A Methodology to Develop Answers to Definitional Questions

Samir Kanaan
TALP Research Center
Universitat Politecnica de Catalunya
Barcelona - Spain
skanaan@lsi.upc.edu

Abstract

Definitional Question Answering task is usually proposed as the retrieval of a set of text fragments that contain information relevant to the target to define. Following this principle, a systematic approach to obtain which information is relevant is proposed, along with a new evaluation process. Finally, the paper describes the elaboration of a corpus of definitions in order to produce a gold standard model for use in the evaluation of definitional question answering systems.

1. Introduction

The question answering track of TREC 13 competition has a definitional subtask that consists in the proposal of several text fragments that supposedly include relevant information about the target of the definition not previously asked in other questions referring that target.

The evaluation of this subtask is based on a list of important ideas about the definition target or "nuggets", which can be essential or simply optional to the final definition.

Many TREC participants have criticized the spirit and the application of this evaluation scheme; see [Hildebrandt et al, 2004] for an example. Most of these critics are based on the arbitriarity of the list of nuggets used to peform the evaluation, both of the nuggets themselves as of their qualification as either essential or optional, and the actual process of manual evaluation, that leads to large differences in the final scoring of participant systems.

The organization of TREC makes judgements of the participant systems available as a kind of corpus of fragments considered right. As [Lin 2005] states for factual question answering, this collection of definitions can not be employed to evaluate definitional question answering systems because it is an incomplete list, and using the (docid,fragment) would underestimate system outputs as there may be other documents containing the same text, while using only (fragment) to evaluate will surely overestimate system responses, judging as right fragments in a wrong context.

This paper proposes a method to develop definitions in a systematic way, trying to avoid some of the problems mentioned above. Guidelines to develop the list of nuggets are also proposed, in order to obtain lists of nuggets consistent and in a systematic way. Besides the usual evaluation process used in TREC 13, an alternative process is exposed that tries to avoid the problematic classification of nuggets into essential or optional among other problems. Finally, the elaboration of a gold standard corpus is described, along with the methodology employed and its possible applications.

2. A Proposal for Definitional Question Answering

The criteria with which definitions will be evaluated should be established before any other aspects of the task are taken into consideration. Our proposal of which is the ideal output that a definitional question answering system should produce is based on the following ideas.

First, a definition should consist in a set of text fragments extracted from the corpus; it does not have to be limited to a single fragment. The fragments in the output should be completely self-explanatory and unambiguous, although the term to define can be asumed and therefore can be omitted. Table 1 shows two examples taken from the evaluation results of TREC 2004.

Term to define	Amtrak
Idea the fragment refers to	Amtrak plans to introduce a high-speed train service named Acela
Non-valid fragment (marked as good in the TREC 2004 evaluation)	Acela
Valid fragment (note the omission of the term "Amtrak")	new high-speed Acela train
Term to define	Fred Durst
Idea the fragment refers to	his style is rap-metal
Non-valid fragment (marked as good in the TREC 2004 evaluation)	rap-metal
Valid fragment	lead singer for the rap-metal band

Table 1. Example of non-valid fragments in TREC 2004 evaluation.

In second place, these text fragments should answer as many as possible questions from the following set (suppose X to be the term to define):

Defining question	**Definee**	**Fragments**
What is X?	quarks	subatomic particles
	James Dean	American actor
Which are the main features of X?	James Dean	rebellious archetypes
	Kurds	world's biggest stateless nation
	Johnny Appleseed	homeless man
	prions	at least eight different strains
For which reasons is X known/famous?	Johnny Appleseed	sowing apple seeds all over the country
	prions	cause mad cow disease and its human variant, Creutzfeldt-Jakob disesase
What does X serve for? What does X do/produce/have achieved?	quarks	two up quarks and a down quark make a proton
	James Dean	films "Rebel Without a Cause", "East of Eden" and "Giant"
With which other concepts is X strongly related?	Kurds	PKK is fighting for autonomy for Kurds in Turkey's southeast
	Fred Durst	lead singer for the rap-metal band Limp Bizkit
Which is X's spatial/temporal situation?	James Dean	1931-1955
		was killed in a car crash in California on Sept. 30, 1955
	Abercrombie and Ficth	Based in Reynoldsburg, Ohio
Which kinds of X there exist?	quarks	three different types whimsically called "colors": red, green and blue
Which are the components of X?	Rat Pack	Frank Sinatra, Dean Martin and Sammy Davis Jr.

