Answer Mining by Combining Extraction Techniques with Abductive Reasoning

Sanda Harabagiu, Dan Moldovan, Christine Clark, Mitchell Bowden, John Williams and Jeremy Bensley
Language Computer Corporation
Richardson, TX 75080

Abstract

Language Computer Corporation's Question Answering system combines the strengths of Information Extraction (IE) techniques with the vastness of axiomatic knowledge representations derived from WordNet for justifying answers that are extracted from the AQUAINT text collection. CICERO LITE, the named entity recognizer employed in LCC's QA system was able to recognize precisely a large set of entities that ranged over an extended set of semantic categories. Similarly, the semantic hierarchy of answer types was also enhanced. To improve the precision of answer mining, the QA system also relied on a theorem prover that was able to produce abductive justifications of the answers when it had access to the axiomatic transformations of the WordNet glosses. This combination of techniques was successful and furthermore, produced little difference between the exact extractions and the paragraph extractions.

Introduction

In 2003, the TREC QA track had two separate tasks: the *main task* and the *passage task*. LCC's QA system participated in both tasks. The main task combined three different question types: factoids, lists and definitions. Factoid questions seek short, fact-based answers in the document collections, e.g. Q1910: *"What are pennies made of?"*. Some factoid questions may not have an answer in the AQUAINT collection, and thus the correct answer in this case is *NIL*. Otherwise, the correct, exact answer is an entity, e.g. *steel* for Q1910. Factoid questions were evaluated similarly in the 2002 TREC QA track. This year however, the score of the main task was computed as a weighted average of the factoid score with the scores obtained for processing list questions and definition questions:

$$\text{Main\_task\_score} = \frac{1}{2} \times \text{factoid\_score} + \frac{1}{4} \times \text{list\_score} + \frac{1}{4} \times \text{definition\_score}$$

This formula shows that in 2003, a QA system with very good performance (e.g. 76%) on factoid questions and with only 28% performance for list questions and 40% on definition questions would have achieved 55% overall performance. Another system, with much worse performance on factoid questions (e.g. 40%) but better performance for list questions (e.g. 60%) and definition questions (e.g. 80%) would have achieved the same performance. In general, a list question requests a *set* of instances of specified types, such as Q2014: *"List brands of pianos."* or Q1940: *"What grapes are used in making wine?"*. The response to a list question is a non-null, unordered and unbounded set of answer instances. In previous years, the cardinality of the set of list elements or instances was specified in the question. In the 2003 TREC, the list questions did not specify the target number of responses. The final answer set for any list question was created from the union of the distinct, correct responses returned by all participants plus the set of answers found by NIST assessors during question development. This final answer set was used for computing the F-measure of the question, which equally weighted the instance recall (IR) and the instance precision (IP). These measures were defined as:

- IR = #instances judged correct and distinct / #answers in the final set
- IP = #instances judged correct and distinct / #instances returned
- F = (2 × IP × IR) / (IP + IR)

The score for the list component was the mean of the F-scores of the list questions. The response to a definition question was also measured by the F-score, but the interpretation of the final set was different.

For each definition question, the assessor has created a list of acceptable information nuggets from the union of the returned responses and the information discovered during question development. Some of the nuggets are deemed essential, i.e. a piece of information that must be in the definition of the target in order to consider it a valid definition. The remaining nuggets in the list are acceptable. Once the list of acceptable nuggets are created, the assessor decides upon the acceptable and essential nuggets returned by each system for each question. Each nugget was matched only once. The definition questions were scored using nugget recall (NR) and an approximation of nugget precision (NP) based on length. These scores are combined using the F-measure in which NR is five times more important

