Natural Language Engineering **11** (4): 343–361. © 2005 Cambridge University Press
doi:10.1017/S1351324905003682 Printed in the United Kingdom

Implementing clarification dialogues in open domain question answering

MARCO DE BONI
School of Computing, Leeds Metropolitan University, Leeds LS2 3QS, UK
e-mail: m.deboni@leedsmet.ac.uk

SURESH MANANDHAR
Department of Computer Science, University of York, York YO10 5DD, UK
e-mail: suresh@cs.york.ac.uk

(*Received 15 August 2003; revised 2 March 2004*)

Abstract

We examine the implementation of clarification dialogues, a mechanism for ensuring that question answering systems take into account user goals by allowing them to ask series of related questions either by refining or expanding on previous questions with follow-up questions, in the context of open domain Question Answering systems. We develop an algorithm for clarification dialogue recognition through the analysis of collected data on clarification dialogues and examine the importance of clarification dialogue recognition for question answering. The algorithm is evaluated and shown to successfully recognize the start and continuation of clarification dialogues in 94% of cases. We then show the usefulness of the algorithm by demonstrating how the recognition of clarification dialogues can simplify the task of answer retrieval.

1 Clarification dialogues in question answering

Open domain Question Answering (QA) Systems, as defined for example in the TREC evaluation framework which currently provides a widely accepted standard baseline for research in this area (see Voorhees (2002) for an outline of the framework and an overview of current systems), aim to determine an answer to a question (which is not limited to a particular task or topic) by searching for a response in a collection of documents such as newspaper articles. To achieve this (see, for example, Harabagiu, Moldovan, Pasca, Surdeanu, Mihalcea, Gîrju, Rus, Lacatusu, Morarescu and Bunescu 2002), systems narrow down the search by using information retrieval techniques to select a subset of documents, or paragraphs within documents, containing keywords from the question and a concept which corresponds to the correct question type (e.g. a question starting with the word "Who?" would require an answer containing a person). The exact answer sentence is then sought by either attempting to unify the answer semantically with the question, through some kind of logical transformation (e.g. Moldovan and Rus 2001) or by some form of pattern matching (e.g. Soubbotin 2002; Harabagiu *et al.* 1999).

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

Often, though, a single question is not enough to meet users' goals: a wider dialogue (which, in the case of TREC-style QA systems is limited to a series of question/answer pairs, and not, as happens in human dialogue, also question/question pairs), either elaborating and building on information gathered, or clarifying previously given information is required, i.e. a dialogue which will enable the users to fully achieve their informational goals. We shall hence refer to such exchanges as clarification dialogues, following the terminology used for example by Purver, Ginzburg and Healey (2002), Purver *et al.* (2003) and Ginzburg (1998), as the questions that constitute them either clarify previous questions or answers or clarify the mental picture the user is trying to build by elaborating on previously asked or given information: the expression "clarification dialogue" indicates that we are in fact a) examining a dialogue, albeit a very limited one, where only one party in the dialogue asks questions and only one party gives answers; and b) we are considering a dialogue which clarifies some concept in the questioner's mind, whether this be by asking for some new information related to the topic investigated or asking for an explanation of something already given.

One example of a clarification dialogue is in the form of questions that seek to clarify the meaning of an answer, for example when the user has not understood a term contained in the answer, as in the following exchange:

(1) Q_1: What is a fairy tale?
A_1: The American Heritage dictionary tells me it is a fanciful tale of legendary deeds and creatures.
Q_2: What does fanciful mean?
A_2: . . .

On other occasions users want to expand on a given answer in order to have more details, as in the following example, where the user, having discovered a need (the necessity to have a license to fish) wants more details about how to go about fulfilling that need (the cost of the license):

(2) Q_1: Do I need a license to fish in the Tiber river?
A_1: Yes.
Q_2: How much?
A_2: . . .

In other cases the user's goal is to form a broad picture about some topic and a number of separate questions are needed in order to achieve the breadth of information required:

(3) Q_1: Where was Frank Sinatra born?
A_1: Hoboken, N.J.
Q_2: Where did he grow up?
A_2: Hoboken, N.J.
Q_3: What kind of childhood did he have?
A_3: . . .

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

The common link between the above dialogue fragments is the fact that the question/answer sequences form coherent units of discourse quite different from an interaction such as the following:

(4) Q: What is caffeine?
A: A stimulant.
Q: What imaginary line is halfway between the North and South Poles?
A: The equator.
Q: Where is John Wayne airport?
A: . . .

