Patterns of Potential Answer Expressions as Clues to the Right Answers

M. M. Soubbotin
InsightSoft-M, Moscow, Russia
MSoubbotin@insight.com.ru

Abstract

The core of our question-answering mechanism is searching for predefined patterns of textual expressions that may be interpreted as answers to certain types of questions. The presence of such patterns in analyzed answer-string candidates may provide evidence of the right answer. The answer-string candidates are created by cutting up relatively-large source documents passages containing the query terms or their synonyms/substitutes.
indicative patterns.
The specificity of our approach is:
- placing the use of indicative patterns in the core of the QA approach;
- aiming at the comprehensive and systematic use of such indicators;
- defining various structural types of the indicative patterns, including nontrivial and sophisticated ones;
- developing accessory techniques that ensure effective performance of the approach.
We believe that the use of indicative patterns for question answering can be considered as a special case of the more general approach to text information retrieval that contrasts with linguistics-oriented methodology.

Introduction

We decided to participate in the TREC-10 Question Answering track with a purpose to test certain specific features of the text processing technology we are developing in the framework of our CrossReader project. This technology is aimed at presenting to a user the needed information directly, i.e. instead of documents, or sources containing potentially relevant information. The query-relevant sentences or short passages are extracted from the processed documents and judiciously arranged; so, new full texts emerge that are focused precisely on the user's subject (Subbotin, 1993; Gilyarevskii, 1993; Perez, 2001).

The latest version of this technology - the TextRoller system - uses not only key words, but also positive and negative patterns for choosing and arranging text items. For the TREC-10 Question Answering task we have developed a variant of our basic technology that searches for candidate answers using key words (from the question text) and chooses the most probable answer using patterns. The participation at TREC-10 was a test for some basic mechanisms of our technology. Now, after this test was successfully passed, these mechanisms will be implemented in the new TextRoller versions.

Basic Features of the Applied Approach

It seems that many systems participating at TREC QA track represent the question (or its reformulations) as a set of entities and relations between them in order to compare these entities and relations with those of candidate answers texts; the answer candidate that correlates at the highest degree with the question text gets the highest score. By contrast, our QA system checks the answer candidates for the presence of certain predefined indicators (patterns) to which scores were assigned beforehand, i.e. independently of the question text analysis. Candidate snippets containing the highest-scored indicators are chosen as final answers.

It is obvious from the above that the applied approach does not require NLP or knowledge-based analysis of the question text. This text is considered as just a string consisting of various substrings. These are used, first of all, for composing queries helping to retrieve passages containing answer candidates. If present in candidate answers texts, they are considered as a condition of applicability of a given indicative pattern for a given question, but they do not influence the score of the pattern (as said above, it is predefined beforehand).

The efficiency of this approach depends on the quantity and diversification of predefined indicative patterns as well as on the recall of passages containing candidate answers.

We could not rely on the presence of predefined patterns in the texts of candidate answers for every question. If case of neither pattern was found, the system used the more common way to choose among candidate answers basing on lexical similarity between the question and an answer snippet. From 289 answer strings that were correct responses 193 did contain the patterns. Non-matching any patterns, but containing question (query) terms were 64. Other (containing minor indicators, such as capitalized words, or randomly selected) - 32.

To some extent, many QA-Track participants (at TRECs 8 and 9) had used what we call the indicative patterns. The specificity of our approach is:

- placing the use of indicative patterns in the core of the QA approach;
- aiming at the comprehensive and systematic use of such indicators;
- defining various structural types of the indicative patterns, including nontrivial and sophisticated ones;
- developing accessory techniques that ensure effective performance of the approach.

In (Sanda Harabagiu et al., 2000) the term "definition patterns" was introduced as "associated with questions that inquire about definitions". This kind of patterns was widely used by our QA system, although in many cases they were effective in combination with some additional indicators (see section "How patterns work"). It is also noteworthy that we did not confine the use of these patterns to questions inquiring about definitions. We assume, in general, that there should not be one-to-one correspondence between a given pattern and a question type. The same pattern can be applicable in answering many types of questions (getting a different score for each question type).

The Library of Indicative Patterns

The indicative patterns used by our QA system are sequences or combinations of certain string elements, such as letters, punctuation marks, spaces, tokens (such as "&", "%", or "$"), digits, and words/phrases that are accumulated in special lists (see Fig. 1).