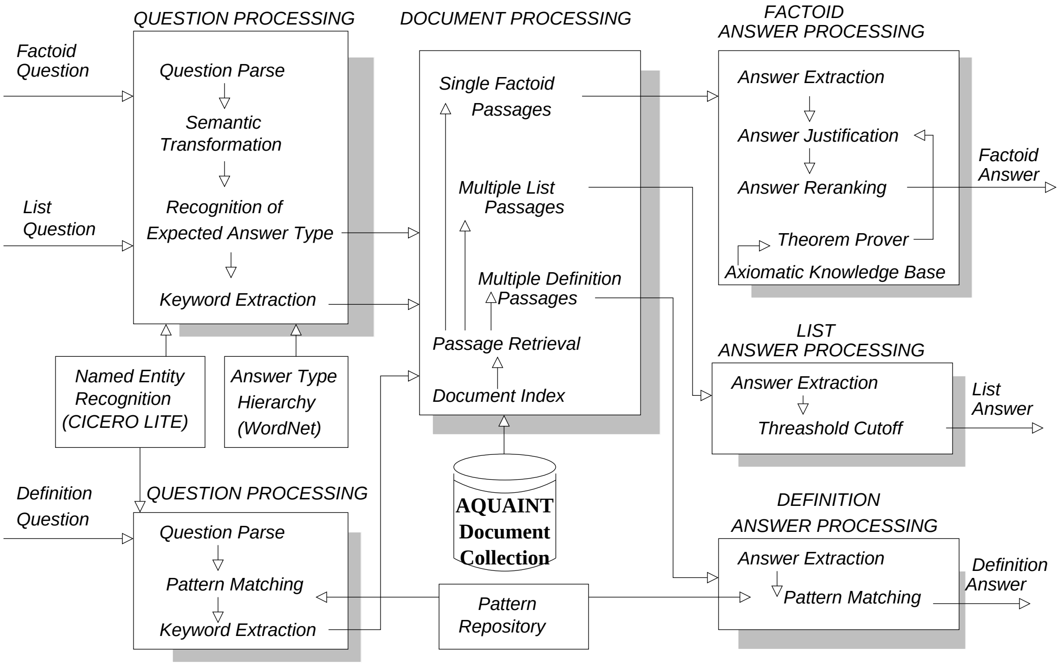

Figure 1: Architecture of LCC's QA system

than NP. The formulae used for measuring the definition score are:

• NR = #essential nuggets returned in response / #essential nuggets

• NP is defined using:

- *allowance* = (100 × [#essential and acceptable nuggets returned]); and

- *length* = (total #non-white space characters in answer strings):

NP = 1, if length < allowance;

NP = 1 - [(length - allowance)]/length, otherwise

• F = (26 × NP × NR) / (25 × NP + NR)

In TREC-2003, 413 factoid questions were evaluated; 37 list questions and 50 definition questions. Table 1 shows the distribution of the corresponding answers. The ideal system had to extract 383 exact answers for factoid questions, identify 30 NIL questions but also discover 549 list instances and 207 essential nuggets of definitions. If the list instances and the definition nuggets are approximated as exact answers, the factoid answers returned by this ideal system would have accounted only for 34% of the entire set of answers. The main-task score, although devised before knowing the cardinality of the final sets for the list and definition questions, attributes a 50% importance to factoid questions even for an ideal system. In other words, answering factoid questions was as important for the mai task of the 2003 TREC QA evaluation as answering list or definition questions.

The passage task, in contrast, used only the factoid questions from the main task. A submission could contain only one response for each question, which could be either the string *NIL* or an extract from the document.

Answer type	*Count*
Answers to factoid questions	383
NIL-answers to factoid questions	30
Answer instances in List final set	549
Essential nuggets for Definition questions	207
Total nuggets for Definition questions	417

Table 1: Distribution of answers in TREC-2003

A document extract is any text snippet of length smaller or equal to 250 bytes. This definition of he extract allowed the evaluations for this task to be performed by following the same procedure as in TREC-2001. A passage could be (a) incorrect, when it did not contain the answer; (b) unsupported, when it contained the answer but the document did not support the answer; or (c) correct. The final score was the fraction of the questions judged correct.

