Answering Definition Questions via Temporally-Anchored Text Snippets

Marius Paşca
Google Inc.
1600 Amphitheatre Parkway
Mountain View, California 94043
mars@google.com

Abstract

A lightweight extraction method derives text snippets associated to dates from the Web. The snippets are organized dynamically into answers to definition questions. Experiments on standard test question sets show that temporally-anchored text snippets allow for efficiently answering definition questions at accuracy levels comparable to the best systems, without any need for complex lexical resources, or specialized processing modules dedicated to finding definitions.

1 Introduction

In the field of automated question answering (QA), a variety of information sources and multiple extraction techniques can all contribute to producing relevant answers in response to natural-language questions submitted by users. Yet the nature of the information source which is mined for answers, together with the scope of the questions, have the most significant impact on the overall architecture of a QA system. When compared to the average queries submitted in a decentralized information seeking environment such as Web search, fact-seeking questions tend to specify better the nature of the information being sought by the user, whether it is the name of the longest river in some country, or the name of the general who defeated the Spanish Armada. In order to understand the structure and the linguistic clues encoded in natural-language questions, many QA systems employ sophisticated techniques, thus deriving useful information such as terms, relations among terms, the type of the expected answers (e.g., cities vs. countries vs. presidential candidates), and other semantic constraints (e.g., the elections from 1978 rather than any other year).

One class of questions whose characteristics place them closer to exploratory queries, rather than standard fact-seeking questions, are definition questions. Seeking information about an entity or a concept, questions such as *"Who is Caetano Veloso?"* offer little guidance as to what particular techniques could be used in order to return relevant information from a large text collection. In fact, the same user may choose to submit a definition question or a simpler exploratory query (*Caetano Veloso*), and still look for text snippets capturing relevant properties of the question concept. Various studies (Chen et al., 2006; Han et al., 2006) illustrate the challenges introduced by definition questions. As such questions have a less irregular form than other open-domain questions, recognizing their type is relatively easier (Hildebrandt et al., 2004). Conversely, the identification of relevant documents and the extraction of answers to definition questions are more laborious, and the impact on the architecture of QA systems is quite significant. Indeed, separate, dedicated modules, or even end-to-end systems are specifically built for answering definition questions (Klavans and Mureşan, 2001; Hildebrandt et al., 2004; Greenwood and Saggion, 2004). The importance of definition questions among other question categories is confirmed by their inclusion among the evaluation queries from the QA track of TREC evaluations (Voorhees and Tice, 2000).

This paper investigates the impact of temporally-anchored text snippets derived from the Web, in answering definition questions and, more generally, exploratory queries. Section 2 describes a lightweight mechanism for extracting text snippets and associated dates from sentences in Web documents. Section 3 assesses the coverage of the extracted snippets. As shown in Section 4, relevant events, in which the question concept was involved, can be captured by matching the queries on the text snippets, and organizing the snippets around the associated dates. Section 5 describes discusses the role of the extracted text snippets in answering two sets of definition questions.

2 Temporally Anchored Text Snippets

All experiments rely on the unstructured text in approximately one billion documents in English from a 2003 Web repository snapshot of the Google search engine. Pre-processing of the documents consists in HTML tag removal, simplified sentence boundary detection, tokenization and part-of-speech tagging with the TnT tagger (Brants, 2000). No other tools or lexical resources are employed.

A sequence of sentence tokens represents a potential date if it consists of: single year (four-digit numbers, e.g., *1929*); or simple decade (e.g., *1930s*); or month name and year (e.g., *January 1929*); or month name, day number and year (e.g., *January 15, 1929*). Dates occurring in text in any other format are ignored. To avoid spurious matches, such as *1929 people*, potential dates are discarded if they are immediately followed by a noun or noun modifier, or immediately preceded by a noun.

