A No-Frills Architecture for Lightweight Answer Retrieval

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

ABSTRACT

In a new model for answer retrieval, document collections are distilled offline into large repositories of facts. Each fact constitutes a potential direct answer to questions seeking a particular kind of entity or relation, such as questions asking about the date of particular events. Question answering becomes equivalent to online fact retrieval, which greatly simplifies the de-facto system architecture.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*; H.3.1 [**Information Storage and Retrieval**]: Context Analysis and Indexing—*linguistic processing*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*

General Terms

Algorithms, Experimentation

Keywords

Web information retrieval, lightweight text analysis, fact repositories, question answering

1. INTRODUCTION

Open-domain question answering (QA) systems generally have modules for question processing, document retrieval, and answer detection and ranking. In particular, document retrieval identifies a relatively small subset of documents from the underlying text collection, which are deemed to be the most relevant to the given question [3]. The de-facto paradigm in QA can be described as *"retrieval of potentially relevant documents"*, then *"extraction of potential answers"*, then *"return of top answers"*, all of which are performed online for each question. Depending on the collection size, 95% or more of the documents in the collection (much more in the case of the Web) are left out of the selected subset for any given query, and thus become invisible to subsequent processing stages for answer mining.

Without neglecting the unquestionable practical advantages of employing document retrieval in QA, this paper explores an alternative approach for reducing the amount of text being searched for answers for each submitted question.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

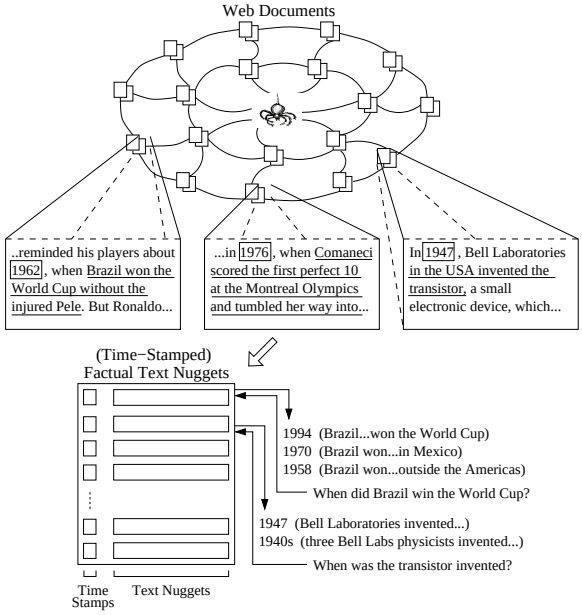

Figure 1: Web-derived fact repositories for QA

The approach can be described as a sequence of *"offline extraction of potentially relevant facts from all documents"* (in one pass, similarly to document indexing), followed by *"online retrieval of relevant facts"* and *"return of top answers associated with prominent facts"*.

2. APPROACH

2.1 Overview

As illustrated in Figure 1, pairs of a sentence fragment (e.g., *"Bell Labs invented the transistor"*) and the date (e.g., *1947*) of the corresponding event are extracted from unstructured text and organized offline into fact repositories rather than document indices. Each fact is equivalent to a pseudo-document that is stored, indexed and retrieved with standard information retrieval techniques. Whenever a question (e.g., *"When was the transistor invented?"*) targets the same type of information as that encoded in the fact repositories, the entities (e.g., *1947*) associated with the matching sentence fragments are returned as candidate answers.

In addition to offering direct answers in the form of entities of the desired type (dates), the proposed fact repositories have several advantages over the traditional QA dataflow.

First, they greatly simplify the system architecture. An entire sequence of previously-proposed processing stages for QA (namely, document retrieval, passage retrieval, answer detection and answer ranking) is replaced with a single stage for fact retrieval. Second, the matching text fragments provide at-a-glance justifications as to why the associated entities have been returned. Third, the results merge textual information and build upon evidence originating from scattered documents.