In (4) there is no relationship between the questions or between the questions and previous answers and hence in seeking an answer there is no immediate reason to take into consideration previously asked questions or previously given answers. In fragments (1) to (3), however, in order to answer the questions correctly it is necessary to take into consideration the previous context in order to satisfy the user's goals. In (1), for example, the user isn't asking for the generic meaning of the word fanciful (the American Heritage Dictionary, for example, gives three separate meanings for the word fanciful) but the specific meaning that word takes in the sentence "it is a fanciful tale of legendary deeds and creatures". Similarly in (2) the question "How much?" makes no sense, and cannot be answered, without reference to the context. Exchanges such as those in examples (1) to (3) therefore have in common the feature that to answer a question satisfactorily some reference must be made to previously asked questions and previously given answers. While a number of researchers have looked at clarification dialogue from a theoretical point of view (e.g. Purver *et al.* 2002, 2003; Ginzburg 1998; Ginzburg and Sag 2000; van Beek, Cohen and Schmidt 1993), or from the point of view of task oriented dialogue within a narrow domain (e.g. Ardissono and Sestero 1996), there has been little work on clarification dialogue for *open domain* question answering systems such as the ones presented at the TREC workshops, where the task that the user is pursuing and the subject matter of the user's investigations are not known *a priori*. Initial work in this direction has consisted of a series of experiments carried out for the (subsequently abandoned) "context" task in the TREC-10 QA workshop (Voorhees 2002; Harabagiu *et al.* 2002) and the initial experiments presented by De Boni and Manandhar (2003b). Here we seek to partially address this problem by looking at a particular aspect of clarification dialogues in the context of open domain question answering: the problem of recognizing that a clarification dialogue is occurring, i.e. how to decide whether the current question is part of an on-going series (i.e. clarifying previous questions) or the start of a new series; we then show how the recognition that a clarification dialogue is occurring can simplify the problem of answer retrieval.

2 The TREC content experiments

The TREC-2001 QA track included a "context" task which aimed at testing systems' ability to track context through a series of questions (Voorhees 2002). In other words,

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

systems were required to respond correctly to a kind of clarification dialogue in which (according to the intentions of the organizers) questions could only be answered satisfactorily if previous questions had been interpreted correctly. In order to test the ability to answer such questions correctly, a total of 42 questions were prepared by NIST staff, divided into 10 series of related question sentences which therefore constituted a type of clarification dialogue; the dialogues varied in length between three and eight questions, with an average of four questions per dialogue. These clarification dialogues were however presented to the question answering systems already classified and hence systems did not need to recognize that clarification was actually taking place. Consequently systems that simply looked for an answer in the subset of documents retrieved for the first question in a series performed well without any understanding of the fact that the questions constituted a coherent series.

In a more realistic approach, systems would not be informed in advance of the start and end of a series of clarification questions and would not be able to use this information to limit the subset of documents in which an answer is to be sought.

3 Analysis of the TREC context questions

We manually analyzed the TREC context question collection to determine what features could be used to determine whether a question was part of a longer question series (see De Boni and Manandhar 2003b), with the following conclusions:

1. Pronouns and possessive adjectives. For example:

   –What does transgenic mean?
   –What was the first transgenic mammal?
   –*When was it born?*

where a question are referring to some previously mentioned object through a pronoun (“it”). The use of personal pronouns (“he”, “it”, . . .) and possessive adjectives (“his”, “her”, . . .) which did not have any referent in the question under consideration was therefore considered an indication of a clarification question. Notice there is no need to use any form of coreference resolution to classify these questions as being part of a wider series.

2. Ellipsis, as in:

   –What type of vessel was the modern Varyag?
   –. . .
   –In what country is this facility located?
   –*On what body of water?*
   –How long is the Varyag?
   –*How wide?*

where the incomplete syntactical construction is an indication that the question referred to some previous question or answer.

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

3. semantic relations between words in question series, as in the following:
   –Which museum in Florence was damaged by a major bomb explosion?
   –*Which galleries were involved?*
   –*How many people were killed?*
   –. . .
   –*How much explosive was used?*

where there is a semantic relation between "museum" and "galleries", and between "explosion", "killing" and "explosive". Questions belonging to a series were "about" the same subject, and this aboutness could be seen in the use of semantically related words. A particular case of semantic relation between words was the repetition of proper nouns, as in:

–What type of vessel was the modern Varyag?
–. . .
–How many aircraft was it designed to carry?
–*How long was the Varyag?*

where the repetition of the proper noun indicates that the same subject matter is under investigation.

4 Experiments in clarification dialogue recognition

We speculated that an algorithm which made use of these features would successfully recognize the occurrence of clarification dialogue. In order to verify this hypothesis we collected two sets of new data on which to test the algorithm: we made use of the first set to carry out an initial evaluation; following this initial evaluation changes were made to the algorithm and therefore a second set of data was necessary in order to test the changes. The collected questions were fed into a system implementing the algorithm, which attempted to recognize the occurrence of a clarification dialogue; the results given by the system were then compared to the previously tagged clarification dialogues. We then conducted a number of experiments to verify the usefulness of clarification dialogue recognition in improving answer retrieval performance in a question answering system.

5 Collection of new data

Data collection was carried out in two stages: a first collection which aimed at testing the algorithm and understanding problems associated both with the algorithm and the collection of the dialogue data itself; a second collection which improved the data collection process in light of the problems noted in collecting the first data and was used to test any modifications of the algorithm made in light of the first collection.

5.1 Dialogue Collection A

Given that the only available data was the collection of "context" questions used in TREC-10, it was felt necessary to collect further data in order to test our algorithm

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

rigorously. This was necessary both because of the small number of questions in the TREC data and the fact that there was no guarantee that an algorithm built for this dataset would perform well on "real" user questions. A collection of 253 questions was therefore put together using the method described below.