The architecture of the QA system

The architecture of LCC's QA system is illustrated in Figure 1. It is to be noted that the question processing module is identical for factoid and list questions, but different for definition questions. To process factoid or list questions, the QA system needs to identify the expected answer type encoded either as a semantic class recognized by CICERO LITE^{TM}, our Named Entity Recognizer, or in a hierarchy of semantic concepts, built using the WordNet hierarchies for verbs and nouns. The expected answer type is typically indicated by the head of one of the question phrases. The recognition of this

QUANTITY	55	ORGANIZATION	15	PRICE	3
NUMBER	45	AUTHORED WORK	11	SCIENCE NAME	2
DATE	35	PRODUCT	11	ACRONYM	1
PERSON	31	CONTINENT	5	ADDRESS	1
COUNTRY	21	PROVINCE	5	ALPHABET	1
OTHER LOCATIONS	19	QUOTE	5	URI	1
CITY	19	UNIVERSITY	3		

Figure 2: Name classes

head is based on syntactic dependencies as well as some semantic dependencies that are approximated from the question parse. For example, in the case of the factoid question Q1997: *"What American revolutionary general turned over West Point to the British?"*, the expected answer type is PERSON, a Named Entity class, determined by the noun *general* which is found in the hierarchy of humans in WordNet. The same expected answer type is found when processing the list question *"Who are professional female boxers?"*. But the main difference among the two kinds of questions stems from the fact that in the case of factoid questions, the system will search for a unique PERSON whereas for list questions, it will try to identify as many PERSONS as possible in any relevant passage.

When processing definition questions, the questions are parsed in order to detect NPs. Then NPs are then matched against a set of patterns. For example, question Q2041: *"What is Iqra?"* is matched against the pattern <What is Question-Phrase>, which is associated with the answer pattern <Question-Phrase, which means Answer-Phrase>. In this case, <Question-Phrase> is *Iqra* and the answer phrase will be extracted from *"Iqra , which means ' read ' in Arabic , was the first word that the arch - angel Gabriel is said to have spoken to Islam 's Prophet Mohammed"*.

The Document Processing module is the same for any of the three forms of questions, as it retrieves relevant passages based on the keywords provided by question processing. For factoid questions, it ranks the candidate passages after filtering out all passages that do not contain a concept of the expected answer type. In the case of list questions, it prefers passages having multiple occurrences of the expected answer type. In the case of definition questions, it allows multiple matches of keywords.

Answer extraction is performed differently for each form of questions. In the case of factoid questions, answers are first extracted based on the recognition of the answer phrase provided by CICERO LITE. If the answer is not extracted as a named entity, it is justified abductively by using a theorem prover that makes use of axioms derived from WordNet as well as other axioms approximating semantic relations or linguistic pragmatics. For example, for the factoid question Q2252: *"What apostle was crucified?"*, because apostles are classified in WordNet under PERSON, the CICERO LITE Named Entity Recognizer is not able to detect names of saints as PERSON, since it was not trained to do so. In the most informative paragraph, two names of apostles are found: the apostle Peter and the apostle Paul. Because the verb *crucified* is not used as a keyword, the candidate answer becomes *the apostle Paul*. But when abduction is performed, the correct answer, *the apostle Peter* is returned as the exact answer.

The extraction of definition answers relies on pattern matching. The answer phrase (AP) is identified based on the results of the parse. For the answer of question Q2041, two nuggets from the AP were evaluated as vital: *Arabic word for read* and *Gabriel's first word to Mohammed.*

List questions are extracted by using the ranked set of paragraphs and their corresponding exact answers. The paragraphs are processes with the goal of finding a cutoff measure based on the semantic similarity of answers. This cutoff measure determines the number of elements in each list answer. For example, the answer to question Q2014: *"List brands of pianos."* is the list *Ivers Pond*, *Baldwin*, *Boesendorfer*, *Steinway*, *Yamaha.*

Extracting answers for factoid questions

Our Question Answering system extracted 289 correct answers out of 500 factoid questions. Out of these, 234 correct answers were obtained by extracting the answer which was identified by the CICERO LITE system or recognizing it from the Answer Type Hierarchy. Table 2 illustrates some of the factoid questions that asked for city names as well as the answers returned by our system.

