

ZED: Explaining Temporal Variations in Query Volume

Maojin Jiang¹, Shlomo Argamon¹, Abdur Chowdhury² and Kush Sidhu²

¹ Laboratory of Linguistic Cognition, Illinois Institute of Technology, Chicago, USA
jianmao@iit.edu, argamon@iit.edu

² America Online, Inc., USA
Cabdur@aol.com, KSidhu35@aol.com

Abstract. We hypothesize that the variance in volume of *high-velocity queries* over time can be explained by observing that these queries are formulated in response to events in the world that users are interested in. Based on it, this paper describes a system, ZED, which automatically finds explanations for high velocity queries, by extracting descriptions of relevant and temporally-proximate events from the news stream. ZED can thus provide a meaningful *explanation* of what the general public is interested in at any time. We evaluated performance of several variant methods on top velocity “celebrity name” queries from Yahoo, using news stories from several sources for event extraction. Results bear out the event-causation hypothesis, in that ZED currently finds acceptable event-based explanations for about 90% of the queries examined.

1 Introduction

Web search is the second most popular activity on the Internet, exceeded only by e-mail. What people are searching for changes over time, due to cycles of interest and events in the world [1]. Of particular interest are *high-velocity* queries, which show a sharp increase in popularity over a short time (*velocity* refers to relative change in number of searches of a query during a given period.). Such queries may reflect important specific events in the world, as people hear about the current news and then search for more information on significant events (Fig. 1).

This paper describes a new system, ZED¹, which, given a celebrity name which is a top velocity query on a given day, automatically produces a description of a recent event involving that celebrity which is likely to be of widespread interest. We call such events *query-priming events*. The set of such explanations can serve on its own as a summary of the current ‘zeitgeist’, or may be used to index into more information about the events and individuals currently deemed most interesting by the querying public.

To the best of our knowledge, our specific task of explaining peak queries has not been addressed previously. We briefly discuss here prior work on extractive text summarization which ZED’s event extraction algorithm is based on. One relevant work is multiple document summarization [2] and its application to news [3]. Event-based text summarization [4, 5] focuses on news summarization, which addresses the issues to detect events in news and to maintain a complete description of event constituents. Another relevant work on query-based summarization [6, 7] takes a user query into account (as in our task) such that summarization is made only on relevant text.

¹ Zeitgeist Event Detector.

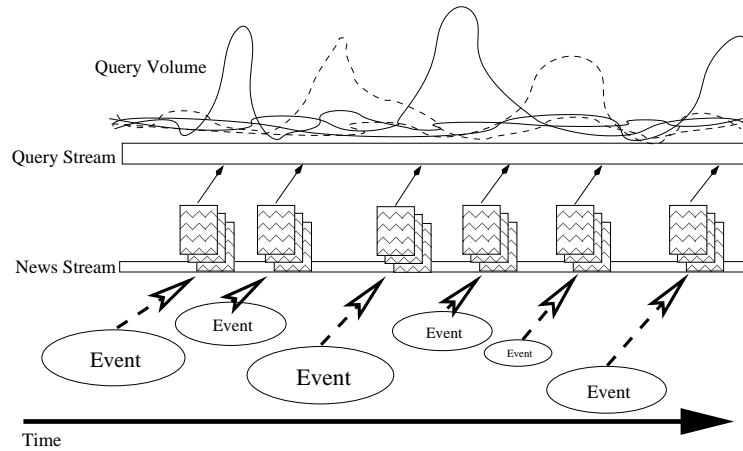


Fig. 1. The ‘event causation’ model of query volume variability. Events in the real world (of greater or lesser significance) cause news stories to be written, which spur user interest in the events, leading to sharp volume increases for related queries.

2 The Approach

Our approach is to enhance traditional text summarization methods with specific cues for finding query-focused sentences describing relevant query-priming events. We first give an overview of the ZED system architecture, then discuss its components in more detail, including variant methods for certain subtasks.