A number of questions on a wide range of topics were chosen from the TREC collection used for the TREC open domain question answering track (not the context question collection) and used as a collection of possible dialogue topics. 24 users were then invited to interact with a question answering system using the "Wizard of Oz" method (see Preece, Rogers, Sharp, Benyon, Holland and Carey (1994) for a description of this methodology). First they were required to choose a topic from the given collection of topics. They were then asked to seek information on the given topic; no details were given as to how much information or what type of information was to be collected and users were told it was up to them what information and how much information was to be gathered. As an "ice-breaker" they were then invited to ask the QA system the exact question on the topic as given in the TREC collection, after which they were free to interact with the system as they felt appropriate. In order to ensure a realistic interaction, which was not dependant on the performance of the Question Answering system itself, the questions were answered by an operator using the Google search engine to find relevant answers. These questions thus collected made up 24 clarification dialogues, varying in length from 3 questions to 23, with an average length of 12 questions (the data is available from the main author upon request). A typical interaction would therefore proceed as follows:

> Question topic chosen: Philadelphia.
> Initial "ice-breaker" question (n. 1526 in the TREC-11 question series): "What is the city of brotherly love?"
> System: "Philadelphia"

At this point the user was free to ask any question to the system and proceeded with:

> User: "Where is Philadelphia" [sic]
> System: "USA"
> User: "Where more precisely?"
> etc.

Unlike the TREC collection, the dialogues we gathered highlighted the tendency of users to make use of cue-words such as "exactly" or "precisely" (as in "where exactly?") to clarify previously asked questions. Other than this there did not appear to be any significant difference in the type of questions asked in the TREC collection and in the collection we gathered (apart from one single user who attempted to have a lifelike dialogue with the system with some very intricate questions). The average length of a dialogue interaction was significantly higher in our collection (an average of 12 questions per dialogue as opposed to an average of 4) with, as would be expected, a wider range of extremes (the shortest dialogue in our collection was of 3 questions, the longest 23, as opposed to 4 and 8 in the TREC context questions).

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

Table 1. *Differences between TREC "context" collection and the new collections*

	Groups	Qs	Av. len	Max	Min
TREC context	10	41	4	8	4
Collection A	24	253	11	23	3
Collection B	16	144	9	17	5

5.2 Dialogue Collection B

It was noted that the method used to collect dialogue data was lacking in rigour due to:

- not having given the users a specific task to accomplish: without a specific goal in mind it was difficult to compare the dialogues
- not having a well-defined procedure for deciding what answer to give the user: follow-up questions depended on the type of answer given to the previous question

More experimental data was therefore gathered. In this case, users were given the task of collecting enough information from the system in order to then be able to write a short paragraph on a given topic which they had no extensive information about beforehand (they were asked if they had any in-depth information about a topic before starting the experiment). They then interacted with the QA system using the wizard-of-Oz methodology in order to gather information. The operator of the QA system consulted the Google search engine for an answer by using the question in its entirety as a query and looking in the retrieved documents in the precise order in which they were given for the sentence snippet which answered the question in the most concise manner; this then constituted the answer which was given by the system.

A total of 16 dialogues were collected: question length was much more consistent than collection A, with an average length of 9 questions per dialogue, a maximum of 17 and a minimum of 5. There appeared to be no significant difference in the type of questions asked compared to the previously collected dialogues.

The differences between the TREC "context" collection and the new collections are summarized in Table 1.

6 Clarification recognition algorithm

Our approach to clarification dialogue recognition looks at certain features of the question currently under consideration (e.g. pronouns and proper nouns) and compares the semantic properties (which could, in a very loose sense, be said to make up the "meaning") of the current question with the semantic properties of previous questions to determine whether they are "about" the same matter. Given:

- a question q_i
- a question window n which determines how far back a question can refer within the clarification dialogue sequence
- n previously asked questions $q_{i-1}..q_{i-n}$

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

we have a function Clarification_Question which is true if a question is considered a clarification of a previously asked question. Empirical work such as Ginzburg (1998) and Purver *et al.* (2002) indicates that questioners do not usually refer back to questions which are very distant, and this was consistent with our data. In particular Purver *et al.* analyzed the English dialogue transcripts of the British National Corpus, finding that clarification request source separation (CSS, the distance between a question and the question or answer which it is attempting to clarify) was at most 15 sentences and usually less than 10 sentences. We therefore set the question window to be the average length of the clarification dialogues in the two sets of data, and hence considered the set of the previously mentioned eight questions, i.e. set the question window $n = 8$. This is consistent with the empirical observations of Purver *et al.* as our maximum distance of 8 questions is equivalent to a CSS of 16 sentences (i.e. eight pairs of questions and answers).

A question is deemed to be a clarification of a previous question if:

1. There are direct references to nouns mentioned in the previous n questions through the use of pronouns (he, she, it, . . .) or possessive adjectives (his, her, its . . .) which have no references in the current question; this was altered after the experiments carried out on the first sample of collected data to also include cue-words such as "precisely", "exactly" etc. clearly indicating a reference to a previous question or answer.
2. The question does not contain a verb phrase.
3. There are explicit references to proper and common nouns mentioned in the previous n questions; or there is a strong sentence similarity between the current question and the previously asked questions. In other words:

Clarification_Question
(q_i, $q_{i-1}..q_{i-n}$)
is true if
1. q_i has pronoun and possessive adjective references
to $q_{i-1}..q_{i-n}$
2. q_i does not contain any verbs
3. q_i has repetition of common or proper nouns in $q_{i-1}..q_{i-n}$ or
q_i has a strong semantic similarity to some $q \in q_{i-1}..q_{i-n}$