1898: What city is Disneyland in? Answer: Anaheim
1916: What city did Duke Ellington live in? Answer: Washington D.C.
1986: What city is Ole Mississippi University in? Answer: Oxford, Miss.
1912: In which city is the River Seine? Answer: Paris

Table 2: Factoid questions asking for city names

Similar to previous TREC QA evaluations, the Named Entity Recognizer had to identify a varied set of semantic classes. Figure 2 lists some of the semantic classes of the names as well as the number of times

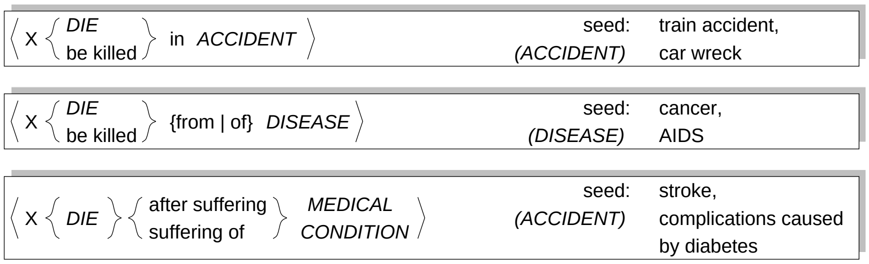

Figure 3: MANNER-OF-DEATH patterns

each of them were recognized correctly when extracting a factoid answer.

For locations, CICERO LITE distinguished among countries, cities, provinces, continents and other locations. After QUANTITY and NUMBER, the semantic classes associated with locations were the most numerous. A special class of names was trained for authored work, like names of books, songs or poems. Table 3 illustrates some of the factoid questions that were answered by this class of names.

1934: What is the play "West Side Story" based on? Answer: Romeo and Juliet
1976: What is the motto for the Boy Scouts? Answer: Be Prepared
1982: What movie won the Academy Award for best picture in 1989? Answer: Driving Miss Daisy
2080: What peace treaty ended WWI? Answer: Versailles
2102: What American landmark stands on Liberty Island? Answer: Statue of Liberty

Table 3: Questions asking for names of authored works

71% of the factoid questions were answered correctly because of a name that was recognized by CICERO LITE. The other 29% of the correctly answered factoid questions were answered by concepts that are represented in the answer type taxonomy employed by our system. The vast majority of these conceptual taxonomies classify concepts into such classes as: DISEASE, DRUGS, COLORS, INSECTS or GAMES.

There is one particular hierarchy that deserves more discussion. It is the MANNER-OF-DEATH category, developed because of previous TREC questions like *"How did Adolf Hitler die?"*. Text mining techniques for identifying such information were developed, based on lexico-semantic patterns from WordNet that were re-enforced in texts. For example, one such pattern is [kill#sense 1(verb) $\rightarrow$ cause $\rightarrow$ die#sense 1(verb)]. Some of the troponyms of the first sense of the verb *kill* are candidates for the MANNER-OF-DEATH hierarchy, e.g., *drown*, *poison*, *strangle*, *assassinate*, *shoot*. However, since not all MANNER-OF-DEATH are lexicalized as verbs, we set out to determine additional

1921: How did Virginia Woolf die? Sentence answer: When someone dies quoting from Virginia Woolf 's own *suicide* note , you think there must be even further options
1927: How did George Washington die? Sentence answer: Washington died from a *throat infection* at age 67 , almost three years after leaving the presidency
1928: how did Patsy Cline die Sentence answer: Who else died in the *plane crash* that killed Patsy Cline
1939: How did Einstein die? Sentence answer: Some 15,000 die from *ruptured abdominal aortic aneurysms* each year
2012: How did Marty Robbins die? Sentence answer: The late country - western singer , who died Dec. 8 , 1982 at age 57 after suffering a *heart attack* six days earlier at his Nashville , Tenn. , home , often would return home after a long day of rehearsing and still have enough voice left in him to deliver a private concert
2072: How did Brandon Lee die? Sentence answer: It was during filming of the original movie version of " The Crow " that actor Brandon Lee _ son of martial arts legend Bruce Lee _ was killed in an on - set shooting accident in 1993
2143: How did John Dillinger die? Sentence answer: On July 22 , 1934 , a man identified as bank robber John Dillinger was *shot to death* by federal agents outside Chicago 's Biograph Theater
2216: How did Dennis Brown die? Sentence answer: Initial reports suggested Brown died of *complications caused by respiratory problems* , but his cause of death had not yet been confirmed
2265: How did Marjorie Kinnan Rawlings die? Sentence answer: Alas , in that same year a *cerebral hemorrhage* dispatched her , only 57
2335: How did Harry Chapin die? Sentence answer: Chapin wrote some strong story - songs , but he was still a work in progress when he died in a *car accident* in 1981
2383: How did Jerry Garcia Die? Sentence answer: The Grateful Dead 's online ' OK ' will likely keep the buzz alive for a group that disbanded after lead singer Jerry Garcia died in 1995 of a *heart attack*
2386: How did Harry Houdini die? Sentence answer: In 1926 , magician Harry Houdini died in Detroit , suffering complications of a *ruptured appendix*