To convert document sentences into a few text snippets associated with dates, the overall structure of sentences is roughly approximated. Deep text analysis may be desirable but simply not feasible on the Web. As a lightweight alternative, the proposed extraction method approximates the occurrence and boundaries of text snippets through the following set of lexico-syntactic patterns:

(P_1): $\langle Date$ [,|-|(|nil] [when] $Snippet$ [,|-|)|.]$\rangle$
(P_2): $\langle$[StartSent] [In|On] $Date$ [,|-|(|nil] $Snippet$ [,|-|)|.]$\rangle$
(P_3): $\langle$[StartSent] $Snippet$ [in|on] $Date$ [EndSent]$\rangle$
(P_4): $\langle$[Verb] [OptionalAdverb] [in|on] $Date\rangle$

The first extraction pattern, P_1, targets sentences with adverbial relative clauses introduced by wh-adverbs and preceded by a date, e.g.:

"By [$_{Date}$ 1910], when [$_{Snippet}$ Korea was annexed to Japan], the Korean population in America had grown to 5,008".

Comparatively, P_2 and P_3 match sentences that start or end in a simple adverbial phrase containing a date. In the case of P_4, the occurrence of relevant dates within sentences is approximated by verbs followed by a simple adverbial phrase containing a date. P_4 marks the entire sentence as a potential nugget because it lacks the punctuation clues in the other three patterns.

The patterns must satisfy additional constraints in order to match a sentence. These constraints constitute heuristics to avoid, rather than solve, complex linguistic phenomena. Thus, a nugget is always discarded if it does not contain a verb, or contains any pronoun. Furthermore, the snippets in P_2 and P_3 must start with, and the nugget in P_4 must contain a noun phrase, which in turn is approximated by the occurrence of a noun, adjective or determiner. The combination of patterns and constraints is by no means definitive or error-free. It is a practical solution to achieve graceful degradation on large amounts of data, reduce the extraction errors, and improve the usefulness of the extracted snippets. As such, it emphasizes robustness at Web scale, without taking advantage of existing specification languages for representing events and temporal expressions occurring in text (Pustejovsky et al., 2003), and forgoing the potential benefits of more complex methods that extract temporal relations from relatively clean text collections (Mani et al., 2006).

3 Coverage of Text Snippets

A concept such as a particular actor, country or organization usually occurs within more than one of the extracted text snippets. In fact, the set of text snippets containing the concept, together with the associated dates, often represents an extract-based, simple temporal summary of the events in which the concept has been involved. Starting from this observation, a task-based evaluation of the coverage of the extracted text snippets consists in verifying to what extent they capture the condensed history of several countries. Since any country must have been involved in some historical timeline of events, a reference timeline is readily available in an exter-

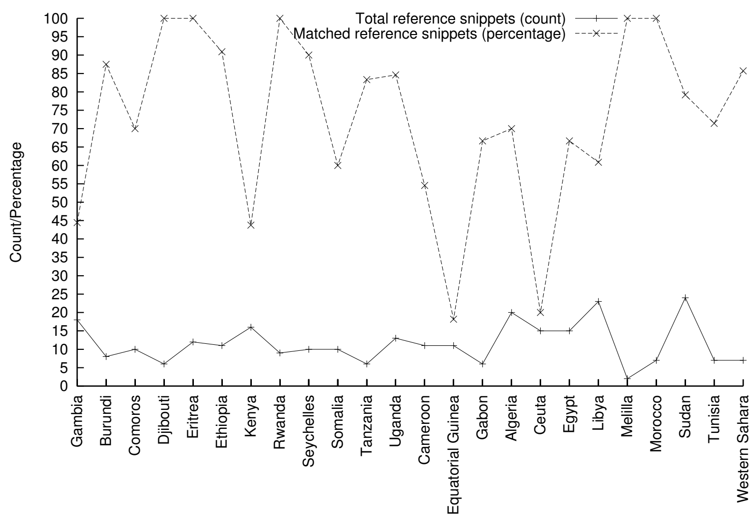

Figure 1: Percentage of reference snippets with corresponding extracted snippets

nal resource, e.g., encyclopedia, as an excerpt covering a condensed history of the country. The reference timeline is compared against the text snippets containing a country such as *Ethiopia*. To this effect, the text snippets containing a given country as a phrase are retained, ordered in increasing order of their associated dates, and evaluated against the reference timeline.

Both the test set of countries and the gold standard are collected from Wikipedia (Remy, 2002). The test set comprises countries from Africa. Since African countries have fewer extracted snippets than other countries, the evaluation results provide more useful, lower bounds rather than average or best-case. Due to limited human resources available for this evaluation, the test countries are a subset of the African countries in Wikipedia, selected in the order in which they are listed on the site. They cover all Eastern, Central and Northern Africa. The *Central African Republic*, the *Republic of the Congo*, and *Sao Tome and Principe* are discarded and *Gambia* added, leading to a test set of 24 country names. The source of the reference timelines is the condensed history article that is part of the main description page of each country in Wikipedia.