This basic algorithm was improved by using a sigmoid function to simulate decaying importance of previous questions in order to avoid a sharp step. The algorithm above can be considered a step function with an abrupt cut-off point, which makes a binary decision on the possible relevance of previous questions to the current question. In particular, a question which is too far back to be included in the question window n will be ignored, no matter how strongly it is related to the current question; furthermore, no differentiation is made between recently asked questions and questions which are significantly more distant in time as long as they are within the allowed window: hence, when looking at similarities between previous questions and the current question, there is no distinction between the question, which immediately preceded the current one and a question which was uttered 6 moves ago in the dialogue. What we really want is a stronger similarity

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

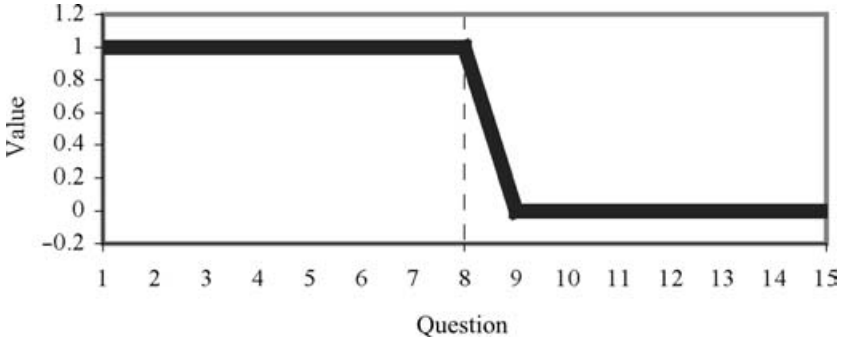


Fig. 1. Initial function used.

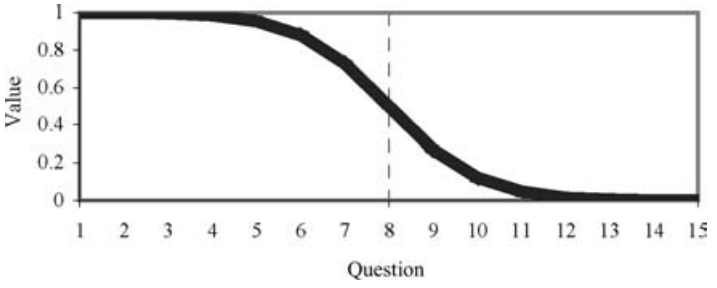


Fig. 2. Function used instead of the step function.

the further away a question is: we are therefore looking for a scaling factor which ensures that in the limit weak similarities between very close questions are more important than strong similarities between very distant questions.

Figure 1 shows the initial function used in the algorithm above, a step function which ignores any questions outside the given window of 8 questions. Figure 2 shows the function used instead of the step function, derived from a basic sigmoid function[1]: this is much more satisfactory as it ensures that questions outside the given window are not simply ignored but considered relevant with a lower degree of probability.

The algorithm above was therefore modified so that when looking at similarities between current and past questions the distance between the questions was also taken into account. In particular the similarity measure was weighted according to formula (a). This ensured that as the distance between questions increased the

[1] Given the basic sigmoid function ($y = 1/(1 + \exp(-x))$) we are seeking a function which rapidly decays once it has reached the given question window n. The sought-after function is therefore:

$$y = 1 - \frac{1}{1 + e^{-(x-n)}}$$

which in our case becomes (fig. 2):

$$y = 1 - \frac{1}{1 + e^{-(x-8)}}$$

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

similarity between the questions had to increase in order to be considered relevant, i.e. part of the same dialogue. We therefore have:

3.a q_i has repetition of common or proper nouns in $q \in q_{i-1} .. q_{i-n}$ or q_i has a strong semantic similarity to some $q \in q_{i-1} .. q_{i-n}$, *weighted according to the distance between* q_0 *and* q *according to the formula* similarity=similarity*(1−(1/(1+exp(−(m−8))))) *where* m *indicates the number of questions between* q_i *and* q

Similarity between questions is therefore measured as a decaying function, allowing for references to questions far back in the dialogue, but without running into the danger of overplaying similarities which, due to the distance between the questions, should not be considered significant.

7 Sentence similarity metric

A major part of our clarification dialogue recognition algorithm is the sentence similarity metric, which looks at the similarity in meaning between the current question and previous questions. WordNet (Miller 1999; Fellbaum 1998), a lexical database which organizes words into synsets, sets of synonymous words, and specifies a number of relationships such as hypernym, synonym, meronym which can exist between the synsets in the lexicon, has been shown to be fruitful in the calculation of semantic similarity. One approach has been to determine similarity by calculating the length of the path or relations connecting the words which constitute sentences (see, for example, Green (1997) and Hirst and St-Onge (1998)); different approaches have been proposed either using all WordNet relations (Budanitsky and Hirst 2001) or only is-a relations (Resnik 1995; Jiang and Conrath 1997; Mihalcea and Moldvoan 1999) (for an evaluation see Budanitsky and Hirst 2001). Miller (1999), Harabagiu *et al.* (2002) and De Boni and Manandhar (2002) found WordNet glosses, considered as micro-contexts, to be useful in determining conceptual similarity. Our sentence similarity measure followed on these ideas, adding to the use of WordNet relations, part-of-speech information, compound noun, proper noun and word frequency information, based on the algorithm examined in De Boni and Manandhar (2003), to which we refer the reader for a more detailed discussion of the individual components and experimental data which justifies their use.

