# Identifying Definitions in Text Collections for Question Answering

**Horacio Saggion**

Department of Computer Science
University of Sheffield
Regent Court
211 Portobello Street
S1 4DP - Sheffield - England
UK
saggion@dcs.shef.ac.uk

**Abstract**

One particular type of question which was made the focus of its own subtask within the TREC2003 QA track was the definition question ("What is X?" or "Who is X?"). One of the main problems with this type of question is how to discriminate in vast text collections between definitional and non-definitional text passages about a particular definiendum (i.e., the term to be defined). A method will be presented that uses definition patterns and terms that co-occurr with the definiendum in on-line sources for both passage selection and definition extraction.

## 1. Introduction

The research is concerned with the problem of finding definitions in vast text collections and with the resources necessary to carry out this task. The problem is related to the TREC QA 2003 definition subtask, where given a huge text collection like AQUAINT (over 1 million texts from the New York Times, the AP newswire, and the English portion of the Xinhua newswire and totals about 3.2 gigabytes of data) and a definition question like "What is Goth?" or "Who is Aaron Copland?", an automatic system has to find text fragments that convey essential and non-essential characteristics of the main question term (e.g., "Goth" or "Aaron Copland"). This is a challenging problem not only because of the many ways in which definitions can be conveyed in natural language texts but also because the definiendum (i.e., the thing to be defined) has not, on its own, enough discriminative power to allow selection of definition-bearing passages from the collection.

In order to find good definitions, it is useful to have a collection of metalanguage statements (i.e., "DEFINIENDUM is a", "DEFINIENDUM consists of", etc.) which implement patterns for identification and extraction of "definiens" (the statement of the meaning of the definiendum). Unfortunately there are so many ways in which definitions are conveyed in natural language that it is difficult to come up with a full set of linguistic patterns to solve the problem. To make matters more complex, patterns are usually ambiguous, matching non-definitional contexts as well definitional ones. For example, a pattern like "Goth is a" to find definitions of "Goth", will match "Becoming a goth is a process that demands lots of effort" as well as "Goth is a subculture".

In this paper we describe a method that uses external sources to mine knowledge which consists of terms that co-occur with the "definiendum" before trying to define it using the given text collection. This knowledge is used for definition identification and extraction (for the complete description of the method the reader is referred to (Saggion and Gaizauskas, 2004)).

## 2. Definition Knowledge

There are two sources of knowledge we rely on for finding definitions: linguistic patterns, which represent general knowledge about how definitions are expressed in language; and secondary terms, which represent specific knowledge about the definiendum outside the target collection.

### 2.1. Linguistic Patterns

Definition patterns or metalanguage statements containing lexical, syntactic, and sometimes semantic information have been used in the past in research in terminology (Pearson, 1998), ontology induction (Hearst, 1992), and text summarization (Saggion and Lapalme, 2002) among others.

When a corpus for specific purposes is available, then patterns can be combined with well formed terms or specific words to restrict their inherent ambiguity. One simple formal defining expositive proposed by Pearson (1998) is "X = Y + distinguishing characteristics" where possible fillers for "X" are well formed terms (those word sequences following specific patterns), fillers for "Y" are terms or specific words from a particular word list (e.g., method, technique, etc.), and fillers for "=" are connective verbs such as "to be", "consist" or "know". The

use of predefined word lists or term formation criteria is, however, not possible in our case because we are dealing with an heterogeneous text collection where the notion of term is less precise than in a corpus of a particular domain.

Dictionaries are good sources for the extraction of definition knowledge. Recent research in classification and automatic analysis of dictionary entries (Barnbrook, 2002) has shown that a limited number of strategies for expressing meaning in those sources exist and that automatic analysis can be carried out on those sources to extract lexical knowledge for natural language processing tasks. Barnbrook (2002) identified 16 types of definitions in the Cobuild student's dictionary and extraction patterns used to parse them (e.g. "A/an/The TERM is/are a/an/the..."). The question remains as whether this typology of definition sentences (and associated extraction patterns) is sufficient to identify definition statements in less structured textual sources.

