Use of Patterns for Detection of Answer Strings: A Systematic Approach

Martin M. Soubbotin, Sergei M. Soubbotin
InsightSoft-M, Moscow, Russia
http://insight.com.ru/

Abstract

The paper describes the Question Answering approach applied first at TREC-10 QA track and developed systematically in TREC 2002 experiments. The approach is based on the assumption that answers can be identified by their correspondence to formulas describing the structure of strings carrying certain (generalized) semantics, supposed by the question type. These formulas, or patterns, are like regular expressions but include elements corresponding to predefined lists of terms. Complex patterns can be constructed from blocks corresponding to such semantic entities as persons' or organizations' names, posts, dates, locations, etc. Using various combinations of blocks and intermediate syntactic elements allows to build a great variety of patterns. Exact position of elements corresponding to the "exact answer" was localized within the structure of each pattern. Each pattern is characterized by a generalized semantics, thus the pattern-matching string must be checked for correlation with the question terms and/or their synonyms/substitutes.

Essentials of the Approach

In 2002 TREC QA track tests we have further developed the approach described in [Soubbotin, 2001]. In general, our method lies in the domain of approaches examining the potential of information extraction for question answering tasks [Srihari, Wei Li, 1999; De Boni, 2001]. The evolution of IE systems, as represented, in particular, at Message Understanding Conferences (MUCs), shows a certain shift from deep text analysis based on computational linguistic and NLP methods to surface techniques [Eagles, 1998]. Our approach can be considered as being in line with this tendency.

More specifically, our approach is based on the use of formulas describing the structure of strings likely bearing certain semantic information. For example, string "FBI Director Louis Freeh" can be recognized, according to one of such formulas, as likely bearing the following information: a person represented by his/her first and last names occupies a (leading) post in an organization. The formula for this string is: a word composed of capital letters; an item from the list of posts in an organization; an item from the list of first names; a capitalized word. We can mark two first items in this formula as "exact answer", if we want to get answer to the question "Who is Louis Freeh?", and two last items, if the question is "Who is FBI head?" (question 1583 at TREC 2002).

First used at TREC-10 QA track, formulas of such kind were called "patterns" [Soubbotin M.M and Soubbotin S.M, 2001]. The term "pattern" is widely used in the field of Information Extraction. Our concept of patterns as structural formulas for strings is obviously different from that in "traditional" IE field, but keeping this difference in mind, we consider it convenient to use this term.

Each pattern is characterized by a certain generalized semantics, because the formulas' items refer to certain semantic categories (e.g., "posts") and not to specific semantic units (e.g., "president", "head", "director"). Therefore, after a string corresponding to a formula is recognized, the next step is to identify the question terms (or their synonyms/substitutes) within it or in its surrounding. To increase the likelihood of getting the right answer, the surrounding of the found string must be checked for the presence of expressions negating its semantics (e.g., "former", "-elect", "deputy", etc., located before or after the term from the list of posts).

After a question's type is defined (e.g., question about a person occupying certain post in an organization, question about husband/wife/relative of a person, question about acronym, etc.), a set of formulas, prepared for this type, is applied to match the strings in question-relevant passages.

Our approach does not need to distinguish linguistic entities in the text. We handle the source text strictly as string, i.e. consisting only of characters. The patterns used in our QA approach are aimed only at recognizing sequences of elements that correspond to the predefined formulas.

As surface patterns, our formulas for strings are similar to wrappers [Adams, 2001; Kushmerick, 2000] and look like regular expressions. However, patterns used by the wrapper techniques are mostly resource-specific, they relate to the document formats rather than the ways information is presented in written texts per se. As for difference from regular expressions, it is worth noting that patterns, that we use, include elements referring to the lists of predefined words/phrases.

Currently, increased attention is seen on surface approaches in QA. In some recent publications surface patterns similar to those used by us were discussed [Magnini, et al., 2002; Brill, et al., 2002; Brill, et al., 2001; Ravichandran and Hovy, 2002; Hovy et al., 2002].