In particular, sentence similarity was considered as a function which took as arguments a sentence s and a second sentence t and returned a value representing how closely related in meaning s was with respect to t in the context of some background knowledge B, i.e.

$$\text{Sentence-similarity}(s, t, B) = n$$

Sentence-similarity $(s_1, t, B) <$ Sentence-similarity(s_2, t, B) represents the fact that sentence s_1 is not as closely related in meaning as s_2 in respect to the sentence t and the context B. In our experiments, B was taken to be the full set of semantic relations given by WordNet. Clearly, the use of a different knowledge base would give different results, depending on its completeness and correctness.

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

To calculate the semantic similarity between a sentence s_1 and another sentence s_2, s_1 and s_2 were considered as sets P and Q of word lemmas. The similarity between each word in the two sentences was then calculated and the sum of the closest matches gave the overall similarity. In other words, given two sets Q and P representing the current question and a previous question, where $Q = \{qw_1, qw_2, \ldots, qw_n\}$ and $P = \{pw_1, pw_2, \ldots, pw_m\}$, the similarity between Q and P is given by

$$\sum_{1<p<n} \text{Argmax}_m \text{ similarity}(qw_p, pw_m)$$

Note that the asymmetry of the function was not an issue as similar results were obtained inverting the order of the arguments. The function $similarity(w_1, w_2)$ maps the lemmas of the two words w_1 and w_2 to a similarity measure m representing how semantically related the two words are; $similarity(w_i, w_j) < similarity(w_i, w_k)$ represents the fact that the word w_j is less semantically related than w_k in respect to the word w_i. In particular $similarity = 0$ if two words are not at all semantically related or if $w_1 \in \text{ST} \vee w_2 \in \text{ST}$, where ST is a set containing 100 stop-words (e.g. "the", "a", "to") which were judged to be too common to be able to be usefully employed to estimate semantic similarity. On the other hand, $similarity = 1$ if the words are the same. In all other cases $similarity(w_1, w_2) = h$, where h is calculated by comparing the words w_1 and w_2 using all the available WordNet relationships (is-a, satellite, similar, pertains, meronym, entails, etc.). Each relationship is given a weighting indicating how related two words are, with a "same as" relationship indicating the closest relationship, followed by synonym relationships, hypernym, hyponym, then satellite, meronym, pertains, entails.

So, for example, given the question "Who went to the mountains yesterday?" and the second question "Did Fred walk to the big mountain and then to mount Pleasant?", Q would be the set {who, go, to, the, mountain, yesterday} and P would be the set {Did, Fred, walk, to, the, big, mountain, and, then, to, mount, Pleasant}.

To calculate similarity the algorithm would consider each word in turn. "Who" would be ignored as it is a common word and hence part of the list of stop-words. "Go" would be related to "walk" in a is-a relationship and receive a score h_1. "To" and "the" would be found in the list of stop-words and ignored. "Mountain" would be considered most similar to "mountain" (same-as relationship) and receive a score h_2: "mount" would be in a synonym relationship with "mountain" and give a lower score, so it is ignored. "Yesterday" would receive a score of 0 as there are no semantically related words in Q. The similarity measure of Q in respect to P would therefore be given by $h_1 + h_2$.

Additional information which has been shown to improve performance of the similarity measure (see De Boni and Manandhar (2003) for a complete discussion) was then considered, in particular:

- *Compound noun information.* The motivation behind this is similar to the reason for using chunking information, i.e. the fact that the word "status" in "status quo" should not be considered similar to "status" as in "status symbol". Compound nouns were identified using WordNet.

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

- *Proper noun information.* The intuition behind this is that titles (of books, films, etc.) should not be confused with the "normal" use of the same words: "blue lagoon" as in the sentence "the film Blue Lagoon was rather strange" should not be considered as similar to the same words in the sentence "they swam in the blue lagoon" as they are to the sentence "I enjoyed Blue Lagoon when I was younger". A simple set of rules seeking sequences of words which were not at the beginning of a sentence and which started with capitalized letters was found to be sufficient to identify proper nouns.
- *Word frequency information.* This is a step beyond the use of stop-words, following the intuition that the more a word is common the less it is useful in determining similarity between sentences. So, given the sentences "metatheoretical reasoning is common in philosophy" and "metatheoretical arguments are common in philosophy", the word "metatheoretical" should be considered more important in determining relevance than the words "common", "philosophy" and "is" as it is much more rare and therefore less probably found in irrelevant sentences. Word frequency data was taken to be the word frequency given in the British National Corpus (see BNCFreq 2003). The top 100 words, making up 43% of the English Language, were then used as stop-words and were not used in calculating semantic similarity.

8 Results

An implementation of the algorithm was evaluated on the TREC context questions used to develop the algorithm and then on the collection of new clarification dialogue questions. Initially, the individual components of the algorithm were evaluated separately to gauge their contribution to the overall performance of the algorithm. The overall algorithm was then evaluated. The evaluation consisted in testing the algorithm's ability to:

- recognize the start of a new series of questions (indicated by N, for "New", in the results table);
- recognize that the current question is clarifying a previous question (indicated by C, for "Clarification", in the table).

We calculated N as the percentage of all questions which the algorithm correctly recognized as being the first in a new series, while C indicated the percentage of questions which were correctly recognized as being clarification questions. In order to understand the overall performance of the algorithm an F-measure or weighted harmonic mean (van Rijsbergen 1979) is used, based on the formula

$$\mathrm{F-measure} = \frac{2NC}{N+C}$$

The individual components were evaluated as follows:

- *Common Words (CW).* This was the baseline method which did not use any linguistic information and simply took a question to be a clarification question if it had any words in common with the previous n questions (excepting the list of stop words which was also used for the semantic similarity algorithm), else took the question to be the beginning of a new series.