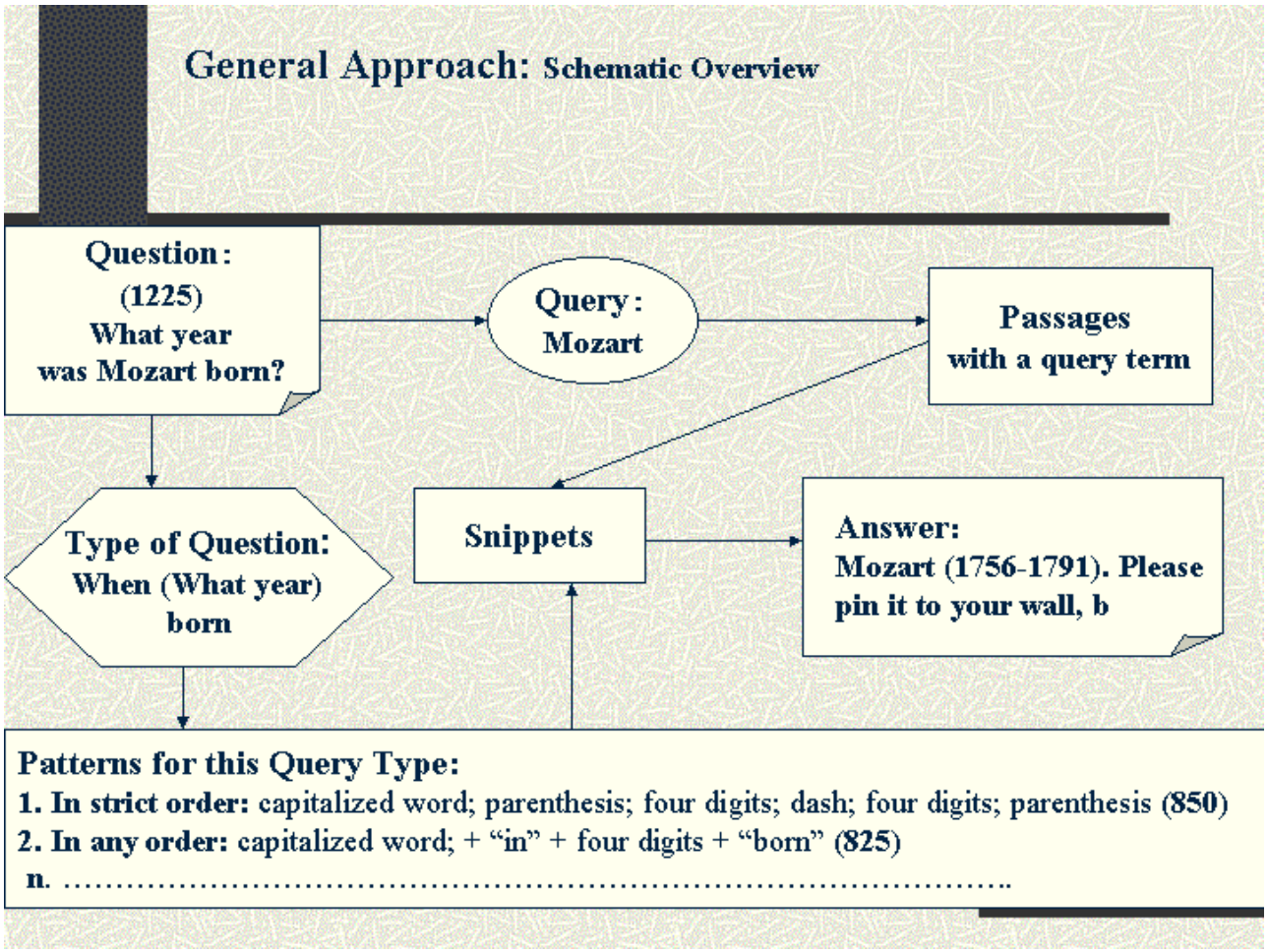

Fig. 1. The general approach

The way we defined indicative patterns is totally heuristic and inductive.

At the initial stage the indicative patterns lists are accumulated basing on expressions that can be interpreted as answers to the questions of a definite type. For example: "Milan, Italy" present in any text passage can be considered (completely independently from the whole sense of the passage) as an answer to the question "Where is Milan". So, a pattern for the "Where" question type may be created: "city name; comma; country name". The string "Mozart (1756-1791)" contains answers to the questions about Mozart's birth and death dates, allowing construction of the pattern: "a person's name; parenthesis; four digits; dash; four digits; parenthesis ". We studied texts systematically with the purpose of identifying expressions that may serve as models for answer patterns. Some patterns components can be used for searching more complex structure patterns. The validity of a pattern for a given question type (and its score) can be tested in large text corpora.

