Large scale testing of a descriptive phrase finder

Hideo Joho
Department of Information Studies
University of Sheffield, Western Bank
Sheffield, S10 2TN, UK
+44 (0)114 222 2675
h.joho@sheffield.ac.uk

Ying Ki Liu
Department of Information Studies
University of Sheffield, Western Bank
Sheffield, S10 2TN, UK

Mark Sanderson
Department of Information Studies
University of Sheffield, Western Bank
Sheffield, S10 2TN, UK
+44 (0)114 222 2648
m.sanderson@sheffield.ac.uk

ABSTRACT
This paper describes an evaluation of an existing technique that locates sentences containing descriptions of a query word or phrase. The experiments expand on previous tests by exploring the effectiveness of the system when searching from a much larger document collection. The results showed the system working significantly better than when searching over smaller collections. The improvement was such, that a more stringent definition of what constituted a correct description was devised to better measure effectiveness. The results also pointed to potentially new forms of evidence that might be used in improving the location process.

Keywords
Information retrieval, descriptive phrases, WWW.

1. INTRODUCTION
Retrieving descriptions of the words and phrases, which are not often found in dictionaries, has potential benefits for a number of fields. The Descriptive Phrase Finder (DPF) is a system that retrieves descriptions of a query term from free text. The system only uses simple pattern matching to detect a description, and ranks the sentences that hold the descriptive phrases based on within document and cross document term occurrence information. The system does not attempt to extract descriptions from text, it simply locates sentences that are hopefully relevant to a user. It is assumed that users are able to read a sentence and locate any description within it. The advantage of using such an approach is that the DPF is much simplified and does not require parsing to find the exact location of the phrase. Due to its simplicity, it achieves a level of domain independence.

The DPF was implemented and succeeded in The DPF was implemented and succeeded in retrieving sentences holding descriptive phrases (DPs) of a wide range of proper nouns. Initial testing on a collection of LA Times articles from the TREC Collection showed that 90% of the queries had at least one correct DP in the top 5 ranked sentences and 94% in the top 10 ([3]). It was shown that the effectiveness of the system was in part due to the large amount of free text being searched. What was not shown by the experiment was if performance could be further improved by searching an even larger text. Consequently, a larger scale experiment was conducted, searching for phrases from the World Wide Web (WWW) using the output of a commercial Web search engine to locate candidate documents that were then processed locally by the DPF.

In addition to increasing the number of documents searched, more queries were tested and different definitions of relevance were tried. The rest of this short paper explains the system and shows the results of the expanded experiment, followed by pointers to future work.

2. THE SYSTEM
The Web-based DPF was composed of two parts: a front-end to an existing Web search engine, which fetched documents; and the system that located sentences holding descriptive phrases.

The Web front end simply routed queries to a Web search engine (Google), and the text of the top 600 documents returned by the engine was fetched, split into sentences (using a locally developed sentence splitter), and those sentences holding the query term were passed onto the DPF.

It ranked sentences on a score calculated from multiple sources of evidence. A detailed description of the DPF is found in [3]. The primary clue to there being a descriptive phrase in a sentence was the presence of a *key phrase* within it. An example key phrase was "such as", which may be found in the sentence: "He used several search engines *such as* AltaVista, HotBot and WebTop to compare the performance". If such a sentence were returned to a user who entered the query "WebTop", they would determine it was a search engine. Specifically, the DPF is searching for the key phrase in proximity to a query noun *(q*n) to locate a descriptive phrase *(d*p) e.g.

- ... *dp* such as *qn* ...

other key phrases used, some suggested by [2], were

- ... such *dp* as *qn* ...
- ... *qn* (and | or) other *dp* ...
- ... *dp* (especially | including) *qn* ...
- ... *qn* (*dp*) ...

- ... *qn* is a *dp* ...
- .. *qn*, (a | the) *dp*, ...

The phrases form the key part of the DPF as they identify well sentences likely to contain descriptions of *q*n. While the number of times a particular *qn* appears in a sentence with a key phrase are small, by searching a large corpus, like the Web, the chances of finding a few (accurately identified) descriptions of *qn* in the form required are high.

Based on results from a testing phase, certain key phrases were found more accurate at locating a descriptive phrase than others. Consequently, when ranking matching sentences, different scores were assigned depending on the accuracy of the key phrase found within. Since unfamiliar words tend to be explained or rephrased at the early part of a document, sentence position was also a factor in the rank score, with earlier sentences given preference. Finally, cross-document information was taken into account. Across all the matching sentences for a particular query, the occurrence of all the terms within the sentences was noted. It was anticipated that terms occurring more frequently within the set of sentences were likely to belong to descriptions.