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

- *Reference Words (RW)*. This method employed point 1 of the algorithm described in section 6 by looking for "reference" keywords such as "he", "she", "this", "so", etc. which clearly referred to previous questions. This did not misclassify any "new" questions.
- *Absence of Verbs (AV) + RW*. This method employed points 1 and 2 of the algorithm described in section 6 by looking for the absence of verbs combined with keyword lookup.
- *Noun Similarity (NS1) + AV + RW*. This method implemented the algorithm described in section 6 by looking at the similarity between nouns in the current question and nouns in the previous questions, in addition to reference words and the absence of verbs. Note that the data was manually checked to identify and correct gross errors in compound noun identification (compound noun identification was carried out following the procedure outlined in De Boni and Manandhar 2002) which were noted in initial experiments (which were responsible for degrading performance in the identification of new sequences of collection A to 67%).
- *Noun Similarity (NS2) + AV + RW*. This differed from the previous method in that it specified a similarity threshold when employing the similarity measure.
- *Decaying Function (DF) + PS + NS2 + AV + RW*. This implemented the full algorithm, with the use of a decaying function to give a weighting to the similarities between questions in a sequence (i.e. using point 3.a above as opposed to 3). Moreover this method employed a similarity threshold which was experimentally set for best performance based on an analysis of the results of the experiments carried out on the TREC data.
- An examination of the results of experiments carried out on test Collection A with the features examined above, indicated that it was also necessary to consider Answer Similarity (ANS) (this feature was not evident in the TREC data collection due to the way the data was put together), for instance clarifying the meaning of a word contained in the answer, or building upon a concept defined in the answer. An example was the question "What did Antonio Carlos Tobim play?" following "Which famous musicians did he play with?" in the context of a series of questions about Frank Sinatra: Antonio Carlos Tobim was referred to in the answer to the previous question, and nowhere else in the exchange. ANS indicated a strong semantic relationship between the current question and the answer given immediately before the question was asked. In the gathered data questions referred to the immediately preceding answer, and not to answers within a given "window"; consequently, our algorithm only considered the immediately preceding answer. The fact that this feature was identified following an analysis the experiments carried out on Collection A made it necessary to gather and experiment on Collection B to verify the strength of the algorithm.

The results on the TREC data, which was used to develop the algorithm, are summarized in Table 2. The results for the same experiments conducted on the collected data were as shown in Table 3.

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

Table 2. *Results on TREC data*

TREC	CW	RW	RW + AV	RW + AV + NS1	RW + AV + NS2	RW + AV + NS2 + DF
N	90%	90%	90%	60%	80%	90%
C	47%	53%	59%	78%	72%	77%
F-measure	62%	67%	71%	68%	76%	83%

Table 3. *Results on collected data*

Collected A	CW	RW	RW + AV	RW + AV + NS1	RW + AV + NS2	RW + AV + NS2 + DF	RW + AV + NS2 + DF + ANS
N	100%	100%	100%	71%	87%	96%	96%
C	64%	62%	66%	91%	89%	93%	96%
F-measure	78%	77%	80%	80%	88%	94%	96%

Table 4. *Same experiments on Dialogue collection B*

	Collected B	*Full Algorithm*
	N	93%
	C	96%
	F-measure	94%

Table 5. *Results on Dialogue collections A and B*

	No. of permutations	Average F-measure	Standard Deviation
Collection A	6	95.9	0.84
Collection B	6	94.3	0.98

The same experiments, using all the features available to the full algorithm, were then carried out on the Dialogue Collection B; the only modification to the algorithm was the addition of cue words to point 1 of the algorithm. Results were similar to the results for Collection A, as can be seen in Table 4.

In the experiments above the data was provided to the algorithm in the order in which it was collected. To verify how the particular order in which the questions were given influenced the results another series of experiments were carried out in which the dialogues from Collection A and Collection B were reshuffled and fed into the full algorithm. Results are given in Table 5.

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

As can be seen, the particular order in which the dialogues were given had a negligible effect on the results. Problems noted were:

- False positives: questions following a similar but unrelated question series. For example, "Are they all Muslim countries?" (talking about religion, but in the context of a general conversation about Saudi Arabia) followed by "What is the chief religion in Peru?" (also about religion, but in a totally unrelated context).
- Absence of relationships in WordNet, e.g. between "NASDAQ" and "index" (as in share index). Absence of verb-noun relationships in WordNet, e.g. between "to die" and "death", between "battle" and "win" (i.e. after a battle one side generally wins and another side loses), "airport" and "visit" (i.e. people who are visiting another country use an airport to get there).

As can be seen from Tables 2–5, the same experiments conducted on the TREC context questions yielded slightly worse results; a failure analysis revealed this was mostly due to the inability to find semantic relationships in WordNet between words in the domain of two of the ten questions: explosives (explosion – explosives – bomb) and wine growing (winery – grape). The small sample size meant that errors in only two of the questions significantly affected the results for the overall performance (the errors in the two questions were in fact responsible for 71% of the mistakes made by the system).

