Processing Definition Questions in an Open-Domain Question Answering System

Marius Paşca
Language Computer Corporation
Dallas, Texas
marius@languagecomputer.com

Abstract

This paper presents a hybrid method for finding answers to Definition questions within large text collections. Because candidate answers to Definition questions do not generally fall in clearly defined semantic categories, answer discovery is guided by a combination of pattern matching and WordNet-based question expansion. The method is incorporated in a large open-domain question answering system and validated by extracting answers to 188 questions from a standard 3-Gigabyte text collection and from Web documents.

Background

The continuous growth and diversification of online text information represents a constant challenge, simultaneously affecting both the designers and the users of information processing systems. Information covering almost any conceivable topic becomes available every day, from a variety of sources including live news-wire services like Reuters and Associated Press, or online encyclopedia like Britannica Online. Assisting the users in finding relevant answers from large text collections is a challenging task, whether the users are interested in the estimated expenses associated with the best MBA programs in California, or the distance between the Earth and Jupiter, or any other factual information.

Situated at the frontier of natural language processing and modern information retrieval, open-domain question answering (QA) is an appealing alternative to retrieval of full-length documents. Users of QA systems formulate their information needs in the form of natural-language questions, thus eliminating any artificial constraints sometimes imposed by a particular input syntax (e.g., Boolean operators). The QA system returns brief answer strings extracted from the text collection, thus taking advantage of the fact that answers to specific questions are often concentrated in small fragments of text documents. The output of a QA system is better adapted to modern information needs. It is the system, and not the user, who is responsible for analyzing the content of full-length documents and identifying small, relevant text fragments. With the introduction of the QA track in the Text REtrieval Conference (TREC), research in open-domain QA gained new momentum. The yearly-organized track provides a useful benchmark against which novel models and architectures can be tested, by extracting answers from a Gigabyte text collection in response to a test set of fact-seeking questions.

Definition Questions

This paper focuses on the problem of finding document snippets that answer a particular category of fact-seeking questions, namely *Definition* questions. Examples of such questions are *"What is autism?"*, *"What is a shaman?"* or *"What are triglycerides?"*. The choice of Definition questions versus other types of questions is motivated by the following factors:

- A considerable percentage of the questions actually submitted on the Web by search engines are Definition questions. Current search engine technology does little to support these questions because most search engines return links to full-length documents rather than brief document fragments that answer the user's question;
- The frequent occurrence of Definition questions in daily usage is confirmed by the composition of the question test sets in the QA track at TREC. The percentages of questions that are Definition questions grew from 1% in TREC-8 to 9% in TREC-9 and 25% in TREC-10;
- Most recent approaches to open-domain QA use named entity recognition as core technology for detecting candidate answers (Abney, Collins, & Singhal 2000; Srihari & Li 2000). The common observation is that the set of acceptable answers to a question like *"Which city has the largest French-speaking population?"*, can be restricted to *town* named entities such as *New York*, *Paris* or *Montreal*. Comparatively, Definition questions are more difficult to process because their candidate answers are unknown.

In addition to the factors listed above, adequate support for Definition questions also encourages the interactions between the user and the QA system, with