The library of patterns can never be complete.

Identifying patterns while studying text corpora is a research field by itself, accumulating special knowledge on cooccurencies of text elements (characters, strings, as well as definite classes of characters and strings). So, it can be found that the string "Mr. " at a certain frequency level precedes one or two capitalized words, and the string "Jr." follows such words, etc.

Thus, we can accumulate the knowledge on "typical" combinations and correlations of strings that correspond to personal names, to a persons age, to locations, dates, activities, etc. This requires the use of sophisticated tools and special methods. This knowledge area can become important not only for QA, but also for other text retrieval tasks. For example, we use such methodology for extracting and ordering of sentences resulting in a coherent description of the requested subject.

The Structure of Indicative Patterns

A pattern may include a constant part and a variable part. The latter can be represented by a query term or even an unknown term (the answer word/phrase proper that occupies a definite position in the sequence of pattern elements).

We distinguish between two pattern categories: the first represents a complete structure while the second is a composite structure of specific pattern elements (see above). For TREC-10 we had prepared 51 lists of various patterns elements; for each question category 5 - 15 of such lists were applied for recognition of potential answers (see Fig. 2).

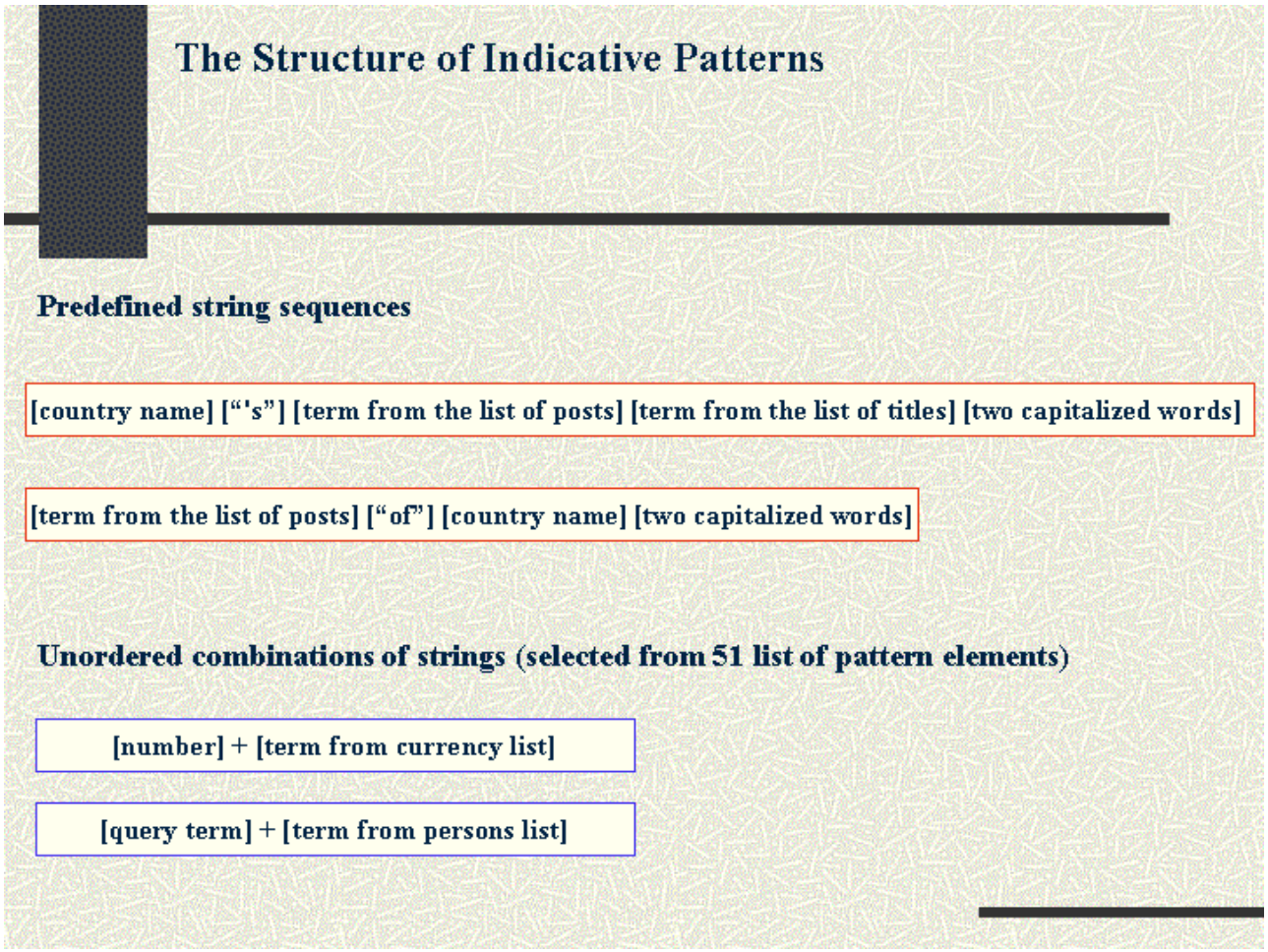

Fig. 2. Structure of the patterns

Usually, patterns with more sophisticated internal structure are more indicative of the answer. So, for answers to the question type "Who is the President (Prime Minister, etc.) of a given country" we found various combinations of elements that can be present in an answer expression: words that signify the name of a country, the post a person occupies, a proper name, a title, punctuation marks, etc. Let us denote countries by "a", posts by "b", proper names (first and last) by "w", titles (e.g. "His Excellency") by "e". The presence of combinations "abeww"; "ewwdb,a", "b,aeww" in an analyzed string indicates a correct answer to this question type with high probability.

The validity of certain simple structure patterns (e.g. the "definitions patterns") is dependent on the presence of additional positive and negative indicators (see below).

We distinguish between 6 basic definition patterns.

Below, these patterns are represented as sequences of their constituent elements. We will denote the primary query word present in the "snippet" with A, and the supposed "actual answer" with X; the pattern elements are divided by a semicolon. We consider two subtypes with the same structure, but where A and X occupy the inverse positions, as belonging to same pattern type.