The performance of the individual components of the algorithm followed a similar pattern in both the TREC collection and collection A, with the individual components cumulatively contributing to increasing ability to recognize clarification or the start of a new dialogue, which confirmed the usefulness of the components. The results on Collection B confirmed the strength of the algorithm even with a slightly different type of dialogue setting.

9 Usefulness of clarification dialogue recognition

Recognizing that a clarification dialogue is occurring only makes sense if this information can then be used to improve answer retrieval performance. We hypothesized that clarification dialogue recognition would in fact enable us to simplify the answer retrieval process (and hence improve performance) by adding constraints to what the answer could be. Noting that a questioner is trying to clarify previously asked questions is in fact important in order to determine the context in which an answer is to be sought: answers to certain questions are constrained by the context in which they have been uttered. The question "What does attenuate mean?", for example, may require a generic answer outlining all the possible meanings of "attenuate" if asked in isolation, or a particular meaning if asked after the word has been seen in an answer (i.e. in a definite context which constrains its meaning). In other cases, questions do not make sense at all out of a context: no answer could be given to the question "where?" asked on its own, while following a question such as "Does Sean have a house anywhere apart from Scotland?" it becomes an easily intelligible query.

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

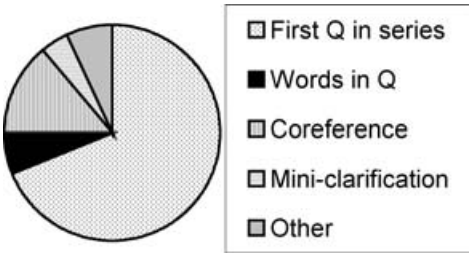


Fig. 3. Search technique used for Question.

The usual way in which Question Answering systems constrain possible answers is by restricting the number of documents in which an answer is sought by filtering the total number of available documents through the use of an information retrieval engine. The information retrieval engine selects a subset of the available documents based on a number of keywords derived from the question at hand. In the simplest case, it is necessary to note that some words in the current question refer to words in previous questions or answers and hence use these other words when formulating the IR query. For example, the question "Is he married?" cannot be used *as is* in order to select documents, as the only word passed to the IR engine would be "married" (possibly the root version "marry") which would return too many documents to be of any use. Noting that the "he" refers to a previously mentioned person (e.g. "Sean Connery") would enable the answerer to seek an answer in a smaller number of documents. Using co-reference resolution to add terms to the query would not however always work: questions such as "Where more precisely?" or "How much?", would require a much more comprehensive ability to understand and manipulate language than current methods for co-reference resolution. In all these examples, however, we hypothesized that, given that the current question is asked in the context of a previous question, the documents retrieved for the previous related question could provide a context in which to initially seek an answer.

To verify the usefulness of constraining the set of documents in which to seek an answer, a subset made of 15 clarification dialogues (about 100 questions) from the given question data was analyzed by taking the initial question for a series, submitting it to the Google Internet Search Engine and then manually checking to see how many of the questions in the series (excepting the very first question) could be answered simply by using the first 20 documents retrieved for the first question in a series. The results are summarized in the figure 3:

- 69% of clarification questions could be answered by looking within the documents used for the first question in the series, thus indicating the usefulness of noting the occurrence of clarification dialogue. In contrast, simulating a standard QA system which simply used the clarification question as a query to be sent to the search engine, only 37% of the answers could be found in the first 20 retrieved documents.
- For the remaining 31% a different approach had to be taken, in particular:
  - 6% could be answered after retrieving documents by using the words in the current question as search terms (e.g. "What caused the boxer uprising?");

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

- 14% required some form of coreference resolution and could be answered by combining the words in the question with the words to which the relative pronouns in the question referred (e.g. "What film is he working on at the moment", with the reference to "he" resolved, which gets passed to the search engine as "What film is Sean Connery working on at the moment?");
- 7% required more than 20 documents to be retrieved by the search engine or other, more complex techniques: for example, a question such as "Where exactly?" required both an understanding of the context in which the question was asked ("Where?" makes no sense on its own) and the previously given answer (which was probably a place, but not restrictive enough for the questioner).
- 4% constituted subdialogues within a larger clarification dialogue (a slight deviation from the main topic which was being investigated by the questioner) and could be answered by looking at the documents retrieved for the first question in the subdialogue.

Recognizing that a clarification dialogue is occurring therefore can simplify the task of retrieving an answer by specifying that an answer must be in the set of documents used to answer the first question in a series. This is consistent with the results found in the TREC context task (Voorhees 2002), which indicated that systems were capable of finding most answers to questions in a context dialogue simply by looking at the documents retrieved for the initial question in a series. As in the case of clarification dialogue recognition, simple techniques can resolve the majority of cases; nevertheless, a full solution to the problem requires more complex methods. The last case indicates that it is not enough simply to look at the documents provided by the first question in a series in order to seek an answer: it is necessary to use the documents found for a previously asked question which is related to the current question (i.e. the questioner could "jump" between topics). For example, given the following series of questions starting with Q_1:

Q_1: When was the Hellenistic Age?
[. . .]
Q_5: How did Alexander the great become ruler?
Q_6: Did he conquer anywhere else?
Q_7: What was the Greek religion in the Hellenistic Age?

where Q_6 should be related to Q_5 but Q_7 should be related to Q_1, and not Q_6. In this case, given that the subject matter of Q_1 is more immediately related to the subject matter of Q_7 than Q_6 (although the subject matter of Q_6 is still broadly related, it is more of a specialized subtopic), the documents retrieved for Q_1 will probably be more relevant to Q_7 than the documents retrieved for Q_6 (which would probably be the same documents retrieved for Q_5).