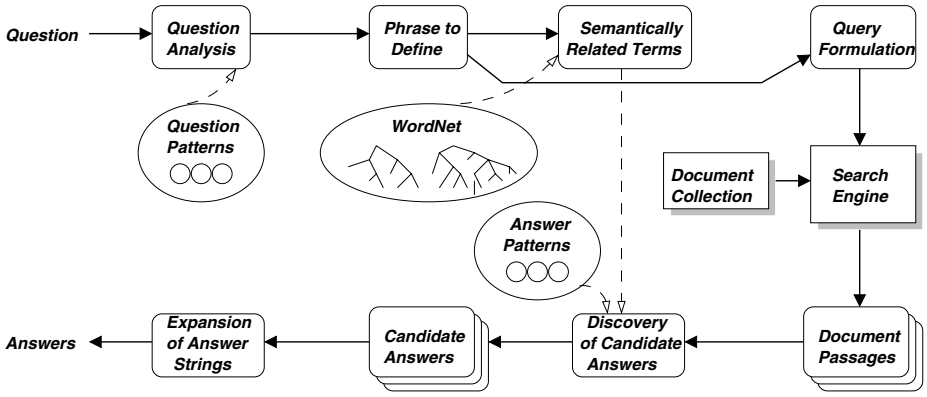

Figure 1: Answering Definition questions from large text collections

immediate advantages on overall effectiveness. Consider the question *"Where does Alcatel have locations in Texas?"* and its answer *"Alcatel, which has been in Texas for 10 years, employs more than 8,000 persons in the Dallas area, within three locations: its U.S. headquarters in Plano, and two other locations in the Telecom Corridor, including a research center"* extracted from `www.consulfrance-houston.org`. A user who is not familiar with the term *Telecom Corridor* may ask a follow-up question *"What is the Telecom Corridor?"*, leading to the extraction of the answer *"Dubbed the Texas Telecom Corridor, the Plano/Richardson area has become a nexus for telecommunications giants"* from `www.northstar.net`. In general, if the extracted answers are unclear or contain unfamiliar terms, the users may obtain further clarifications by submitting Definition questions. Furthermore, the discovery of definitions in a large text collection is beneficial to the construction of domain-specific thesauri, or the addition of new terms to an existing dictionary.

Discovery of Answer Anchors

Most difficulties in answering Definition questions arise from the lack of a clearly defined semantic category that circumscribes the candidate answers. Recent approaches tend to categorize general fact-seeking questions according to their expected answers. Thus the questions *"What was the 33rd US president?"* and *"Who is the mayor of New York City?"* have a common semantic category for their expected answers, namely person's names. Questions asking about named entities generally require access to knowledge of statistical (Berger *et al.* 2000; Ittycheriah *et al.* 2001) and/or linguistic (Cardie *et al.* 2000; Moldovan *et al.* 2000; Hovy *et al.* 2001) nature for reliable identification of the candidate answers. In contrast, candidate answers of Definition questions rarely fall in separate semantic categories. Therefore the availability of large-coverage named entity recognizers cannot be exploited to find candidate answers to questions like Q358: *"What is a meerkat?"* or Q386: *"What is anorexia nervosa?"*, both of which are part of the test set from the TREC-9 QA track. In addition, the small number of available keywords is a characteristic of Definition questions that further complicates the identification of candidate answers. Whereas for longer questions the concentration of the keywords in a small fragment of a document passage may be a clue that candidate answers are located nearby, there is no such concentration for one-keyword questions like *"What is a meerkat?"*.

The peculiar characteristics of Definition questions are taken into account in the system architecture shown in Figure 1. During question analysis, the part-of-speech tagged question is matched against a set of question patterns of the form:

What <be-verb> a <PhraseToDefine>?
→ *Match: What is anorexia nervosa?*
Who <be-verb> <HumanEntity>?
→ *Match: Who is Barbara Jordan?*

When the matching succeeds on any question pattern, the system marks the input as a Definition question. The phrase to define (e.g., *anorexia nervosa* for Q386: *"What is anorexia nervosa?"*) is transformed into a conjunctive Boolean query, e.g. (*anorexia* $\wedge$ *nervosa*). The query is passed to an external search engine providing transparent access to the underlying document collection from which the answers are to be extracted. The highest-ranking candidate answers discovered in the passages are expanded and returned to the users.

Because the precise discovery of candidate answers is hindered by the lack of a clear semantic category of the answers, a practical approach to answering Definition questions is to approximate the task of finding relevant answer strings by the simpler task of finding nearby words or *answer anchors* in the text documents. The answer anchors alone do not necessarily constitute meaningful answers. However, if the anchors are located in the immediate proximity of the actual answers, then 50-byte or 250-byte strings expanded around the answer anchors are likely to be relevant. For example,

Table 1: Identifying answers to Definition questions with pattern matching

Answer pattern	Phrase to define (QP)	Detected candidate answer (AP)
<AP> such as <QP>	What is autism?	*developmental disorders* such as autism
<AP> (also called <QP>)	What is bipolar disorder?	*manic - depressive illness* (also called bipolar disorder)
<QP> is an <AP>	What is caffeine?	caffeine is an *alkaloid*
<QP> , a <AP>	What is a caldera?	caldera , a *volcanic crater*

the detection of the answer anchor *such* for the question Q903: *"What is autism?"* enables the extraction of the 50-byte answer string *"developmental disorders such as autism"*, expanded around the answer anchor.

The first source for finding answer anchors in the retrieved passages is the matching of the passages on answer patterns. Table 1 illustrates answer strings extracted via answer patterns from the document collection used in the TREC QA track. Answer patterns are developed offline. They capture word sequences that are likely to introduce definitions or explanations of the phrase to define. The simplicity of the patterns provides robustness, a very desirable property for a QA system working on unrestricted text. About a dozen answer patterns allow for the extraction of correct answers to 58 Definition questions from TREC-9. Unfortunately, simplicity also leads to incorrect text snippets being sometimes returned due to spurious pattern matching. For instance, given the question Q236: *"Who is Coronado?"* the fifth answer, extracted from a Los Angeles Times document, is *"San Diego may find more sympathy among its southern neighbors , such as Coronado , with which it has a close rapport"*. The extraction of this answer is due to the false matching of the first pattern from Table 1 where QP is *Coronado* and AP is *neighbors*. Answer patterns are not sufficiently reliable to be used alone for answer finding.