1. <A; is/are;[a/an/the]; X>
<X; is/are;[a/an/the]; A>
Example: "Michigan's state flower is the apple blossom".

2. <A; comma; [a/an/the]; X; [comma/period]>
<X; comma; [a/an/the]; A; [comma/period]>
Example: "Moulin Rouge, a cabaret ".

3. <A; [comma]; or; X; [comma]>
Example: "shaman, or tribal magician,".

4. <A; [comma]; [also] called; X [comma]>
< X; [comma]; [also] called; A [comma]>
<X; is called; A>
<A; is called; X>
Example: "naturally occurring gas called methane".

5. <X, dash; A; [dash] A; dash; X; [dash]>
Example: "nepotism - hiring relatives for the better jobs".

6. <X; parenthesis-; A; parenthesis >
<A; parenthesis; X; parenthesis >
Example: "myopia (nearsightedness)".

As said above, these patterns were used not only for answering the "definition questions", but also for "Who-", "Where-", and other question types.

The expressions matching a pattern often show no structural similarity with the question text (see, for instance, the example in Fig 1). A lot of expressions in text corpora convey information that can be interpreted as answering a certain question without any special intention to do it (e.g. the standard beginning of agency news: "Milan, Italy..." answers the question "Where is Milan?".

How Patterns Work

Presence of certain patterns in the snippet-candidate serves as an almost guaranteed indication of the right answer (see Fig. 1). Their high score lets to choose the answer string with confidence. Lower score patterns cannot guarantee the correct response as they can be present in a number of candidate answer strings both correct and wrong. For some of such patterns we used additional indicators of validity. When there are several candidates with the same pattern, the system checks the text of candidate answers (and their surrounding) for presence of such additional indicators.

This is the case, in particular, for "definition patterns". Among the additional indicators for them there are such as absence of an article, presence of a stop-list word or other word lists.

Some recently-emerged text-processing techniques claim for using patterns while identifying relevant content. The best known is wrapper induction, an information-extraction technique that is considered an alternative to NLP-based methods (Kushmerick, 2000; Kushmerick, 1999; Adams, 2001). Wrappers demonstrate that extensive linguistic knowledge is not necessary for successful IE. For example, research on a collection of email conference announcements shows that speaker's names are often prefixed by "Who" and many names begin with the title "Dr." (Fraitag, 2000).

However, wrapper induction in its present-day form is resource specific, it extracts information from particular Web sites. It uses specific features related to the document formats rather than the ways information is commonly presented in written texts.