The evaluation procedure is concerned only with recall, but is still highly subjective. It requires the manual division of the reference text into dated events. In addition, the assessor must decide which details surrounding an event are significant, and must be matched into the extracted snippets in order to get any credit. The actual evaluation consists in matching each dated event from the reference timeline into the extracted timeline. During matching, the extracted snippets are analyzed by hand to decide which snippets, if any, capture the reference event, significant details around it, and the time stamp.

On average, 1173 text snippets are returned per country name, with a median of 733 snippets. Figure 1 summarizes the comparison of reference snippets and extracted snippets. The continuous line corresponds to the total number of reference snippets that were manually identified in the reference timeline; *Melilla* has the smallest such number (2), whereas *Sudan* has the largest (24). The dotted line in Figure 1 represents the percentage of reference snippets that have at least one match into the extracted snippets, thus evaluating recall. An average of 72% of the reference snippets have such matches. For 5 queries, there are matches for all reference snippets. The worst case occurs for *Equatorial Guinea*, for which only two out of the 11 reference snippets can be matched. Based on the results, we conclude that the text snippets and the associated dates provide a good coverage in the case of information about countries. The snippets can be retrieved as answers to questions asking about dates (*When*, *What year*) as described in (Paşca, 2007), or as answers to definition questions as discussed below.

4 Answering Definition Questions

Input definition questions are uniformly handled as Boolean queries, after the removal of stop words as well as question-specific terms (*Who* etc.). Thus, questions such as *"Who is Caetano Veloso?"* and *"Who won the Nobel Peace Prize?"* are consistently converted into conjunctive queries corresponding to *Caetano Veloso* and *won Nobel Peace Prize* respectively. The score assigned to a matching text snippet is higher, if the snippet occurs in a larger number of documents. Similarly, the score is higher if the snippet contains fewer non-stop terms in addition to the question term matches, or the average distance in the snippet between pairs of query term matches is lower. A side effect of the latter heuristic is to boost the snippets in which the query terms occur as a phrase, rather than as scattered term matches.

When they are associated to a common date, retrieved snippets transfer their relevance score onto the date, in the form of the sum of the individual snippet scores. The dates are ranked in decreasing order of their relevance scores, and those with the highest scores are returned as responses to the question, together with the top associated snippets. Within a set of text snippets associated to a date, the snippets are also ranked relatively to one another, such that each returned date is accompanied by its top supporting snippets. The ranking within a set of snippets associated to a date is a two-pass procedure. First, the snippets are scanned to count the number of occurrences of non-stop unigrams within the entire set. Second, a snippet is weighted with respect to others based on how many of the unigrams it contains, and the individual scores of those unigrams.

In the output, the snippets act as useful, implicit text-based justifications of why the dates may be relevant or not. As such, they implement a practical method of fusing together bits (snippets) of information collected from unrelated documents. In some cases, the snippets show why a returned result (date) is relevant. For example, *1990* is relevant to the query *Germany unified* because *"East and West Germany were unified"* according to the top snippet. In other cases, the text snippets quickly reveal why the result is related to the query even though it may not match the original user's intent. For instance, a user may ask the question *"When was the Taj Mahal built?"* with the well-known monument in mind, in which case the irrelevance of the date *1903* is self-explanatory based on one of its supporting snippets, *"the lavish Taj Mahal Hotel was built"*.

5 Evaluation

The answers returned by the system are ranked in decreasing order of their scores. By convention, an answer to a definition question comprises a returned date, plus the top matching text snippets that provide support for that date. Ideally, a snippet should only contain the desired answer and nothing else. In practice, a snippet is deemed correct if it contains the ideal answer, although it may contain some other extraneous information.

5.1 Objective Evaluation

A thorough evaluation of answers to definition questions would be complex, prone to subjective assessments, and would involve significant human labor (Voorhees, 2003). Therefore, the quality of the text snippets in the context of definition questions is tested on a set, DefQa1, containing the 23 *"Who is/was [ProperName]?"* questions from the TREC QA track from 1999 through 2002. In this case, each returned answer consists of a date and the first supporting text snippet.