In our quest for a more reliable method of identifying the candidate answers, i.e. the text snippets that answer the question, we opt for a hybrid method combining pattern matching and question expansion based on semantic resources (Paşca 2001a). The approach integrates information from a general-purpose knowledge source, namely the WordNet database (Miller 1995) into the search for candidate answers. The advantages of a stand-alone, structured knowledge base are combined with the discovery of unexpected information in small document fragments that would be otherwise submerged in a large text collection and therefore virtually inaccessible to users.

When the lexical concept corresponding to the phrase to define is found in WordNet, the question is automatically expanded to enhance the search for candidate answers. The expansion collects terms that are linked to the phrase to define through semantic relations. From the variety of semantic relations encoded in WordNet, including synonymy, hypernymy (generalization), hyponymy (specialization), holonymy (whole name), we choose to use the hypernymy relation and expand the phrase to define with its immediate hypernyms. The occurrence of a collected hypernyms in the retrieved document passages counts as an answer anchor, in addition to the anchors identified through pattern matching. Table 2 presents relevant answer strings extracted from the TREC-10 document collection, after the WordNet-based automatic expansion of Definition questions.

Table 2: Using WordNet to find answers to Definition questions

Phrase to define	Hypernym in WordNet	Detected candidate answer
What is a shaman?	{priest, non-Christian priest}	Mathews is the *priest* or shaman
What is a nematode?	{worm}	nematodes , tiny *worms* in soil .
What is anise?	{herb, herbaceous plant}	anise , rhubarb and other *herbs*

Table 3: Incorrect answer strings extracted due to imprecise discovery of answer anchors

Question	[*50-byte answer string*] shown in surrounding context
What is a carcinogen?	one of its drugs contained [*a carcinogen , a substance that encourages the*] growth of cancer
What is ozone depletion?	Because of the concern about ozone depletion, [*the nation 's of the world have agreed that*]
What is myopia?	night myopia [*, a condition that hinders focusing in the dark*]

Discovery of Candidate Answers

As shown in Table 3, the imprecise detection of answer anchors in the documents often results in answer strings that are slightly shifted and thus miss the actually correct text snippets. Ideally, the selected answer anchors should answer the user's question directly, i.e. without any string expansion around the anchors. Moreover, the most relevant answer strings should be assigned a high relevance score and ranked accordingly in the final answer list. As a preliminary step toward discovering answer candidates rather than answer anchors, three extensions are proposed here based on answer patterns:

parsing of the text snippets which match against an answer pattern; imposing a set of priorities on the answer patterns; and controlled WordNet-based expansion.

Answer sentence parsing The discovery of an answer anchor through pattern matching is followed by the parsing of the document sentence in which the anchor occurs (the same probabilistic parser is used for processing the submitted questions (Moldovan *et al.* 2000)). The answer anchor is adjusted to the closest base noun phrase, thus eliminating many "obvious" errors due to the erroneous selection of stop words (*a*, *the*, *in* etc.). Depending on the answer pattern, the noun phrase is searched in the document before (e.g., *...environmental catastrophes such as ozone depletion*) or after (e.g., *carcinogen , a substance that...*) the phrase to define (*ozone depletion* and *carcinogen* respectively). The answer strings are expanded in the same direction, starting from the adjusted answer anchor. The strings from Table 4 are returned after parsing the relevant text snippets; the answer anchors have been adjusted to the new locations indicated by the base noun phrases.

Table 4: Relevant answer strings extracted after adjustment of answer anchors

Question to define	[*50-byte answer string*] shown in surrounding context
What is a carcinogen?	one of its drugs contained a carcinogen , a [*substance that encourages the growth of cancer*]
What is ozone depletion?	[*long - term environmental catastrophes such as*] ozone depletion
What is myopia?	on [*the back of the eye , curing nearsightedness , or myopia .*]

Pattern priorities Each pattern is given a certain priority. Answer strings expanded around anchors that were obtained with a higher-priority pattern are assigned a higher relevance score than the strings obtained from lower-priority patterns. For example, the pattern *"AP , also called QP"* is assigned a higher priority than *"QP is a AP"*. Currently, the patterns are grouped into three classes of priority. The last two answer strings from Table 4 are returned within the first five answers if the pattern priorities are enabled.