Table 4: Questions asking for MANNER-OF-DEATH

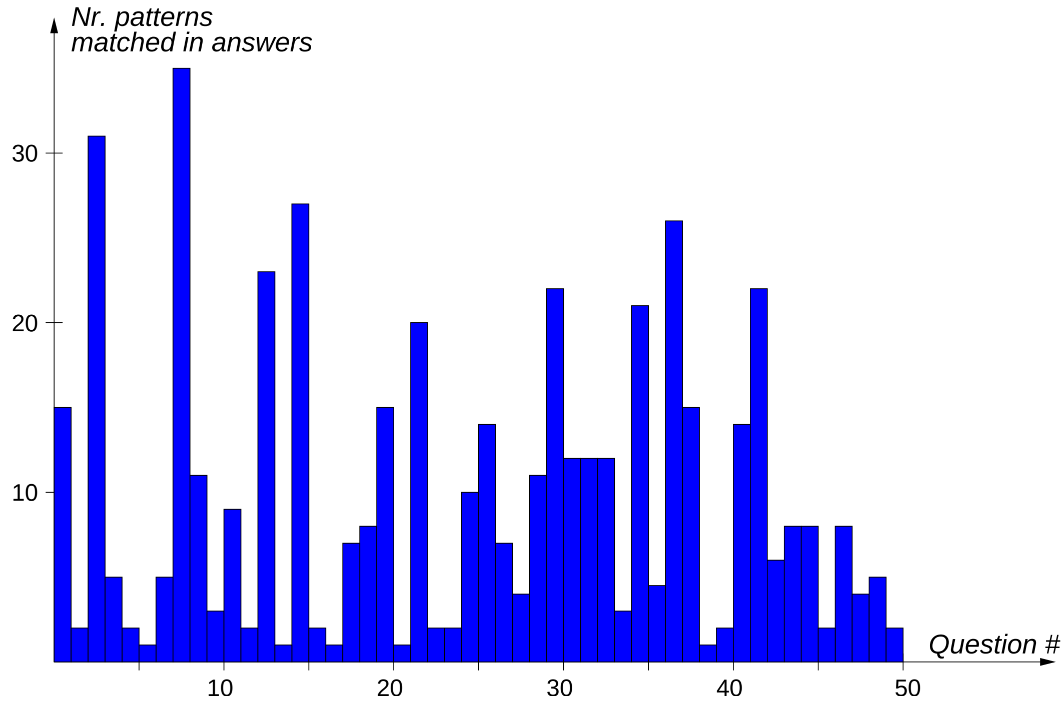

Figure 4: Number of answers extracted via patterns for each definition question

patterns that detect manners of death.

Especially for the cases when the cause of death is not lexicalized by a single noun or verb from the WordNet dictionary, we have developed a technique for acquiring (1) dictionaries for the cause-of-death; as well as (2) patterns that recognize manners of death. For this reason, we started with (a) a set of seed patterns and (b) a set of possible death causes. Figure 3 lists some of the seed patterns and their corresponding death causes.

By using the multi-level bootstrapping technique reported in (Riloff & Jones 1999) we populated this taxonomy with 100 concepts, which were manually verified. Table 4 lists factoid questions that are resolved by patterns extracted for MANNER-OF-DEATH questions.