Table 2. Examples of the application of defining questions to several definees.

- What is X? (Genus; see [Sager and L'Homme, 1994])
- Which are the main features of X? (Species; see [Sager and L'Homme, 1994])

- For which reasons is X known/famous?
- What does X serve for? / What does X do/produce/have achieved?
- With which other concepts is X strongly related (should include the nature of this relation)?
- Which is X's spatial/temporal situation? (Including birth/death dates and places if X is a human being)
- Which kinds of X are there?

The definitional question answering system described in [Blair-Goldesohn, 2003] is also based on the concepts of Genus and Species listed in table 2, along with other categories. In [Swartz, 1997], a definition of what should be a definition is presented, along with the description of Genus and Species.

Depending on the nature of the term to define, some of the previous questions may not make sense; the expected definition should contain text fragments that answer all the questions that apply to the term. Table 2 illustrates the meaning of each proposed question with several examples.

Note that one important consequence of this scheme is that nuggets asked in previous questions shall not be discarded, as they usually belong to the inner core of the definition of the target, and removing them from it lets the task rather incomplete. This does not imply, however, that in further competitions the removal of previously asked nuggets cannot be applied; it is simply that the definition corpus should include all possible nuggets for a target.

If the term to define X has been associated to another term Y, it is possible to include further information about term Y if it is relevant to define X, as the examples in table 3 illustrate.

Definee	**Defining fragments**
Kurds	Kurds PKK is fighting for autonomy for Kurds in Turkey's southeast
Kurds	PKK leader Abdullah Ocalan
Fred Durst	lead singer for the rap-metal band Limp Bizkit
Fred Durst	Bizkit's last album (the 1.5 million-selling ``Three Dollar Bill, Y'all")

Table 3. Examples of valid associative chains.

The application of each of the questions proposed to a term to define may produce zero, one or more possible pieces of information as answer; each of these possible pieces of information, which will be called "nuggets" in order to follow TREC nomenclature, will be the unit of evaluation. Section 3 explains the details of this evaluation.

3. Evaluation of Definitions

In this section two evaluation methods are presented: the method used in last TREC evaluation and an alternative method. Both are based on a list of nuggets which serves to score the answers given by the systems.

TREC evaluation of definitional questions is based on a list of nuggets divided into two categories: essential nuggets (vital) and optional nuggets (okay). The answer is expected to contain all the vital nuggets in the list, and the presence of okay nuggets only justifies a longer answer. Measures used are described in figure 1 (extracted from [Voorhees, 2004]); as it can be seen, recall is more important than precision in the final f-score measure, and okay nuggets only increase precision marginally, so their role in this evaluation is rather small.

Classification of nuggets into vital or okay may lead to some difficult cases. For example, suppose that the term to define is "James Dean" and that one vital nugget is "Appeared in three films: 'Rebel Without a Cause', 'East of Eden' and 'Giant'". How will be evaluated a fragment that only mentions one of those films? A possible solution would be to divide the previous nugget intro three independent ones, but then would they be vital or okay? They clearly do not have the same relevance as the original nugget with the three films.

Let
r: number of vital nuggets in system response
a: number of okay nuggets in system response
R: number of vital nuggets in the list elaborated by judges
L: length of the system response, i.e., number of non-whitespace characters of all the fragments in the response

Then

$$recall = r / R$$

$$allowance = 100 * (r + a)$$

$$precision = \begin{cases} 1 & si\ L < allowance \\ 1 - \dfrac{L - allowance}{L} & otherwise \end{cases}$$

$$f - score_{b=5} = \frac{10 * precision * recall}{9 * precision + recall}$$

Figure 1. TREC 2004 evaluation of definitional questions

An alternative evaluation scheme that tries to overcome the previous problem and also avoids the difficult task of classifying a nugget as either vital or okay is presented next. It is also based on a list of nuggets, but decomposing nuggets like the one from the previous example ("Appeared in three films...") into three independent nuggets. Each nugget of the list, instead of being classified as either vital or okay, would be assigned a relative weight, so that the sum of the relative weights of all the nuggets of a term equals 1. Table 4 shows an example of this scheme, where it can be seen that each film of James Dean is a separate nugget with a relative importance of 0.1.

This approach requires an important question to be solved: how to assign weight to the different nuggets obtained. As different targets will have different definitional features associated with, and their relative importance will certainly vary.

A possible solution for this problem is to measure the distance of the question target to the main concept of the nugget counting the number of hops in an ontology (for example WordNet, see [Miller 1995]) from the former to the latter; the higher the number of hops required, the lesser the importance. Obviously, all the weights would have to be normalized to make them sum 1.