Controlled concept expansion In the hybrid approach, some of the answer anchors are selected by expanding the question phrase based on WordNet. When the concepts selected for expansion are too general, like *substance*, *device* or *element*, the risk is to return answer strings that are inconclusive to be judged as relevant. Examples of such answer strings include *"a substance called platelet - derived growth factor , or PDGF"* for Q1033: *"What are platelets?"*, or *"as well as other elements , like phosphorus"* for Q1131: *"What is phosphorus?"*. To reduce the number of errors induced by automatic WordNet expansion, the expansion is not permitted for very general concepts.

Evaluation

The performance on Definition questions was tested by extracting answers from the text collection used in the QA track of the Text REtrieval Conference. Table 5 describes the structure of the collection.

Table 5: Description of the TREC QA track document collection

Source	No. Docs
Los Angeles Times	131,896 docs
Foreign Broadcast Info Service	130,471 docs
Financial Times	210,157 docs
Associated Press Newswire	242,918 docs
Wall Street Journal	173,252 docs
San Jose Mercury News	90,257 docs
Complete Collection (3 Gigabyte)	736,794 docs

Following the standardized scoring methodology proposed in the QA track, individual questions are assigned a score equal to the reciprocal answer rank. The answer rank is the rank of the first correct answer returned by the system (Voorhees & Tice 2000). Thus a question receives a score of 1, 0.5, 0.33, 0.25, 0.2, if the first correct answer is returned at rank 1, 2, 3, 4 and 5 respectively. By convention, the score for questions with no correct answer among the first five returned is 0. In other words, either the system returns a correct answer within the first five answer strings, or it does not receive any credit.

The experiments are performed on the subset of questions from the TREC-9 and TREC-10 QA track that are Definition questions. The test set contains a total of 188 questions (66 questions from TREC-9 and 122 from TREC-10). Table 6 illustrates 50-byte answers found in the TREC QA text collection. In contrast, Table 7 illustrates some of the answers that are extracted from Web documents rather than local text collections. In order to access Web documents, the architecture of the QA system is adapted such that questions are transformed into queries that are passed to external search engines. The QA system fetches locally the top Web documents returned by the search engines and then explores their contents in search for answers (Paşca 2001b). The selection of the Google search engine (Brin & Page 1998) to access Web documents is justified by the size of its index and the quality of search results.

The overall performance on 50-byte answers extracted from the 188 test questions is shown in Table 8. When the answers are extracted from the local TREC collection, the precision score is 0.566. The percentage of questions with a correct answer at any rank among the first five answers returned is 67.5%. The performance degradation incurred when answers are extracted from the Web rather than the local collection can be explained by inherent difficulties related to

Table 6: Answers extracted from the QA track text collection for TREC-9 questions

Question	50-byte answer	Source
Q241: What is a caldera?	caldera , a volcanic *crater* 19 miles long and 9	SJMN91-06111037
Q358: What is a meerkat?	meerkat , a type of *mongoose* , thrives in its	LA100789-0110
Q386: What is anorexia nervosa?	the most common *eating disorders* are anorexia	AP891106-0242

Table 7: Answers extracted from Web documents for TREC-10 questions

Question	50-byte answer	Source
Q926: What are invertebrates?	An invertebrate is an *animal without a backbone*	http://atschool.eduweb.co.uk/ sirrobhitch.suffolk/invert/inverteb.htm
Q982: What are xerophytes?	or xerophytes are *plants which use little water*	http://www.ag.usask.ca/cofa/departments/ hort/hortinfo/yards/appropri.html
Q1061: What is acetaminophen?	by a mild *analgesic* like acetaminophen . Thus ,	http://www.cirp.org/library/ pain/howard/

Web searching such as heterogeneity, network communication problems, and limits of current search engine technology. Nevertheless, with 45% of the questions being correctly answered from the Web, Table 8 shows that high-precision answer extraction is feasible on top of present search engine technology. Answers successfully extracted for other, non-TREC test questions are shown in Table 9.