Patterns and Question Types

The IE task, as presented at its main forum - the Message Understanding Conferences (MUCs), is focused on certain topics, or domains (Terrorism, Management Successions, Natural Disasters, Outbreaks of Infectious Diseases, etc.). The QA task requires another way to categorize the addressed Information.

The usual praxis of TRECs' QA tracks participants is to predefine a set of potential question types. The questions accumulated from several TRECs represent a good source for defining question types on a more or less detailed basis. The paradigm of "information categories" defined by question types (in contrast to "topic/domain" paradigm) allows to create systematically a variety of patterns, basing on potential semantic relationships inside each question category.

So, for the question type "Who is person X?" we can presuppose - among the main alternative possibilities - that this person is known for the (top-level) position he/she occupies in a organization, company or government; for his/her contributions as author, inventor, founder, etc.; as outstanding figure in a professional area; as wife/husband/relative of a well-known person; as involved in well-known event (e.g., as a criminal/perpetrator). In each case, a relationship is established between two or more entities: person, post, and organization/company; author and work; etc. The same entities are present if the Who-questions refer to posts, authors, etc. (e.g.,

"Who occupies the post Y in the organization Z?".)

For most Where-questions, we can suggest geographical items as answers. This is achieved by constructing structural formulas like: item from the list of cities/towns/counties, etc.; comma; item from the list of countries/states. There are question types suggesting as answers combinations of digits with units of measurement or currencies names. Completeness of lists corresponding to "semantic" pattern elements is evidently important (e.g., the list of currencies must include not frequently used words, such as "dlrs").

The type of the processed question is defined basing both on its interrogative and on the presence of words/expressions that are included in the list of characteristic terms for the corresponding question type.

Complex Patterns

Complex patterns are formulas for strings expressing relationships between several semantic entities. There are some basic, typical, frequently used ways of expressing certain relationships between semantic entities in written texts of a language covering the most ways these relationships are expressed in text corpora. There are also less usual ways for expressing these relationships. (Our preliminary investigations show that information on companies' leaders (name, post, company) in more than a half of cases is expressed by strings corresponding to 5 main groups of our complex patterns). Thus, one can gradually embrace less frequent string structures ensuring the more complete covering of certain relationships by a set of patterns.

Each basic way to express a predefined relationship between semantic entities has a great variety of variants. For example, blocks corresponding to people names can include items from first names list, capitalized words, specific name elements such as "bin", "van", etc., capital letters, dots, abbreviations like "Sr." and "Jr.". Multiple ways of writing people's names are reflected in corresponding block formulas (e.g., only first names - for children and pets; combinations of first, middle and last names; people's names of various nationalities; names wits initials, etc.). Blocks corresponding to dates are composed from prepositions, articles, digits, month names, commas, dashes, brackets, special words/phrases like "early", "in the period of", "years ago", "B.C.", etc.

As some non-obligatory elements can be present or not in the corresponding strings, it is important to foresee the most complete set of possible variants for each basic pattern.

Semantic entities (e.g. personal and organizations' names, posts, locations) can be represented in a complex pattern by lists of elements with explicit semantics (words/phrases fitting in a corresponding semantic category), as well as by elements that do not per se bear any definite semantics (e.g., capitalized words); these elements can represent – with higher or lower probability - a certain semantic entity due to their presence in the complete pattern structure.

The validity of a pattern is dependent of its elements and structure.
According to our observations, the more complex a pattern's internal structure, the higher is its validity (reliability). As a rule, complex patterns containing many elements are more valid: the neighboring elements mutually confirm each other. If pattern elements corresponding to a person's name include only such indicators as capitalized words, validity of this pattern is low.

But if a capitalized word is preceded by an item from the list of first names plus an item from titles list, then the validity increases significantly. For resolving the ambiguity of capitalized words - whether they are proper names - the system recognizes the usually capitalized words, such as names of months, etc. (For issues of proper names identification see, for example, [Bikel, 1997; Viitanen, 2002]).

However, for some question types, structurally simple, small-sized patterns can be used to match answers for these questions with high likelihood (e.g., a sequence of four digits for "In what year"-questions; digits plus units of measurement for questions regarding length, area, weight, speed, etc.) So, for answering the question 1634 "What is the area of Venezuela?", a simple pattern allows to match the string "340,569 square miles".