Defining question	Nugget	Nugget weight
What is “James Dean”?	Actor	0.2
Which are the main features of “James Dean”?	Rebellious character	0.08
	Tormented	0.05
For which reasons is “James Dean” famous?	“Rebel Without a Cause”	0.06
	“East of Eden”	0.06
	“Giant”	0.06
	Changed the roles of actors in movies	0.1
With which other concepts is “James Dean” strongly related?	Film director Elia Kazan	0.05
Which is “James Dean”’s spatial/temporal location?	Born in Fairmount, Indiana	0.01
	Born in 1931	0.04
	Screen debut in 1951	0.05
	Died in 1955 (at age 24)	0.1
	Died on California	0.04
	Died in a car accident	0.1

Table 4. Example of weighted nugget list.

The measurements used to compute the evaluations using weighted nuggets would be rather similar to those used in current TREC evaluation, with recall being the percentage of information retrieved and precision being a measure of the density of information of the answer strings. Figure 2 details the new expressions for recall, precision and f-score.

Let
r: sum of relative weights of nuggets in system response
a: number of okay nuggets in system response
R: number of vital nuggets in the list elaborated by judges
L: length of the system response, i.e., number of non-whitespace characters of all the fragments in the response

Then

$$recall = r / R$$

$$allowance = 100 * (r + a)$$

$$precision = \begin{cases} 1 & si\ L < allowance \\ 1 - \dfrac{L - allowance}{L} & otherwise \end{cases}$$

$$f - score_{b=5} = \frac{10 * precision * recall}{9 * precision + recall}$$

Figure 2. Alternative evaluation of definitional questions

This alternative scheme of evaluation exhibits the following advantages:

- It allows to easily split a compound nugget into independent, "atomic" nuggets.
- It avoids the difficult binary classification into vital/okay, allowing for a smooth relevance assignment to nuggets. The fact that a nugget be vital or okay may have a great influence on evaluation results, while a difference of a small percentage in the relative weight assigned to a nugget would only have a marginal influence on final evaluation results.

- The relative importance of a nugget can be easily boosted if the nature of the definition target requires it.
- It gives all nuggets some relevance in recall, as all of them are supposed to give relevant information on the definition target.
- There is a systematic, predictable method to assign weights to the nuggets proposed by judges.

4. A Corpus of Definitions

Definitional question answering is a relatively recent task and it lacks two important elements: first, a clear, objective and systematic definition of the task; second, a significatively large corpus of definitions that allows research groups to evaluate their systems.

The development of a corpus of definitions is a very time-consuming task, as each definition consists of several nuggets, which may be expressed in different ways throughout the text corpus, and therefore it is almost impossible to guarantee that all posible forms of a nugget are collected. However, depending on the use that will be given to the corpus, see section 5, it would not be as important to collect an exhaustive list of fragments where a nugget appears.

What has been done as a first step in this direction is to improve the existing corpus of system outputs for TREC 2004 in order to have guaranteed some properties:

- All fragments have been reviewed and it has been verified that all of them are right fragments of the corpus and the corresponding document effectively contains such a fragment. Many fragments had to be modified (where tokenized, all in lowercase, or even lemmatized), and some of them did not appear in the corresponding document, so they were removed.
- For each fragment in the corpus of definitions, all the ocurrences of the same text where retrieved from the text corpus and it was manually verified that in fact they referred to the same idea; in that case, they were added to the corpus, so that the evaluation with (docid,text) would not underestimate system responses.
- Empty nuggets in the system responses where filled with text fragments manually located using keyword search.

As it is exposed in section 5, this definition corpus already allows an automatic evaluation of definitional question answering systems by counting matching (docid,text) fragments in the system response. The evaluation can be slightly underestimated, as there might be variations of the same idea in the text corpus not included in the definition corpus, but it is almost impossible to find all the variations of the same idea.

A second stage of the elaboration of a corpus of definitions would require to apply a set of systematic guidelines, as those exposed in section 3, in order to obtain a definition corpus as regular and systematic as possible, with objective criteria in the elaboration of the list of nuggets, and assignment of weights to the nuggets instead of classifying them into vital/okay.

The features of this corpus would be:

- It would have to grow far beyond 65 question targets in order to reach some statistical significance. An initial size of 200 question targets has been estimated.
- It would not be biased towards persons/organizations, but instead would contain a balanced mix of persons, organizations and common names, in order to approach more difficult and irregular definitions (those of common names).