2.1 System Architecture

ZED comprises three components: *document indexing*, *query searching*, and *event extraction*. The system works as follows. First, news stories coming in from the stream are indexed. Then, each top velocity query containing a celebrity name is fed into a search engine to retrieve the most relevant news stories, restricting attention to stories which (a) came in within the last k days, and (b) contain somewhere the full query. After a set of most-relevant stories is obtained, each of these stories file split into individual sentences, and certain irrelevant sentences filtered out (e.g., numerical lists). Then sentences are ranked for rough relevance, based on a linear combination of three measures (as in MEAD [2]): *centroid distance*, *query relevance*, and *sentence position*. Sentences with high overall scores are then processed by one of several *sentence selection* criteria to choose the ‘best’ sentence purporting to describe the event which prompted the input query to suddenly rise in volume. Finally, the headline from the story in which this top sentence is found is returned together with the sentence as the explanation for the query under consideration. The various modules within ZED are described in more detail next.

2.2 News Story Retrieval

The AIRE IR engine [8] is used to index and retrieve news stories. News story files are parsed, tokenized, and downcased with stopwords removed, but without stemming. Token positions are recorded to facilitate phrase searching for the full name in the query.

For retrieval, only queries which constitute a celebrity name *in toto* are considered. Query tokens are therefore not stemmed or elided. Only story files containing the full query as a phrase at least once are considered as possibly relevant. This serves to remove many false positives, and is based on the assumption that a complete news report on a person should mention the full name at least once. Also, only stories with a post date within the two days prior to the query date are considered since we hypothesize that a top-velocity name query reflects only latest events in news and two days should be a reasonable approximation. Relevance ranking is then done using the individual query terms via the standard bm25 [9] ranking function, giving headline words more weight by adding 10 to their term frequencies. This simple heuristic rule originates in our observation that a news report is more focused on events of a celebrity if his or her name appears in the headline. The top 10 retrieved documents are then passed on to the next module.

2.3 Sentence Splitting and Filtering

All relevant news story files returned from previous step are then segmented into sentences. Since many articles contain ‘sentences’ that aren’t really descriptive (lists of sports scores, for example), we remove at this stage all sentences which:

- Contain fewer than five words (this removes too short sentences in that they often do not contain enough information about an event), **or**
- Contain fewer than two function words, based on a preconstructed list of 571 common English function words (this ensures the sentence is regular English text), **or**
- Contain no query terms.

The position $\text{pos}(s)$ of each sentence s in the story is then recorded. This set of tagged sentences is then passed on to the sentence ranking module.

2.4 Sentence Ranking

The set of candidate event description sentences is then ranked by a quick approximate quality measure based on previous methods for multi-document summarization. We use a combination of three measures of sentence quality that are based on those used by the MEAD summarization system [2]:

Centroid similarity: Here we compute the word-frequency ‘centroid’ of the retrieved set and measure relevance of each sentence based on its similarity to this centroid. The centroid is computed by computing, for each content word in the retrieved set, the sum of its tf-idf score over all 10 documents. The 80 words with the highest scores are retained, and the resultant vector of tf-idf sums is the ‘centroid’ of the document set. Then the centroid similarity $f_C(s)$ for each sentence s is computed as its cosine distance from the centroid.

Query similarity: The query similarity measure $f_q(s)$ evaluates how directly relevant a sentence s is to the input query. It is calculated as the tf-idf cosine distance between a sentence and the input query.

Sentence position: The third measure is a LEAD-like measure of importance based on the idea that the more important sentences tend to appear near the beginning of news stories. The position measure, $f_P(s)$, is defined as reciprocal of square root of $pos(s)$, giving higher values to sentences nearer the beginning of a story.

After these three quality measures are computed for each sentence s , its overall score is computed by a manually-adjusted linear combination of the measures:

$$f(s) = 8f_C(s) + 12f_q(s) + f_P(s) \quad (1)$$

As in MEAD [2], the parameters in (1) are obtained by manually examining different scores of some sample sentences by using different coefficient values. In the future, optimal values of them may be set by applying some machine learning approach to well-established human-made summaries. The sentences are then ranked according to $f(s)$, duplicates are removed, and the top 10 sentences selected for further processing.

2.5 Sentence Selection

The next step is to select, from the candidate sentences provided by sentence ranker, the most likely sentence to constitute an effective explanation as a query-priming event description. We compare three methods for selecting the final event describing sentence:

Ranking: Select the sentence with the highest ranking score $f(s)$.