Correlation between the Pattern-Matching String and the Question Semantics

As said above, multiple strings can correspond to each pattern structure. The suitable string can be recognized if words/phrases of the question (or their synonyms/substitutes) are present inside this string or In its surrounding. If a string is matched by complex, multi-elements pattern, the presence of question terms can be checked within it at certain predefined positions. This simplifies the task of verifying the suitability of a matched string. By contrast, strings matched by small-sized patterns usually do not contain all the terms expressing the semantics of a given question type. In this case, to verify sufficiency of correlation between a pattern-matching string and semantics of a question, the surrounding of the pattern must be explored.

The simpler a pattern's structure, the more significant is how question words are located in the surrounding of the pattern-matching string, i.e. at which distance, right or left to the pattern, in which position to other potentially present pattern-matching strings, etc. Other important factors are the number and total weight of question words present in the pattern's surrounding. The weight (rank) assigned to a question word/phrase (or to its substitute) was defined basing on its relative "specificity" in the documents corpus. The highest rank was assigned to quoted expressions and (chains of) capitalized words. Specificity of other words was determined basing on their occurrence frequency in the corpus.

Thus, the relative simplicity of a pattern's structure is compensated by the complexity of rules that should verify the candidate answer's correlation with the question semantics. For each question type, the patterns are grouped into two subsets: complex and (relatively) simple.

We think that the straightforward use of surface patterns for QA without applying a set of heuristic rules for checking the patterns surrounding (see [Hovy et al., 2002]) cannot ensure sufficiently reliable results.

The total score assigned to candidate answers is based both on pattern's reliability and on evaluation of question words's presence inside a pattern-matching string or in its surrounding.

Overview of the QA Process

The process flow includes the following main stages.

Defining the question types for all questions - basing on interrogatives and on the lists of characteristic terms for the corresponding question type.
Ordering the questions aimed at first processing the question types for which there are more reliable patterns.
Forming the query from question terms; ranking query terms according to their "specificity".
Modifying the query, if the search failed or if an answer's score is beneath a predefined threshold (single words can be used instead of phrases; terms from lists of substitutes are added).
Identifying the pattern-matching strings for this question type - applying first a set of complex patterns, then a set of simple patterns.
Checking for correlation between the pattern and the question's semantics.
Identifying the exact answer part in the pattern-matching string.
Calculating the total score for each candidate answer.
Selecting the top-ranking candidate.
Creating a record for the submission file.

Analysis of the Results

Analysis of our successes and failures at TREC 2002 allows to see some characteristic peculiarities of patterns approach for QA.

Our confidence-weighted score is 0.691. The way the obtained answers were ordered was based on the predefined order of question types. So, we have suggested that a simple but highly reliable pattern for questions of the types "(In) what year" and "When" will match in most cases the right strings (taken into account the correlation with the question semantics). As a result, our first 29 recognized answers belonged to questions of these types, among which only 3 answers were wrong. Of course, this influenced positively our confidence-weighted score.

Noteworthy, our answer to the question 1617 ("When did the Klondike gold rush occur?") was assessed as wrong. Our answer "1896" was based on the presence of this string in the sentence containing 3 from 4 question words: "In 1896, a prospecting party discovered gold in Alaska, a finding that would touch off the Klondike gold rush." Another this group answer assessed as wrong is "Victorian era" to the question "When was Benjamin Disraeli prime minister?" (the answer was got from the sentence "Benjamin Disraeli was the most famous Conservative leader of the Victorian era"). These examples show that answers obtained by use of patterns, even if they are not correct, are not senseless, and in many cases are semantically close to right answers. We consider this feature as important for real use of a patterns-based QA system.

Some answers assessed as unsupported demonstrate the same feature of the pattern method. To the question 1476 ("Who was the Roman god of the sea?") correct answer "Neptune" was obtained by matching the string "Neptune, the god of the sea". This string was present within the sentence describing the decoration of a building and was assessed as unsupported apparently on this ground. We think that the possibility to extract the right answer from non-relevant documents/passages, in fact, can be regarded as extending the capacity of the QA system.