Table 8: Precision score for 50-byte answers

Documents from which answers are extracted	Questions with a correct answer in top 5 returned	Precision score
TREC QA track document collection	127/188 (0.675)	0.566
Web documents	86/188 (0.457)	0.372

In a separate experiment, we used the same question set but disabled the expansion of the returned strings around answer anchors. In this case, the highest-ranking answer anchors constitute the output of the system. When the answer sentence parsing, pattern priorities and controlled concept expansion are enabled, the precision score for anchor-based answers extracted from the TREC document collection improves from 0.236 from 0.292. Additional experiments show that the performance degradation when extracting anchor-based answers rather than 50-byte answers is more pronounced for Definition questions than for questions asking about well-defined semantic categories (locations, persons, cities, quantities etc.). Therefore the systematic discovery of precise answers to Definition questions remains an open research issue.

Conclusions and Future Work

The work presented here can be extended in at least two directions. First, knowledge sources other than WordNet can be used to guide the search for the answers. Frequently, novel terms or domain-specific terms are not found in WordNet, in which case the search of the text collection relies solely on pattern matching. Online encyclopedia represent a valuable resource that should be explicitly exploited when answering Definition questions and other types of fact-seeking questions.

A second direction is the integration of heterogeneous sources of information as inputs to the QA system. If the QA system has access to WordNet, the Web, and a static collection of unrestricted texts, it should be able to select the information source to be searched, according to the question. For instance, it is possible but inefficient to search for the answer to a question like *"What is Valentine's Day?"* in a large document collection, especially if there are tens of thousands of text snippets in the collection containing both keywords *Valentine* and *Day* in close proximity to each other. Comparatively, an information source like WordNet provides immediately the desired answer (in WordNet *"Valentine's Day"* is a synonym with *"February 14"* and it is defined as *"a day for the exchange of tokens of affection"*) without any overhead or expense of valuable computational resources.

Acknowledgments

The author would like to thank Sanda Harabagiu, Dan Moldovan and Mihai Surdeanu for various comments on earlier versions of the paper. This work was supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program under contract number 2002-H-753600-000.

References

Abney, S.; Collins, M.; and Singhal, A. 2000. Answer extraction. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-2000)*, 296–301.

Table 9: Answers extracted from Web documents for various questions

Question	250-byte answer	Source
Who is Mohammad Atta?	A British woman who learned to fly with terror suspect Mohammad Atta is helping the FBI in its investigation into the terrorist attacks on the US	http://news.bbc.co.uk/hi/english/uk/england/newsid_1560000/1560305.stm
Who is Donald Rumsfeld?	this week (and much of the last one , too) the administration figure hogging the front page of national newspapers was Defense Secretary Donald Rumsfeld	http://www.time.com/time/pow/article/0,8599,109422,00.html
What is Al Qaeda?	The US government issued an indictment in November 1998 alleging that Osama bin Laden heads an international terrorist network called " Al Qaeda , " an Arabic word meaning " the base	http://www.pbs.org/wgbh/pages/frontline/shows/binladen/who/alqaeda.html
What is the Taliban?	The Taliban is the first faction laying claim to power in Afghanistan , that has targeted women for extreme repression and punished them brutally for infractions	http://www.phrusa.org/research/health_effects/exec.html

Berger, A.; Caruana, R.; Cohn, D.; Freitag, D.; and Mittal, V. 2000. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd International Conference on Research and Development in Information Retrieval (SIGIR-2000)*, 192–199.

Brin, S., and Page, L. 1998. The anatomy of a large scale hypertextual web search engine. In *7th International World Wide Web Conference*.

Cardie, C.; Ng, V.; Pierce, D.; and Buckley, C. 2000. Examining the role of statistical and linguistic knowledge sources in a general-knowledge question-answering system. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-2000)*, 180–187.

Hovy, E.; Gerber, L.; Hermjakob, U.; Lin, C.; and Ravichandran, D. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the Human Language Technology Conference (HLT-2001)*.

Ittycheriah, A.; Franz, M.; Zhu, W.; and Ratnaparkhi, A. 2001. Question answering using maximum-entropy components. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*.

Miller, G. 1995. WordNet: a lexical database. *Communications of the ACM* 38(11):39–41.

Moldovan, D.; Harabagiu, S.; Paşca, M.; Mihalcea, R.; Gîrju, R.; Goodrum, R.; and Rus, V. 2000. The structure and performance of an open-domain question answering system. In *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics (ACL-2000)*.

Paşca, M. 2001a. *High-Performance, Open-Domain Question Answering from Large Text Collections*. Ph.D. Dissertation, Southern Methodist University, Dallas, Texas.

Paşca, M. 2001b. Unveiling next generation search technologies: Answer extraction on the Web. In *Proceedings of the 2nd International Conference on Internet Computing (IC-01)*.

Srihari, R., and Li, W. 2000. A question answering system supported by information extraction. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-2000)*.

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