Table 1 contains a sample of the test questions. The right column shows actual text snippets retrieved for the definition questions, together with the associated date and the rank of that date within the output. In an objective evaluation strictly based on the answer keys of the gold standard, the MRR score over the DefQa1 set is 0.596. The score is quite high, given that the answer keys prefer the genus of the question concept, rather than other types of information. For instance, the answer keys for the TREC questions Q222:*"Who is Anubis?"* and Q253:*"Who is William Wordsworth?"* mark *poet* and *"Egyptian god"* as correct answers respectively, thus emphasizing the genus of the question concepts *Anubis* and *William Wordsworth*. This explains the strong reliance in previous work on hand-written patterns and dictionary-based techniques for detecting text fragments encoding the genus and differentia of the question concept (Lin, 2002; Xu et al., 2004).

Question	(Rank) Relevant Date: Associated Fact
Q218: Who was Whitcomb Judson?	(1) 1893: First patented in 1893 by Whitcomb Judson, the Clasp Locker was notoriously unreliable and expensive
	(2) 1891: the zipper was invented by Whitcomb Judson
Q239: Who is Barbara Jordan?	(1) February 21 1936: Barbara Jordan was born in Houston, Texas
	(2) January 17 1996: Barbara Jordan died in Austin, Texas, at the age of 59
	(4) 1973: Barbara Jordan was diagnosed with multiple sclerosis and was confined to a wheelchair
	(5) 1976: Barbara Jordan became the first African-American Woman to deliver a keynote address at a political convention
	(7) 1966: Barbara Jordan became the first black representative since 1883 to win an election to the Texas legislature
	(8) 1972: Barbara Jordan was elected to the US Congress
Q253: Who is William Wordsworth?	(1) 1770: William Wordsworth was born in 1770 in the town of Cockermouth, England
	(2) April 7 1770: William Wordsworth was born
	(4) 1798: Romanticism officially began, when William Wordsworth and Samuel Taylor Coleridge anonymously published Lyrical Ballads
	(5) 1802: William Wordsworth married Mary Hutchinson at Brompton church
	(7) 1795: Coleridge met the poet William Wordsworth
	(8) April 23 1850: William Wordsworth died
	(11) 1843: William Wordsworth (1770-1850) was made Poet Laureate of Britain
Q346: Who is Langston Hughes?	(1) 1902: Langston Hughes was born in Joplin, Missouri
	(2) May 22 1967: Langston Hughes died of cancer
	(5) 1994: The Collected Poems of Langston Hughes was published
Q351: Who is Charles Lindbergh?	(1) 1927: aviation hero Charles Lindbergh was honored with a ticker-tape parade in New York City
	(2) 1932: Charles Lindbergh's infant son was kidnapped and murdered
	(3) February 4 1902: Charles Lindbergh was born in Detroit
	(5) August 26 1974: Charles Lindbergh died
	(7) May 21 1927: Charles Lindbergh landed in Paris
	(8) May 20 1927: Charles Lindbergh took off from Long Island
	(9) May 1927: an airmail pilot named Charles Lindbergh made the first solo flight across the Atlantic Ocean
Q419: Who was Jane Goodall?	(1) 1977: Goodall founded the Jane Goodall Institute for Wildlife Research
	(2) April 3 1934: Jane Goodall was born in London, England
	(3) 1960: Dr Jane Goodall began studying chimpanzees in east Africa
	(8) 1985: Jane Goodall 's twenty-five years of anthropological and conservation research was published

Table 1: Temporally-anchored text snippets returned as answers to definition questions

5.2 Subjective Evaluation

Beyond the snippets that happen to contain the genus of the question concept, the output constitutes supplemental results to what other definition QA systems may offer. The intuition is that prominent facts associated with the question concept provide useful, if not direct answers to the corresponding definition question, with the twist of presenting them together with the associated date. For instance, the first answer to Q239:*"Who is Barbara Jordan?"* reveals her date of birth and is associated with the first retrieved date, *February 21 1936*. In the objective evaluation, this answer is marked as incorrect. However, some users may find this snippet useful, although they may still prefer the seventh or eighth text snippets from Table 1 as primary answers, as they mention *Barbara Jordan*'s election to a state legislature in *1966*, and to the Congress in *1972*. As an alternative evaluation, the top five matching snippets for each of the top ten dates are inspected manually, and answers such as the birth year of a person are subjectively marked as correct. Overall, 59.1% of the snippets returned for the DefQa1 questions are deemed correct, which shows that the answers capture useful properties of the question concepts.