10 Conclusion

In order to fully meet users' needs and goals attention must be paid to the occurrence of clarification dialogues: in the case of clarification dialogue the meaning of, and the

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

intention behind, users' questions is constrained by previously asked questions and previously given answers. But this in turn constrains the answer: to satisfactorily answer a question it is therefore necessary to refer to the previous exchange of questions and answers. From this it follows that recognizing that the question that is currently being examined by a question answering system is part of a clarification dialogue is an important task to be carried out before attempting to find an answer. An algorithm was developed based on the "context" question sequences prepared by NIST for the TREC-QA task in order to recognize the occurrence of clarification dialogue. The algorithm was then tested on a new collection of clarification dialogues which was gathered for this purpose. The component parts of the algorithm were examined in detail and it was shown to have a good performance. Finally, it was shown experimentally that in automated open-domain question answering the use of an algorithm to recognize that a clarification dialogue is occurring can simplify the task of answer retrieval by constraining the subset of documents in which an answer is to be sought.

References

Ardissono, L. and Sestero, D. (1996) Using dynamic user models in the recognition of the plans of the user. *User Modeling and User-Adapted Interaction* **5**(2): 157–190.

BNCFreq. 2003. *English word frequency list*. `http://www.eecs.umich.edu/~qstout/586/bncfreq.html` (last accessed March 2003).

Budanitsky, A. and Hirst, G. (2001) Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. *Proceedings of the NAACL 2001 Workshop on WordNet and other lexical resources*. Pittsburgh, PA.

De Boni, M. and Manandhar, S. (2002) Automated discovery of telic relations for WordNet. *Proceedings of the First International WordNet Conference*. Mysore: India.

De Boni, M. and Manandhar, S. (2003a) The use of sentence similarity as a semantic relevance metric for question answering. *Proceedings of the AAAI Symposium on New Directions in Question Answering*. Stanford, CA.

De Boni, M. and Manandhar, S. (2003b) An analysis of clarification dialogue for question answering. *Proceedings of the HLT-NAACL Conference*, Edmonton, Canada.

Fellbaum, C. (1998) *WordNet, An electronic lexical databas*e. Cambridge, MA: MIT Press.

Ginzburg, J. (1998) Clarifying utterances. In: Hulstijn, J. and Nijholt, A. (editors), *Proceedings of the 2nd Workshop on the Formal Semantics and Pragmatics of Dialogue*. Twente.

Ginzburg, J. and Sag, I. A. (2000) *Interrogative investigations*. Stanford: CSLI Publications.

Green, S. J. (1997) Automatically generating hypertext by computing semantic similarity. Technical Report n. 366. Computer Systems Research Group: University of Toronto.

Harabagiu, S., Miller, A. G. and Moldovan, D. (1999) WordNet2 – a morphologically and semantically enhanced resource. *Proceedings of SIGLEX-99*, pp. 1–8. University of Maryland.

Harabagiu, S., Moldovan, D., Pasca, M., Surdeanu, M., Mihalcea, R., Gîrju, R., Rus, V., Lacatusu, F., Morarescu, P. and Bunescu, R. (2002) Answering complex, list and context questions with LCC's question-answering server. *Proceedings of TREC-10*, Gaithersburg: NIST Publications.

Hirst, G. and St-Onge, D. (1998) Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (editor), *WordNet: an electronic lexical database*. Cambridge MA: MIT Press.

Jiang, J. J. and Conrath, D. W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. *ROCLING X International Conference*. Taipei, Taiwan: Academia Sinica.

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press

Lin, D. (1998) An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

Mihalcea, R. and Moldovan, D. (1999) A method for word sense disambiguation of unrestricted text. *Proceedings of ACL '99*. Maryland, NY.

Miller, G. A. (1999) WordNet: a lexical database. *Communications of the ACM* **38**(11): 39–41.

Moldovan, D. and Rus, V. (2001) Logic form transformation of WordNet and its applicability to question answering. *Proceedings of the 39th conference of ACL*. Toulouse, France.

Purver, M., Ginzburg, J. and Healey, P. (2002) On the means for clarification in dialogue. In: Smith, R. and van Kuppevelt, J. (editors), *Current and New Directions in Discourse and Dialogue*. Kluwer Academic.

Purver, M., Healey, P. G. T., King, J., Ginzburg, J. and Mills, G. J. (2003) Answering clarification questions. *The 4th SIGdial Workshop on Discourse and Dialogue*. Sapporo.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T. (1994) *Human Computer Interaction*. Wokingham: Addison-Wesley.

Resnik, P. (1995) Using information content to evaluate semantic similarity. *Proceedings of the 14th IJCAI*. Montreal, Canada.

Soubbotin, M. M. (2002) Patterns of potential answer expressions as clues to the right answers. *Proceedings of TREC-10*. Gaithersburg: NIST.

van Beek, P., Cohen, R. and Schmidt, K. (1993) From plan critiquing to clarification dialogue for cooperative response generation. *Computational Intelligence* **9**: 132–154.

Voorhees, E. (2002) Overview of the TREC 2001 question answering track. *Proceedings of TREC-10*. Gaithersburg: NIST Publications.

https://doi.org/10.1017/S1351324905003682 Published online by Cambridge University Press