1), 2006.