12 answers were assessed as "inexact." The exactness of answers (as well as the percentage of right answers) can be increased by further completing the library of patterns and lists of predefined words/phrases. For question about the cost of the international space station we

obtained the answer "dlrs 40 billion"; the exact answer is "at least dlrs 40 billion." Our patterns for this question type include such blocks as currencies names, digits, numerals, and items from the list of adjusting expressions ("more than", "not less than", etc.); the expression "at least" was missing in this list.

The number of right answers was 271. From the 209 wrong answers 148 were "no answer". In the vast majority of these cases the passages where the answer strings might be matched were not found. This was mainly due to that the system was working primarily with the top 50 documents collections supplied for each question. Excluding the "NIL" answers, we can evaluate the rate of wrongly identified answer strings: 48 for 352 questions (13,6%).

After the end of the TREC test we have upgraded our QA system to process large documents collections more efficiently. Now, selective processing of questions to which the answers had not been recognized shows that, to a great extent, the right answers can be obtained instead of wrong "NIL."

Further Work

The similarity between the TREC-11 QA task (that was focused on getting exact answers) and information extraction tasks was an incentive for use of our surface patterns in the framework of the IE technology. Using a modification of the approach applied at TREC-11 QA tests we developed a domain-independent system that extracts information from unstructured texts and populates a database. This system, named “ExactAnswer”, identifies entities such as persons, organizations, locations, and other types of data as well as relationships between entities (e.g. persons in relation to organizations). The tests conducted on various kinds of unstructured texts show high degree of accuracy (over 95%).

Adding more power to the patterns method remains our continued task. We use patterns also in the software products that are developed in the framework of our long-term project (http://insight.com.ru/), aimed not only at extracting of text units, but also at combining them into complex structures, such as single-document and multi-document summaries, discourse and reasoning chains. We also intend to examine the theoretical aspects of patterns considered as structural formulas for text strings. Primarily, we mean a specific dimension of studying languages - as they are represented in written texts - aiming at revealing correlations between the structure of text strings and their semantics.

References

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

Bikel, D.M., et al., 1997. Nymble: a High-Performance Learning Name-finder. Proceedings of the Fifth Conference on Applied Natural Language Processing,Morgan Kaufmann Publishers, pp. 194-201.

Brill, Eric, Jimmy Lin, Michele Banko, Susan Dumais and Andrew Ng. Data-Intensive Question

Answering.
http://www.ai.mit.edu/people/jimmylin/publications/Brill-etal-TREC2001.pdf

Brill, Eric, Susan Dumais and Michele Banko. An Analysis of the AskMSR Question-Answering System.
http://research.microsoft.com/~sdumais/EMNLP_Final.pdf

De Boni, Marco. Information Extraction, Query-Relevant Summarization and Question Answering: an Overview. 2000-2001.

EAGLES. Preliminary Recommendations on Semantic Encoding. Interim Report. Information Extraction, May 1998.
http://www.ilc.pi.cnr.it/EAGLES96/rep2/node30.html

Hovy, Eduard, Ulf Hermjakob, Deepak Ravichandran. A question answer typology with surface text patterns.
http://www.isi.edu/~ravichan/papers/hlt2002isi.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

Magnini, Bernardo, Matteo Negri, Roberto Prevete, and Hristo Tanev. Towards Automatic Evaluation of Question/Answering Systems.
http://tcc.itc.it/research/textec/topics/question-answering/lrec2002.pdf

Ravichandran, Deepak and Eduard Hovy. Learning Surface Text Patterns for a Question Answering System. Proceedings of the ACL Conference, 2002.

Soubbotin, M. M. and Soubbotin, S.M. Patterns of Potential Answer Expressions as Clues to the Right Answers. TREC Proc., 2001.

Srihari, Rohini K. and Wei Li. Information Extraction Supported Question Answering. TREC, 1999.

Viitanen Sirke. Named entities in BRIEFS. 2002.
http://www.ling.helsinki.fi/users/stviitan/prosem.html