Recency: Select the most recent sentence from the candidate set of most highly ranked sentences, breaking ties by $f(s)$.

Grammar: This strategy uses a heuristic method for evaluating sentences based on its syntactic structure. Robust Accurate Statistical Parsing (RASP) system [10] is used to parse each candidate sentence, giving a syntactic parse tree represented as a set of grammatical dependency relations between sentence words. Based on the parse tree, three features of the sentence’s syntactic structure are determined. First is the sentence’s *complexity*—if it consists of a single clause it is ‘simple’, otherwise it is ‘complex’. Second is the *position* of the query term(s) in the sentence—whether in the main clause or a sub-clause. Third is the syntactic *role* of the query term(s) in the sentence, whether in the subject, object, a prepositional phrase (‘pp’), or other syntactic constituent.

The idea is that these features give some indication of the usefulness of the sentence under consideration as a description of a query-priming event. Based on this sort of consideration, we manually constructed the preference ordering for the various possible feature combinations, as shown in Table 1.

After the ‘best’ sentence is selected according to one of the above criteria, it is returned with the headline of its story as the final event description.

2.6 Comparison Baselines

ZED was compared to two baseline methods. The first, termed FirstSentence, based on the assumption that journalists will tend to put a summary sentence at the beginning

Table 1. Preference order for sentences based on grammatical features. The candidate sentence with the highest position in this order is chosen as a priming event description.

Complexity	Position	Function	Complexity	Position	Function	Complexity	Position	Function
1. simple	main	subject	5. simple	main	pp	9. complex	sub	pp
2. complex	main	subject	6. complex	main	pp	10. simple	main	other
3. simple	main	object	7. complex	main	object	11. complex	main	other
4. complex	sub	subject	8. complex	sub	object	12. complex	sub	other

of a news report, simply chooses the first sentence from the most relevant story, together with the story’s headline, as the event description. The second baseline method, FirstRelevant, incorporates the constraint that the explanation contain the name in the query. It chooses the earliest sentence from the top-ranked document that contains at least one of the query terms, together with the story’s headline, as the event description.

3 Evaluation

3.1 Corpus, Testbed and Assessment Metrics

The corpus we used consisted of news stories from the month of September 2005, downloaded from six on-line news-sources: AP, BBC, CNN, NYTimes, Time and Yahoo, in four relevant categories: Top Stories, Entertainment, Sports, and Obituaries. There were a total of 25,142 articles in the full corpus, giving nearly 900 on average per day. In experiment, after top 10 relevant stories are split into sentences and after sentence filtering, on average, a query gets 24 candidate sentences for ranking and selection.

We constructed an evaluation testbed by taking all celebrity name queries from Yahoo Buzz’s “top velocity” queries for the week of 5 September through 11 September, 2005, termed *FirstWeek* and comprising 141 queries, and the week of 24 September through 30 September, 2005 termed *LastWeek* and comprising 128 queries.

Summaries generated by each of the five methods above (three ZED variants and the two baselines) were evaluated (using randomized ordering) by two human assessors (referred to as A and B respectively). Three types of metrics were used for evaluation:

Relevance: This is a traditional binary metric, wherein each explanation is adjudged either relevant or irrelevant to its query. The relevance criterion was whether the summary refers (even tangentially) to a possible priming event for that query.

Ranking: In this metric, the five explanations from different methods that were returned for a query were ranked from 1 to 5, with 1 being the best, and 5 being the worst. Irrelevant explanations were not ranked. Identical explanations, which were rather common, received the same rank, then the next one down received the appropriate rank; e.g., when two explanations both received rank 1 next best received rank 3.

Quality: Finally, the quality of each explanation was evaluated numerically according to three criteria: *completeness*, *sufficiency*, and *conciseness*. Each of these was evaluated on a coarse 3-point scale (from 0 to 2), as follows:

Completeness refers to how fully the explanation describes an event involving the query individual. A score of 2 meant a complete explanation of some event involving the individual (even circumstantially), a score of 1 meant that some essential

information about the event was missing, and a score of 0 meant that it did not describe an event involving the query individual at all.