5.3 Alternative Objective Evaluation

A separate objective evaluation was conducted on a set, DefQa2, containing the 24 definition questions asking for information about various people, from the TREC QA track from 2004. Although correctness assessments are still subjective, they benefit from a more rigorous evaluation procedure. For each question, the gold standard consists of sets of responses classified according to their importance into two classes, namely *vital* nuggets, containing information that the assessors feel must be returned for the overall output to be good, and *non-vital*, containing information that is acceptable in the output but not necessary.

Following the official 2004 evaluation procedure (Voorhees, 2004), a returned text snippet is considered vital, non-vital, or incorrect based on whether it conceptually matches a vital, non-vital answer, or none of the answers specified in the gold standard for that question. The overall recall is the average of individual recall values per question, which are computed as the number of returned vital answers, divided by the number of vital answers from the gold standard for a given question. In this case, a returned answer is formed by a date and its top three associated text snippets. If a vital answer from the gold standard matches any of the three snippets of a returned answer, then the returned answer is vital.

The overall recall value over DefQa2 is 0.46. The corresponding F-measure, which gives three times more importance to recall than to precision as specified in the official evaluation procedure, is 0.39. The score measures favorably against the top three F-measure scores of 0.46, 0.40, and 0.37 reported in the official 2004 evaluation (Voorhees, 2004). The two better scores were obtained by systems that rely extensively on human-generated knowledge from resources such as WordNet (Zhang et al., 2005) and specific Web glossaries (Cui et al., 2007). In comparison, the text snippets retrieved in this paper provide relevant answers to definition questions with the added benefit of providing a temporal anchor for each answer, and without using any complex linguistic resources and tools.

The scores per question vary widely, with the retrieved snippets containing none of the vital answers for six questions, all vital answers for other six, and some fraction of the vital answers for the remaining questions. For example, one of the retrieved text snippets is *"US Air Force Colonel Eileen Marie Collins was the first woman to command a space shuttle mission"*. The snippet is classified as vital for the question about *Eileen Marie Collins*, since it conceptually matches a vital answer from the gold standard, namely *"first woman space shuttle commander"*. Again, even though the standard evaluation does not require a temporal anchor for an answer to be correct, we feel that the dates associated to the retrieved snippets provide very useful, additional, condensed information. In the case of *Eileen Marie Collins*, the above-mentioned vital answer is accompanied by the date *1999*, when the mission took place.

6 Related Work

Previous approaches to answering definition questions from large text collections can be classified according to the kind of techniques for the extraction of answers. A significant body of work is oriented towards mining descriptive phrases or sentences, as opposed to other types of semantic information, for the given question concepts. To this effect, the use of hand-written lexico-syntactic patterns and regular expressions, targeting the genus and possibly the differentia of the question concept, is widespread, whether employed for mining definitions in English (Liu et al., 2003; Hildebrandt et al., 2004) or other languages such as Japanese (Fujii and Ishikawa, 2004), from local text collections (Xu et al., 2004) or from the Web (Blair-Goldensohn et al., 2004; Androutsopoulos and Galanis, 2005). Comparatively, the small set of patterns used here targets text snippets that are temporally-anchored. Therefore the text snippets provide answers to definition answers without actually employing any specialized module for seeking specific information such as the genus of the question concept.

Several studies propose unsupervised extraction methods as an alternative to using hand-written patterns for definition questions (Androutsopoulos and Galanis, 2005; Cui et al., 2007). Previous work often relies on external resources as an important or even essential guide towards the desired output. Such resources include WordNet (Prager et al., 2001) for finding the genus of the question concept; large dictionaries such as Merriam Webster, for ready-to-use definitions (Xu et al., 2004; Hildebrandt et al., 2004); and encyclopedias, for collecting words that are likely to occur in potential definitions (Fujii and Ishikawa, 2004; Xu et al., 2004). In comparison, the experiments reported in this paper do not require any external lexical resource.