CICERO LITE and the answer taxonomies alone are responsible for correctly extracting 234 answers. 65 additional correct answers are due to the theorem prover we employed, which was reported in (Moldovan *et al.* 2003). The role of the theorem prover is to boost the precision by filtering out incorrect answers, that are not supported by an abductive justification. For example, question Q2217: *"What country does Greenland belong to?"* is answered by *"Greenland, which is a territory of Denmark"*. Denmark is recognized as a COUNTRY name, therefore is extracted as an answer. Moreover, the gloss of the synset of {territory, dominion, province} is *"a territorial possession controlled by a ruling state"*. The logical transformation for this gloss is:

[control:v#1(e,x1,x2) & country:n#1(x1)
& ruling:a#1(x1) & possession:n#2(x2)
& territorial:a#1(x2)]

in which each lexeme has the format:

[root : part-of-speech # WordNet-sense].

All lexemes are predicalized, but verb lexemes have a special role: they have one argument e which stands for the eventuality of the event, state or action they represent (cf. Davidsonian treatment of actions) and their arguments stand for: x1=subject, x2=object. The subject and the object are recognized as predicates having the arguments x1 and x2 respectively. The same arguments are shared by the modifiers of the subjects and objects. Whenever the genus of the gloss was either one of the synset elements or one of its morphological variations (e.g. territorial for territory) the head of the genus indicates a specialization of the verbal predicate. In this case, the control is exercised by a possession, therefore the logical form of the gloss for sysnset {territory, dominion, province} can be specialized too:

[possess:v#2(e,x1,x2) & COUNTRY:n#1(x1)
& ruling:a#1(x1) & territory:n#2(x2)]

This specialized logical transformation also uses the unification between [control:v#1(e,x1,x2) & possession:n#2(x2) & territorial:a#1(x2)] and [possess:v#2(e,x1,x2) & territory:n#2(x2)]. Additionally, by using the logic form of the gloss of verb *belong*, which is *"be in the possession of"*, the predicate possess:v#2(e,x1,x2) may be replaced with belong:v#1(e,x1,x2), which resolves the abduction that proves question Q2217, if the answer is *Denmark*. The verb possess:v#2(e,x1,x2) and belong:v#1(e,x1,x2) express a form of meronymy which is not specifically

Id	*Pattern*	*Freq.*	*Usage*	*Question*
25	person-hyponym QP	0.43%	The doctors also consult with former Italian Olympic skier Alberto Tomba , along with other Italian athletes	1907: Who is Alberto Tomba?
9	QP , the AP	0.28%	Bausch Lomb , the company that sells contact lenses , among hundreds of other optical products , has come up with a new twist on the computer screen magnifier	1917: What is Bausch & Lomb?
11	QP , a AP	0.11%	ETA , a Basque language acronym for Basque Homeland and Freedom _ has killed nearly 800 people since taking up arms in 1968	1987: What is ETA in Spain?
13	QA , an AP	0.02%	The kidnappers claimed they are members of the Abu Sayyaf , an extremist Muslim group , but a leader of the group denied that	2042: Who is Abu Sayaf?
21	AP such as QP	0.02%	For the hundreds of Albanian refugees undergoing medical tests and treatments at Fort Dix , the news is mostly good : Most are in reasonably good health , with little evidence of infectious diseases such as TB	What is TB?

Table 5: Examples of definition patterns, usage and frequency of occurrence

encoded in WordNet. This form of meronymy corresponds to the semantics of territories being part of countries. The glosses of those verbs indicate that this meronymy is rather viewed as a form of possession, which due to the verb *control* from the gloss of *territory* shows preference to the more different relation of governing or ruling of territories by their countries. It is to be noted that Greenland is encoded in WordNet, its gloss *"a self-governing province of Denmark"* would have led to the same justification as the one determined by the text snippet retrieved from the AQUAINT corpus, creating a contradiction between the concepts *self-governing* and *control by possession.* Currently, the abductive processes implemented in the COGEX theorem prover (Moldovan *et al.* 2003) do not handle such contradictions. If the relation of provinces/territories belonging to countries would have been encoded as meronymy, then Greenland should have been encoded as part of Denmark in WordNet. (Currently, it is encoded as a part of the Atlantic Ocean)). The latter two meronyms show preference for geographical membership rather than country/nation possession.

For question Q2217, the absence of the abductive justification would have produced *Ethiopia* as the answer, because of the text snippet *"the high ice desert of Greenland and the tributaries of the Blue Nile in Ethiopia"*.