We have collected, through corpus analysis and linguistic intuition, a useful set of lexical patterns to locate definition-bearing passages. The purpose of these patterns is on the one hand to obtain definition contexts for the definiendum outside the target collection in order to mine knowledge from them, and on the other hand to use them for extracting definiens from the target collection. 36 patterns for general terms and 33 patterns for person profiles have been identified, a sample can be seen in Table 1, patterns used in this work contain only lexical information.

### 2.2. Secondary Terms

Terms that co-occur with the definiendum (outside the target collection) in definition-bearing passages seem to play an important role for the identification of definitions in the target collection. For example, there are 217 sentences referring to "Goth" in the AQUAINT collection, only a few of them provide useful definitional contexts, we note that the term "subculture" usually occurs with "Goth" in definitional contexts on the Web, and there are only 6 sentences in AQUAINT which contain both terms. These 6 sentences provide useful descriptions of the term "Goth" such as "the Goth subculture" and "the gloomy subculture known as Goth". So, the automatic identification of specific knowledge about the definiendum seems crucial in this task.

Our method considers nouns, verbs and adjective as candidate secondary terms. Sources for obtaining definition-passages outside AQUAINT for mining secondary terms are the WordNet lexical database (Miller, 1995), the site of Encyclopedia Britannica (`http://www.britannica.com`), and general pages on the web. The passages are obtained automatically from the Web by using the Google API (`http://www.google.com/apis`) exact search facility for each definition pattern.

Terms that co-occur with the definiendum are obtained following three different methods: (i) words appearing in WordNet glosses and hypernyms of the definiendum are extracted; (ii) words from Britannica sentences are extracted only if the sentence contains an explicit reference to the definiendum; (iii) words from other Web sentences are extracted only if the sentences match any definition pattern. Extracted terms are scored based on their frequency of occurrence. Table 2 shows top ranked terms mined from on-line sources for "Aaron Copland" (famous American musician who composed the ballet "Appalachian Spring") and "golden parachutes" (compensation given to top executives that is very generous).

## 3. Identifying Definitions in Texts

In order to select text passages from AQUAINT we rely on the Okapi probabilistic document retrieval system (Robertson and Walker, 1999). Passages are retrieved by querying Okapi with the original definition question expanded with the list of associated secondary terms. For TREC QA 2003 we used the top 20 documents retrieved, this number was identified after experiments measuring end-to-end performance of our factoid-QA system, this number of candidate passages for definitional-QA proved to be too small, as recent experiments have revealed.

We perform a linguistic analysis of each passage which consists of: tokenisation, sentence splitting, matching using the definiendum and any of the definiendum's secondary terms, and pattern matching using the definition patterns. We restrict our analysis of definitions to the sentence level. A sentence is considered a definition-bearing sentence if it matches a definition pattern or if it contains the definiendum and at least three secondary terms.

We perform sentence compression extracting a sentence fragment that is a sentence suffix and contains main and all secondary terms appearing in the sentence, this is done in order to avoid the inclusion of unnecessary information the sentence may contain. For example the definition of "Anthony Blunt" extracted from the sentence.

> *The narrator of this antic hall-of-mirrors novel, which explores the compulsive appeal of communism for Britain's upper classes in the 1930s, is based on the distinguished art historian Anthony Blunt, who was named as a Soviet spy during the Thatcher years.*

is

| General patterns | Person patterns |
|---|---|
| define TERM as | PERSON known for |
| TERM and others | PERSON who was |
| TERM consists of | PERSON a member of |

Table 1: The first column contains patterns for general or common terms. The second column contains patterns for person profiles.

| Definiendum | Secondary terms |
|---|---|
| Aaron Copland | music, american, composer, classical, appalachian, spring, brooklyn, etc. |
| golden parachutes | plans, stock, executive, compensation, millions, generous, top, etc. |

Table 2: Terms tha co-occur with the definiendum in definition-bearing passages.