Sufficiency refers to how strongly the event described would motivate people (those interested in the queried individual) to use the specific query, based on the importance of the event and centrality of the individual's participation in it. A score of 2 meant a strong motivation for querying on that individual, 1 meant that there would be just weak possible motivation to query, and 0 meant that there would be no such motivation.

Conciseness refers to how much extraneous information is included in the explanation besides the relevant event (if present). A score of 2 meant that nearly all the information in the explanation was about the relevant event, 1 meant that there was much irrelevant information, though the explanation was mainly about the relevant event, and 0 meant that the central focus of the explanation was on irrelevant matters.

These three values were summed, to give a total *Quality* score, ranging from 0 to 6.

3.2 Results and Discussion

Event descriptions generated by the five methods for each of the two weeks' data were evaluated separately, so we could also evaluate consistency of system performance for news of somewhat different periods of time.

We first evaluated the consistency of evaluation between the two raters, examining agreement percentage and Cohen's kappa [11]. For relevance, agreement percentages were 97.6% and 89.1% for FirstWeek and LastWeek, respectively, corresponding to kappa values of 0.91 and 0.69. For ranking, where ratings from an ordered set of values, linearly-weighted kappa [12] was used. Agreement percentages were lower at 73.0% and 78.3% for FirstWeek and LastWeek, with linearly-weighted kappas of 0.63 and 0.65. These results show substantive agreement between the raters; in what follows, we present results just for rater A.

Table 2 shows the precision for each variant, where precision is defined as the ratio of the number of relevant explanations returned over the total number of explanations returned. Results are slightly different for the two weeks studied. For FirstWeek, Grammar is beaten by Ranking, though the difference is small; all ZED variants do clearly improve over the baselines. On LastWeek, Grammar has a definite advantage, and the other two ZED variants are not clearly better than just choosing the first-ranked relevant sentence. Overall, the Grammar variant perhaps has a slight edge. It is clear, however, that using the first sentence of the most relevant document is not useful.

Histograms showing the rank distributions for the five variants are given in Fig. 2. We first of all see that, as precision also indicated, the FirstSentence heuristic does not work very well at all. Among the other four methods, however, we see a difference between the two weeks that were evaluated, with Grammar dominating for FirstWeek and FirstRelevant dominating for LastWeek. The good performance of FirstRelevant may lie in the hypothesis that we infer that many reporters tend to mention the name of major character in the first sentence that describes the major event in a news report. However, it is unclear to what extent these differences are really significant, particularly in view of the results of the quality evaluation, given in Table 2. These results show a different

Table 2. Average overall explanation precision (as well as with both week’s data pooled), quality, and average completeness, sufficiency, and conciseness, for the five query explanation methods.

Method	FirstWeek					LastWeek					Pooled Prec.
	Prec.	Qual.	Comp.	Suff.	Conc.	Prec.	Qual.	Comp.	Suff.	Conc.	
Grammar	0.972	3.14	1.03	1.24	0.87	0.852	3.65	1.13	1.28	1.23	0.915
Recency	0.972	3.13	1.09	1.21	0.83	0.828	3.41	1.08	1.19	1.14	0.903
Ranking	0.979	3.29	1.15	1.25	0.89	0.813	3.59	1.12	1.25	1.21	0.900
FirstRelevant	0.950	3.27	1.15	1.26	0.86	0.828	3.63	1.16	1.24	1.23	0.892
FirstSentence	0.234	1.13	0.37	0.46	0.3	0.375	1.75	0.57	0.62	0.56	0.301

pattern, with Grammar attaining much higher quality for LastWeek, while Ranking is preferred somewhat for FirstWeek. Grammar’s quality dominance in LastWeek is due to greater sufficiency and conciseness, perhaps attributable to its ability to distinguish simple from complex sentences, and to ensure that the query appears in a salient syntactic position in the sentence. Regardless, all of the methods (other than FirstSentence) appear to find reasonable event descriptions, on average.

Overall, the three methods implemented in ZED outperform both baselines, though only FirstRelevant is a competitive baseline. FirstSentence’s abysmal performance indicates that a LEAD-like system will not work for this task. Among ZED’s three methods, however, results are inconsistent, though Recency does seem less useful than the more content-based measures. Grammatical cues do seem to provide some particular leverage, though more work is needed to elucidate this point. Future work will clearly need to address the development of larger evaluation sets for this task as well as examining inter-rater reliability measures for the evaluations.