Extracting answers for definition questions

Our QA system extracted 485 answers in response to the 50 definition questions evaluated in TREC-2003. We submitted two runs, one of which consisted of exact answers and the other of the corresponding sentence-type answers. Out of 485 answers, the assessors have found a total of 68 (exact) and 86 (sentence) vital matches from a total of 207 they expected and a total of 110 (exact) and 144 (sentence) matches out of 417 they had in their final set. The definition questions evaluated in TREC-2003 can be classified in:

- questions asking about people;
- questions asking about other types of names; and
- questions asking about general concepts.

The questions asking about people started with the question stem *Who* and contained the name of a person. There were 30 such questions, 22 of which had the person name in the format *first name - last name*, e.g. *Aaron Copland*, *Allen Iverson* or *Albert Ghiorso.* One question had the name in the format *first name - last name1 - last name 2*, i.e. *Antonia Coello Novello.* Three questions had the name as a single word, signifying that they are very well known: *Nostradamus*, *Absalom* and *Abraham.* In the latter case, the context was also specified: *"Abraham in the Old Testament"*. Two other person names were names of old kings or princes: *Vlad the Impaler* and *Akbar the Great*, having the format *first name - the attribute.*

There were 14 questions asking about other types of names. Four asked about different organizations, e.g. *Bausch & Lomb*, *ETA*, *Friends of the Earth* or *Destiny's Child.* Two asked about cities, e.g. *the Hague* but also nicknames of cities, e.g. *Bollywood.* Two asked about medical or biology terms, e.g. *TB* or *Ph.* Three asked about words in foreign languages: e.g. *Iqra* in Arabic, *Schadenfreude* in German, and *Kama Sutra* in Sanskrit. Six definition questions asked about general concepts, e.g. *fractals*, *golden parachute* or *quasar.*

To produce answers for definition questions, our system uses 38 patterns. Out of these, 23 patterns had at least a match for the tested TREC questions. Table 5 illustrates the most popular patterns. Figure 4 illus-

Cecilia Bartoli broke her right ankle slipping on ice outside the Zurich Opera but still intends to sing her first Donna Elvira in Mozart 's "Don Giovanni" this Sunday
Sally Wolf gives a movingly tormented performance as Donna Elvira and is at her best in the great spleen - venting aria "i tradi quell" alma ingrata
Donna Elvira's cold fury seemed to emanate as much from the natural personality of the singer (*Veronique Gens*) as from the nature of the role
In the title role, the resonant bass Ferruccio Furlanetto led a strong cast that included the powerful bass Rene Pape as, surprisingly, a vocally agile Leporello; the exquisite soprano Renee Fleming as a Donna Anna to cherish; and the luminous - voiced soprano *Marina Mescheriakova* as Donna Elvira
A newcomer, the Norwegian soprano *Solveig Kringelborn*, sang Donna Elvira with a clean intensity
Lott is celebrated for her Mozart roles (and recorded her Fiordiligi in "Cosi fan tutte," the Countess in "The Marriage of Figaro," and Donna Elvira in "Don Giovanni"; earlier this season the Met broadcast her elegant portrayal of the Countess) and for her Strauss (she has taken her Arabella , Marschallin , Countess Madeleine , and *Christine Storch* to most of the major opera houses of the world)

Table 6: Answers for question Q2002

trates the number of answers extracted through pattern matching for each of the 50 definition questions.