Overview of the Process Flow

Preconditions for effective use of the method are:
- Detailed categorization of question types (for example, we distinguish between nine "Who"-question types ("Who-Post"; "Who-Author", etc);
- The great variety of patterns for each type (for "Who-Author"-type we have 23 patterns);
- A sufficiently large number of candidate answers to each question (usually we get several hundreds or thousands of candidate snippets).

Multiple overlapping answer-string candidates ("snippets") are created by cutting source documents passages containing the query terms or their substitutes (see Fig. 3).

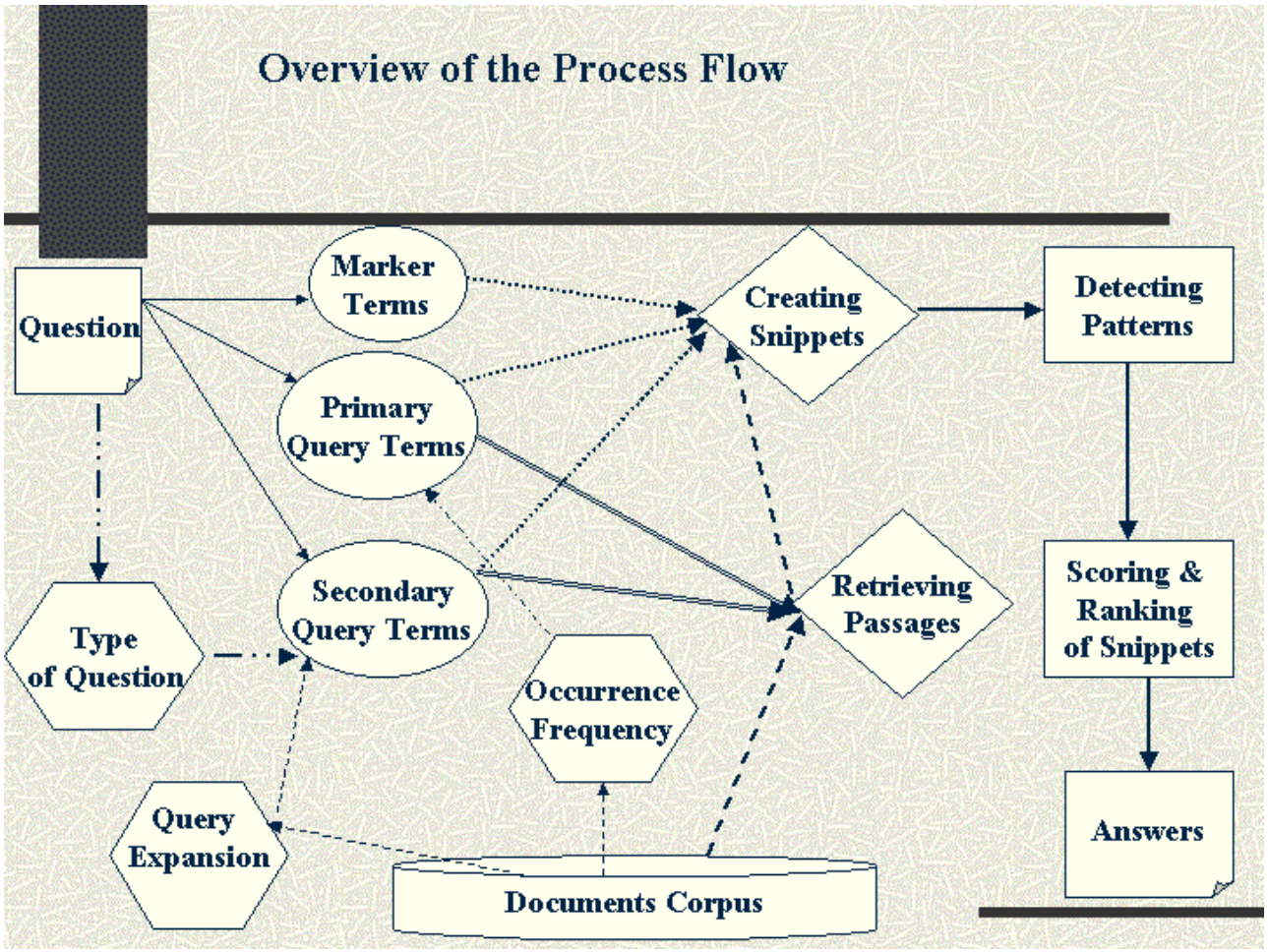

Fig. 3. Overview of the process flow

Using of specific question words (in contrast to common words) as query terms ensures in most cases that the question subject is addressed in the source passages.