Consequently, sentences holding a high number of commonly occurring words were given further preference in the ranking. The last two pieces of information not only improved the accuracy of ranking, but also enabled the system to produce reasonable results when no key phrases were matched. A training phase where the optimum balance between the sources of information was run on existing training data created from the LA Time corpus described in [3].

It may be reasonable to question why such a simple approach to extracting information from free-text sources be taken when more principled NLP-based techniques are well-established (e.g. [4], [5]). There are a number of reasons:

- Any simple approach is likely to be much faster than one that requires operations such as parsing.
- We believe that the use of simple but accurate methods searching over very large corpora provides a new means of determining lexical relations from corpora that are worthy of further exploration.

3. INITIAL STUDY

A pilot study was conducted, searching ten queries using the top hundred documents returned by Google. Of the ten queries, six had the best description located in the top two ranked sentences, two more queries had a good description in the top two. For all queries, a sentence holding a descriptive phrase was returned in the top five ranked sentences.

4. DEFINING RELEVANCE

In this and the previous evaluation described in [3], relevance was defined as a sentence that told the user anything about the query term: a liberal view of relevance (described here as *binary relevanc*e). The results from the pilot, under this interpretation, showed the system performed well. Consequently a more stringent form of relevance was devised. A sample answer for each query was solicited from users: for example, "the Prime Minister of Great Britain" for Tony Blair. Those *key answers* were taken as an acceptable criterion of highly relevant descriptive phrases. Sentences ranked by the system were then compared to the key answer. Correctness of DPs is not enough for this aim. Only a DP that described a query as well as a key answer was regarded as relevant. To illustrate, the sentence "Tony Blair is the current Prime Minister of the United Kingdom." was regarded as relevant, but "Tony Blair is a political leader" was not.

5. THE MAIN EXPERIMENT

A total of 146 queries were tested in the main experiment: 50 of which were evaluated based on key answers; 96 using binary evaluation. In the binary test, the DPF returned a relevant (descriptive) sentence in the top twenty sentences for all 96 queries. On average sixteen of the sentences returned were relevant to each query. The minimum number of relevant was six and maximum was twenty. Across the 96 queries, at least one relevant sentence was found in the top five for every tested query. This is a significant improvement over the previously reported experimental results where 90% of queries were answered in the top five.

Using more stringent key answer based relevance, the system succeeded in retrieving at least one relevant sentence in the top five for 66% of the queries, at least one in the top ten for 82%, and one in the top twenty for 88%.

These results show that the DPF searching the Web (1 billion documents) works dramatically better than the previous experiment using LA Times (100,000 documents). As was shown in previous work, the size of the collection impacts on the effectiveness of the system. This is because by searching a larger collection, there is a better chance of locating a relevant descriptive phrase in the format of one of the searched for key phrases. However in the previous work, there appeared to be an upper bound on the accuracy of the descriptive phrases alone. By searching a much larger collection it is speculated that the cross document term occurrence statistics used contributed significantly to improving the effectiveness of the system.

6. CONCLUSION

An existing descriptive phrase system was adapted to work with a Web search engine to locate phrases describing query words. The system was found to be highly effective at locating good descriptions: finding at least one high quality descriptive phrase in the top 10 returned sentences for 82% of test queries.

7. FUTURE WORK

We plan to undertake a number of further experiments, examining through tests, the ability of people to locate descriptions within the retrieved sentences. In addition, it was notable that the results of the full experiment were not as good as those from the pilot study. One difference between the two tests was the number of web documents examined: 100 top-ranked documents in the pilot; 600 for the expanded experiment. Given that a search engine generally retrieves more relevant documents in the higher ranks, there is likely to be more noise lower down. It is also significant that the search engine used was Google, which uses the *page rank* authority measure ([1]) to enhance its ranking. Therefore, we speculate that use of an authority measure

can be used to further improve the quality of our DPF. This will be investigated in future work.

8. REFERENCES

[1] Brin, S., Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine, in Proceedings of the 7th International WWW Conference, April 1998, Brisbane, Australia.

[2] Hearst, M.A. Automated Discovery of WordNet Relations, in WordNet: an electronic lexical database, C. Fellbaum (ed.), MIT Press, 131-151, 1998.

[3] Joho, H., Sanderson, M. Retrieving Descriptive Phrases from Large Amounts of Free Text, in Proceedings of the 9th ACM CIKM Conference, November 2000, McLean, VA, 180-186.

[4] Radev, D.R., McKeown, K.R. Building a Generation Knowledge Source using Internet-Accessible Newswire, in Proceedings of the 5th ANLP Conference, March 1997, Washington, D.C., 221-228.

[5] Srihari, R & Li, W. A Question Answering System Supported by Information Extraction, in Proceedings of the 8th ANLP Conference, April-May 2000, Seattle, Washington.