4 Conclusions

We have presented ZED, a system which addresses the novel task of finding explanations of query-priming events in a news stream. This system thus provides an alternative view of “What is the interesting news today?” based on what recent events users as a whole have found most compelling. In the future, we intend to explore methods for improving the quality of ZED’s event descriptions. A more precise characterization of the circumstances in which one or another of the selection methods is preferred, if one can be found, may lead to improvements (for example, by helping us refine syntactic preferences). Also, lexical cues (such as ‘died’, ‘just released’) may help the system to recognize ‘significant’ events. Supervised machine learning may also be applied to build better models to combine evidence from different cues.

References

1. Chien, S., Immorlica, N.: Semantic similarity between search engine queries using temporal correlation. In: Proc. WWW-05, Chiba, Japan (2005) 2–11
2. Radev, D., Blair-Goldensohn, S., Zhang, Z.: Experiments in single and multidocument summarization using MEAD. In: Proc. Document Understanding Conference. (2001)

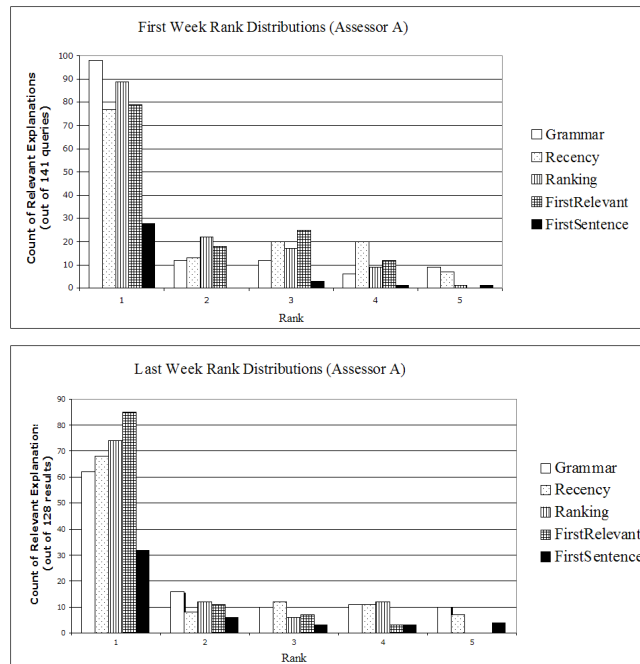


Fig. 2. Histograms of the respective ranks (by assessor A) of the explanations generated by the five event extraction methods for both weeks' data.

3. Radev, D.R., Blair-Goldensohn, S., Zhang, Z., Raghavan, R.S.: Newsinessence: a system for domain-independent, real-time news clustering and multi-document summarization. In: Proceedings of HLT '01. (2001) 1–4
4. Filatova, E., Hatzivassiloglou, V.: Event-based extractive summarization. In: ACL Workshop on Summarization, Barcelona, Spain (2004)
5. Vanderwende, L., Banko, M., Menezes, A.: Event-centric summary generation. In: Proc. Document Understanding Conference at HLT-NAACL, Boston, MA (2004)
6. Saggion, H., Bontcheva, K., Cunningham, H.: Robust generic and query-based summarization. In: Proceedings of EACL '03. (2003) 235–238
7. Amini, M.R.: Interactive learning for text summarization. In: Proceedings of the PKDD'2000 Workshop on Machine Learning and Textual Information Access. (2000) 44–52
8. Chowdhury, A., Beitzel, S., Jensen, E., Sai-lee, M., Grossman, D., Frieder, O., et. al.: IIT TREC-9 - Entity Based Feedback with Fusion. TREC-9 (2000)
9. Robertson, S.E., Walker, S., Hancock-Beaulieu, M.: Experimentation as a way of life: Okapi at TREC. *Information Processing and Management* **36** (2000) 95–108
10. Briscoe, E.J., Carroll, J.: Robust accurate statistical annotation of general text. In: Proceedings of LREC. (2002) 1499–1504
11. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. **20** (1960) 37–46
12. Maclure, M., Willett, W.: Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*. **126** (1987) 161–169