7 Conclusion

Without specifically targeting definitions, temporally-anchored text snippets extracted from the Web provide very useful answers to definition questions, as measured on standard test question sets. Since the snippets tend to capture important events involving the question concepts, rather than phrases that describe the question concept, they can be employed as either standalone answers, or supplemental results in conjunction with answers extracted with other techniques.

References

I. Androutsopoulos and D. Galanis. 2005. A practically unsupervised learning method to identify single-snippet answers to definition questions on the Web. In *Proceedings of the Human Language Technology Conference (HLT-EMNLP-05)*, pages 323–330, Vancouver, Canada.

S. Blair-Goldensohn, K. McKeown, and A. Schlaikjer, 2004. *New Directions in Question Answering*, chapter Answering Definitional Questions: a Hybrid Approach, pages 47–58. MIT Press, Cambridge, Massachusetts.

T. Brants. 2000. TnT - a statistical part of speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP-00)*, pages 224–231, Seattle, Washington.

Y. Chen, M. Zhou, and S. Wang. 2006. Reranking answers for definitional QA using language modeling. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)*, pages 1081–1088, Sydney, Australia.

H. Cui, M. Kan, and T. Chua. 2007. Soft pattern matching models for definitional question answering. *ACM Transactions on Information Systems*, 25(2).

A. Fujii and T. Ishikawa. 2004. Summarizing encyclopedic term descriptions on the Web. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 645–651, Geneva, Switzerland.

M. Greenwood and H. Saggion. 2004. A pattern based approach to answering factoid, list and definition questions. In *Proceedings of the 7th Content-Based Multimedia Information Access Conference (RIAO-04)*, pages 232–243, Avignon, France.

K. Han, Y. Song, and H. Rim. 2006. Probabilistic model for definitional question answering. In *Proceedings of the 29th ACM Conference on Research and Development in Information Retrieval (SIGIR-06)*, pages 212–219, Seattle, Washington.

W. Hildebrandt, B. Katz, and J. Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of the 2004 Human Language Technology Conference (HLT-NAACL-04)*, pages 49–56, Boston, Massachusetts.

J. Klavans and Smaranda Mureşan. 2001. Evaluation of Definder: A system to mine definitions from consumer-oriented medical text. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL-01)*, pages 201–203, Roanoke, Virginia.

C.Y. Lin. 2002. The effectiveness of dictionary and web-based answer reranking. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)*, pages 1–7, Taipei, Taiwan.

B. Liu, C. Chin, and H.T. Ng. 2003. Mining topic-specific concepts and definitions on the Web. In *Proceedings of the 12th International World Wide Web Conference (WWW-03)*, pages 251–260, Budapest, Hungary.

I. Mani, M. Verhagen, B. Wellner, C. Lee, and J. Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)*, pages 753–760, Sydney, Australia.

M. Paşca. 2007. Lightweight Web-based fact repositories for textual question answering. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM-07)*, Lisboa, Portugal.

J. Prager, D. Radev, and K. Czuba. 2001. Answering what-is questions by virtual annotation. In *Proceedings of the 1st Human Language Technology Conference (HLT-01)*, pages 1–5, San Diego, California.

J. Pustejovsky, J. Castano, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, and G. Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, Tilburg, Netherlands.

M. Remy. 2002. Wikipedia: The free encyclopedia. *Online Information Review*, 26(6):434.

E.M. Voorhees and D.M. Tice. 2000. Building a question-answering test collection. In *Proceedings of the 23rd International Conference on Research and Development in Information Retrieval (SIGIR-00)*, pages 200–207, Athens, Greece.

E. Voorhees. 2003. Evaluating answers to definition questions. In *Proceedings of the 2003 Human Language Technology Conference (HLT-NAACL-03)*, pages 109–111, Edmonton, Canada.

E.M. Voorhees. 2004. Overview of the TREC-2004 Question Answering track. In *Proceedings of the 13th Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland. NIST.

J. Xu, R. Weischedel, and A. Licuanan. 2004. Evaluation of an extraction-based approach to answering definitional questions. In *Proceedings of the 27th ACM Conference on Research and Development in Information Retrieval (SIGIR-04)*, pages 418–424, Sheffield, United Kingdom.

Z. Zhang, Y. Zhou, X. Huang, and L. Wu. 2005. Answering definition questions using Web knowledge bases. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 498–506, Jeju Island, Korea.