In the literature we find approaches attempting to distinguish between the main (primary) and additional (secondary) query words. In (Sneiders, 1998) this distinction is discussed as applied to searching for answers to FAQs, where the answers are represented as sentences. Primary keywords are the words that convey the essence of the sentence. They cannot be ignored. Secondary keywords are the less-relevant words for a particular sentence. They help to convey the meaning of the sentence but can be omitted without changing the essence of the meaning.

We accept this distinction by assuming that the primary terms are question-specific words and are almost inevitably present in the passage that treats the same subject as the question. We use certain criteria of the specificity, including the minimal occurrence in the documents corpus (for example, "mortarboard" in the question "Where on the body is a mortarboard worn").

In some question categories, primary query words do not convey the question subject completely, requiring secondary searching terms. Such terms are, for example, the words signifying a certain

post in questions of the type "Who is (the) X of (the) Y?", where X is a post, and Y is the name of a country, company, organization, etc.

For some question types the secondary query terms should be supplemented by their related words. For this purpose we use a query-expansion technique. Our expansion module extracts query-related terms not from the full documents, but from the short relevant text passages.

The retrieved passages are cut into 50-byte snippets. They are cut around the query words, as well as around other question words that have not served as query terms (these are denoted as "markers" in Fig. 1).

All the snippets are analyzed to identify patterns that are indicative of a potential answer (as described above).

The Results and the Perspectives of the Approach

Our mean reciprocal rank (strict): 0.676; mean reciprocal rank (lenient): 0.686. System was confident for 372/492 (75 %) of the questions. Of those, 289/372 (77 %) were correct responses. Two thirds of correct answer strings were obtained using patterns thus proving the feasibility of the applied approach.

We believe that the use of indicative patterns for question answering can be considered as a special case of the more-general approach to text information retrieval that contrasts with the linguistics-oriented methodology.

Generally, text documents contain information that is included not intentionally, but due to its indirect interconnections to what the author directly conveys. This implicit information can be addressed systematically by a set of patterns. We are conducting investigations that will allow us to develop appropriate tools for this.

Aiming at the practical implementation of indicative patterns approach, we are currently developing advanced versions of our TextRoller technology that uses both query terms and patterns while assembling new "full texts" from appropriate passages of the processed documents. The patterns are used not only for choosing passages, but also for ensuring their judicious arrangement, as well as domain specificity and readability of the constructed text.

Bibliography

Adams Katherine C.. The Web as a Database. New Extraction Technologies & Content Management. Online, March/April 2001, pp. 28 - 32

Fraitag, Dayne. Two Approaches to Learning for Information Extraction. Talk at UC, San Diego (San Diego, CA) on October 16, 2000

Gilyarevskii R., M. Subbotin. Russian Experience in hypertext: automatic compiling of coherent texts. Journal of the American Society for Information Science, 1993, v.44, 4, pp. 185-193

Harabagiu Sanda, Dan Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan Bunescu, Roxana Girju, Vasile Rus and Paul Morarescu. Falcon : Boosting Knowledge for Answer Engines. Proceedings of the Text Retrieval Conference (TREC-9), 2000

Kushmerick, Nicolas. Cleaning the Web. IEEE Intelligent Systems, vol. 14, No. 2 (1999): http://www.cs.ucd.ie/staff/nick/home/research/download/kushmerick-ieeeis99.pdf

Kushmerick, Nicolas. Wrapping up the Web.
Synergy: Newsletter of the EC Computational Intelligence and Learning Cluster Issue 2 (Spring 2000) http://www.dcs.napier.ac.uk/coil/news/feature46.html

Perez, Ernest. Finding Needle in the Textstacks. Online, September/October 2001

Sneiders, E.. [The Development of an FAQ Answering System].
Information Systems in the WWW Environment. IFIP TC8/WG8.1 Working Conference, 15-17 July 1998, Beijing, China. Published by Chapman & Hall on behalf of IFIP, pp. 298-319

Subbotin M., D. Subbotin. INTELTEXT: Producing Coherent Linear Texts While Navigating in Large Non-Hierarchical Hypertexts. Lecture Notes in Computer Science, N 753, Springer-Verlag, 1993, pp. 281-289

Acknowledgments: The authors are thankful to Mr. Daniel Kamman for help with preparation of the English version of this article.