- Completeness of its text fragments: a search as exhaustive as possible would be made throughout the whole Aquaint corpus in order to retrieve each possible fragment representative of a given nugget.
- Completeness of its nuggets: although this point is impossible to guarantee, for each target several information sources will be consulted (specialized web pages on the subject) in order to retrieve answer the defining questions proposed above. Only nuggets with supporting fragments in the Aquaint document collection will be included.
- The format of the corpus would be the same of the file with all the fragments of the nuggets for TREC 2004 questions provided by the organization of TREC (list of targets, each with a list of nuggets, each with a list of document identifier and text fragment).

4. Applications of the Corpus of Definitions

The main goal of the elaboration of this corpus of definitions is to allow automatic evaluation of systems, in order to improve their development cycle and therefore their performance. There are several ways in which this corpus of definitions could be used to evaluate systems.

First, the corpus can be used to count which fragments of a system response match with any of the supporting fragments for a given nugget; the number of nuggets contained in a system response would be the number of matches in the nugget list, assuming that both the document identifier and the text fragment match.

In this first approach, TREC evaluation would be followed, or possibly the variation exposed in figure 2 if nuggets are assigned a relative weight.

Another automatic evaluation method for definitions called POURPRE is described in [Lin and Demner-Fushman 2005]. It is a variation of the ROUGE ([Lin and Hovy 2003]) method of evaluation of automatic summaries, and the authors afirm that it is highly correlated with the manual evaluation of TREC 2004 (a Kendall's tau of 0.88).

Basically this method employs the nuggets in the corpus to count the unigram matches of each fragment of the systems response; the degree of coincidence indicates the probability of the presence of that nugget in the response.

Finally, the authors are researching on the application of the evaluation framework QARLA ([Amigó et al 2005]), designed for evaluation of automatic summaries, that uses several selections of the fragments in the corpus of definitions as reference models and allows the combination of different metrics to evaluate a system's response against the reference models.

5. Conclusions and Future Work

The task of definitional question answering has not reached its maturity yet, first because there is no clear and systematic definition of the task, second because the evaluation method has not achieved consensus, and third because there is no gold standard against which compare results of the systems under development.

One possible method of developing definitions in a systematic way is to elaborate them as a set of responses to a list of questions that pretend to capture the main features of the definee.

An improvement on the evaluation method would be to substitute the vital/okay classification of nuggets by a weighting scheme.

Finally, a proposal for the construction of a corpus of definitions is exposed, with a first stage already developed and a second, more ambitious stage that aims at the obtention of a gold standard corpus that would allow different automatic evaluation methods to be applied and would serve of reference to systems under development.

References

[Amigó et al 2005] Enrique Amigó, Julio Gonzalo, Anselmo Peñas and Felisa Verdejo. 2005. QARLA: A Framework for the Evaluation of Text Summarization Systems. *Proceedings of the ACL 2005.*

[Blair-Goldesohn, 2003] S. Blair-Goldesohn, K.R. McKeown, A.H. Schlaikjer. 2003. A Hybrid Approach for Answering Definitional Questions. *Technical Report CUCS-006-03*. Columbia University.

[Hildebrandt et al, 2004] W. Hildebrandt, B. Katz, J. Lin. 2004. Answering Definition Questions Using Multiple Knowledge Sources. *HLT/NAACL 2004.*

[Lin 2005] Jimmy Lin. 2005. Evaluation of Resources for Question Answering Evaluation. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005).*

[Lin and Demner-Fushman 2005] Jimmy Lin and Dina Demner-Fushman. 2005. Automatically Evaluating Answers to Definition Questions. *Technical Report LAMP-TR-119/CS-TR-4695/UMIACS-TR-2005-04.* University of Maryland.

[Lin and Hovy 2003] Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries using n-gram Co-ocurrence Statistics. *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2003).*

[Miller 1995] G. Miller. 1995. WordNet: a Lexical Database. *Communications of the ACM 38(11),* pp. 39-41.

[Sager and L'Homme, 1994] J. C. Sager and M.C. L'Homme. 1994. A model for definition of concepts. Terminology, pages 351-374.

[Swartz, 1997] N. Swartz. 1997. Definitions, dictionaries and meanings. Posted online at http://www.sfu.ca/philosophy/definitn.htm.

[Voorhees, 2004] E.M. Voorhees. 2004. Overview of the TREC 2004 Question Answering Track. *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004).*