2.2 Offline Fact Extraction

For robustness and scalability, the extraction relies only on lightweight tools and minimal resources. Initially, the Web documents are processed to filter out HTML tags. The resulting text-only data is tokenized, split into sentences and part-of-speech tagged using the TnT tagger [1]. A sequence of sentence tokens represents a potential date if it has one of the following formats: single year (four-digit numbers); or simple decade; or month name and year; or month name, day number and year. Dates occurring in text in any other format are ignored. The occurrence and boundaries of factual text fragments are approximated through a set of lexico-syntactic patterns targeting adverbial clauses and phrases, e.g.: $\langle Date$ [,|-|(|nil] [when] $Fragment$ [,|-|)|.]$\rangle$

2.3 Online Fact Retrieval

The fact retrieval stage is a four-step inverted fragment search aiming to: 1) *match* the query against the actual texts of the facts; 2) *score* the matching fragments individually relatively to the query; 3) *aggregate* the best-matching fragments associated with the same date; within each group, *combine* (i.e., sum) the scores of the matching fragments into an aggregated score assigned to the common date; and 4) *select* the dates with the highest scores.

The initial step includes all query terms by default, thus implementing a conjunctive Boolean search. This sometimes results in empty sets of answers, due to lexical mismatches between the question words, on one side, and words in individual text fragments, on the other. Two extensions are added to alleviate the mismatch. First, minimal morphological processing of questions expand the first base-form verb in the query with its Past Tense, based on standard morphological inflection rules refined with 176 English irregular verbs. Second, the fact retrieval stage is modified to issue a series of progressively more general Boolean queries, rather than a single Boolean query. More general queries are generated through iterative removal of question keywords applied as long as the returned answer set is empty, until at least a matching factual fragment is returned.

3. OPEN-DOMAIN EVALUATION

3.1 Experimental Setting

The metrics measured in the evaluation are the following: 1) MRR: the standard mean reciprocal rank score (or 0 if all answers are incorrect); 2) Q_{All}: total number of questions in the test set; 3) Q_{Ri}: number of questions with the first correct answer at rank i; 4) Q_{Ri-j}: number of questions with the first correct answer at some rank from i through j.

To assess the impact of the two extensions meant to reduce the lexical mismatch, they are switched on and off in four system configuration settings, namely KnMn (no extensions enabled), KyMn (only iterative removal of keywords enabled), KnMy (only morphological processing of base form verbs enabled), and KyMy (both extensions enabled).

Table 1: Accuracy of answers retrieved from the temporal fact repository (Q=test question set)

Q_{All}	Config	Q_{R1}	Q_{R2}	Q_{R3}	Q_{R1-5}	MRR
199	KnMn	73	6	1	82	0.389
	KnMy	95	8	1	105	0.503
	KyMn	104	15	3	127	0.574
	KyMy	112	15	2	131	0.608

The document collection contains half a billion documents in English from a 2003 Web repository snapshot maintained by Google. The test question set collects all *When* and *What year* of the 1893 main-task queries, from the Question Answering track of past editions of TREC from 1999 through 2002. Since 8 of the queries in the original set were discarded by the TREC organizers, the set consists of 199 temporal, but otherwise open-domain questions. The organizers also provide a gold standard of answer keys, as well as an automated evaluation script, in addition to the query set.

3.2 Results

Table 1 shows the accuracy of the returned answers. The impact of the system configuration is progressive. In the KnMy configuration, the first returned answer is correct for 22 additional questions when compared to KnMn. Similarly, KyMn correctly answers 31 more questions at rank 1 than KnMn does. The KyMy configuration gives the best results: 112 (56%) questions have a correct answer at rank 1, and 131 (65%) questions have a correct answer at ranks 1 through 5. The resulting MRR score is 0.608.

Another data-driven approach to QA is evaluated in [2] on the same set of temporal questions. The answers are extracted freely from the Web through an external search engine, rather than confined to a more limited, local text collection. The authors report an MRR score of 0.447, and indicate that the score is consistently above the sixth highest score of the systems participating in the TREC QA track. Therefore, our MRR score of 0.608 on temporal questions is comparable to the best performing systems.

4. CONCLUSION

Through a simplified architecture that consists of offline fact extraction and online fact retrieval, factual text fragments provide answers directly to temporal questions. The corresponding MRR score on a set of TREC questions is 0.608, which compares favorably to the performance on the same test question set of the top-performing systems.

5. REFERENCES

[1] T. Brants. 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, 2000.

[2] L. Lita and J. Carbonell. Instance-based question answering: A data driven approach. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, pages 396–403, Barcelona, Spain, 2004.

[3] S. Tellex, B. Katz, J. Lin, A. Fernandez, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th ACM Conference on Research and Development in Information Retrieval (SIGIR-03)*, pages 41–47, Toronto, Canada, 2003.