Answering list questions

To answer the 37 list questions evaluated in TREC-2003, our QA system considered a threshold-based cut-off of the answers extracted. The general idea was that by using concept similarities between the candidate answers we could decide on the threshold value for submitting the elements of the list. Given that for a question we extract N list answers, we first compute the similarity between the first answer and the last answer, S_{1N}. In general, to compute the similarity between a pair of answers (A_i, A_j) we consider a window of three noun or verb concepts to the left and to the right of the exact answer: $W_i = (C^i_{-3}, C^i_{-2}, C^i_{-1}, C^i_{+1}, C^i_{+2}, C^i_{+3})$ and $W_j = (C^j_{-3}, C^j_{-2}, C^j_{-1}, C^j_{+1}, C^j_{+2}, C^j_{+3})$. Then we separate the concepts in nouns and verbs obtaining N_i, N_j, V_i, and V_j. The similarity is measured by the formula: $sim(A_i, A_j) = \frac{1}{2}(sim^N(N_i, N_j) + sim^V(V_i, V_j))$, where $sim^N(N_i, N_j) = \frac{1}{P_{(N_i,N_j)}} \sum sim^C(n_i, n_j)$, with $P_{(N_i,N_j)}$ representing all the possible pairs (n_i, n_j) in which $n_i \in N_i$ and $n_j \in N_j$; and $sim^V(V_i, V_j) = \frac{1}{P_{(V_i,V_j)}} \sum sim^C(v_i, v_j)$, with $P_{(V_i,V_j)}$ representing all the possible pairs (v_i, v_j) in which $v_i \in V_i$ and $v_j \in V_j$.

The concept-based similarity is computed as: $sim(n_i, n_j) = argmax \;\; sim(c_i, c_j)$, where (c_i, c_j) are all possible combinations of the WordNet senses of n_i and n_j, and

$$sim(c_i, c_j) = \begin{cases} 1, & if \;\; c_i = c_j \\ 0, & if \;\; c_i \;\; and \;\; c_j \;\; do \;\; not \;\; belong \\ & \quad to \;\; the \;\; same \;\; hierarchy \\ sim_{LC}(c_i, c_j), & otherwise \end{cases}$$

where $sim_{LC}(c_i, c_j)$ is the Leacock-Chodorow similarity (Leacock & Chodorow 1998) defined as:

$$sim_{LC}(c_i, c_j) = \log \frac{len(c_i, c_j)}{2D}$$

where $len(c_i, c_j)$ is the shortest path between c_i and c_j and D is the overall depth of the WordNet taxonomy.

A threshold value of $Z = C \times S_{1N}$ is computed using the similarity between the first and the last concept answer multiplied with a constant C. The cut-off is determined as the largest value t that satisfies: $\frac{1}{t} \sum_{i=1}^{t} S_{1i} > Z$. When the cutoff is determined, it represents the length of the list of answers that is submitted.

The best precision and recall was obtained for Q2002: *"Name singers performing the role of Donna Elvira in performances of Mozart's "Don Giovanni""*. Table 6 lists the answers extracted for this question. There were 5 correct answers out of 6 submitted, corresponding to a precision of 0.833 and a recall of 0.625; the combined F-score was 0.714.

Performance evaluation

Table 7 summarizes the scores provided by NIST for our system. We have submitted two different runs. They differ only in the way definition answers were extracted. In the first submission, only exact answers were extracted whereas in the second the whole sentence containing the answer was submitted.

Table 7 illustrates the contribution of the factoid, list and definition components to the overall scores of the main tasks.

	factoid	list	definition	all
Main task submission 1	70.0%	39.2%	36.1%	53.8%
Main task submission 2	70.0%	39.6%	44.2%	55.9%
Passage task	68.5%	N/A	N/A	N/A

Table 7: Results in TREC-2003 evaluations

The score of the second submission was slightly higher than that of the first submission because of the better score obtained for the definition questions, which in this case were in the format of an entire sentence. This allowed more vital nuggets to be identified by the assessors, thus obtaining a better score. Another important observation stems from the fact that factoid questions in the main task were slightly better evaluated than in the passage task. We explain this fact by our belief that the passage might have contained multiple concepts similar to the answer, and thus produced a more vague evaluation context.

Acknowledgments

We would like to thank Drs. Mihai Surdeanu and Marius Paşca for their valuable contributions. This work, as well as our broader research on QA, is supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program.

References

Leacock, C., and Chodorow, M. 1998. *WordNet: An Electronic Lexical Database and Some of its Applications.* MIT Press. chapter Combining local context and WordNet similarity for word sense identification, 265–283.

Moldovan, D.; Clark, C.; Harabagiu, S.; and Maiorano, S. 2003. Cogex: A logic prover for question answering. In *Proceedings of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference (HLT-2003)*, 87–93.

Riloff, E., and Jones, R. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, 474–479.