*art historian Anthony Blunt, who was named as a Soviet spy during the Thatcher years.*

All candidate definitions are proposed as answers unless they are too similar to any of the previous extracted answers. We measure similarity of a candidate definition to a previously extracted definition from the collection using tf*idf and a cosine similarity measure.

## 4. Evaluation

The method described here was used in the recent TREC QA 2003 competition. The subtask required finding answers for 50 definition questions. The set consisted of 30 "Who" definition questions and 20 "What" definition questions. TREC assessors created for each question a list of acceptable information nuggets (pieces of text) from all returned system answers and information discovered during question development. Some nuggets are considered essential (i.e., a piece of information that should be part of the definition) while others are considered non-essential. During evaluation, the assessor takes each system response and marks all essential and non-essential nuggets contained in it. A score for each question consists of nugget-recall (NR) and nugget-precision (NP) based on length. These scores are combined in the F-score measure with recall five times as important as precision. We obtained a combined F-score of 0.236. The F-score of the systems that participated in the competition is 0.555 (best), 0.192 (median), 0.000 (worst). Our method was considered among the top 10 out of 25 participants.

After submission of our answers to TREC we discovered that, when extracting secondary terms, we omitted the extraction of proper nouns. This has an impact on the discovery of relevant secondary terms not only for defining people but also for defining common things. We also discovered that having disconsidered name aliases of the definiendum (e.g., "Tomba" instead of "Alberto Tomba") there are good definition-bearing sentences with no chance of being selected (e.g., "Tomba, a three-time Olympic champion..."). In many cases, answers could not be extracted because the definition patterns and filters were far too restrictive to cover these definitions (recall problem). A final problem to be considered is the number of documents/passages to be examined by the extraction system. After fixing the problems mentioned above and having relaxed the extraction patterns, we re-evaluated our method using different numbers of returned passages. Results of the experiment are shown in Table 3. Both, pattern relaxation and increased number of documents examined have a positive impact on the method's performance.

## 5. Related Work

The problem of the definition has its roots in Aristotle with considerations about its basic constituents: *genus* and *differentia*. Many studies have concentrated on different aspects of the definition (e.g., (Chaurand and Mazière, 1990) from philosophical to terminological considerations, (Wilks et al., 1995) for an account of the use of definitions computationally). In the recent TREC 2003 QA definition subtask evaluation, participants used various techniques similar to those presented here. Top ranked groups report on the use of some form of lexical resource like WordNet, the Web for answer redundancy, patterns for definition identification and sophisticated linguistic tools (Kouylekov et al., 2003; Harabagiu et al., 2003). BBN's definitional system (Xu et al., 2003) that obtained the best performance in TREC QA relies on the identification, extraction, and ranking of *kernel facts* about the *question target* (i.e. definiendum) followed by a redundancy removal step. The system uses sophisticated linguistic analysis components such as parsing and coreference resolution. First, sentences containing the question target in the top 1000 documents retrieved by

| Definition filter | 20 passages | 500 passages |
|---|---|---|
| (definiendum and three secondary terms) or pattern | 0.2915 | 0.3806 |
| ((definiendum or alias) and one secondary term) or pattern | 0.3427 | 0.4350 |

Table 3: Effect of relaxation of definition filters and document rank on system performance (F-score).

an information retrieval system are identified; then, kernel facts are identified in those sentences using criteria such as the presence of copula or appositive constructions involving the question target, or matching of a number of structural patterns (e.g., *TERM is a NP*), or containing special predicate-argument structures (e.g., *PERSON was born on DATE*), or presence of specific relations (e.g., *spouse of*, *staff of*); finally, kernel facts are ranked by a metric that takes into account their type and their similarity (using tf*idf metric) to a question profile constructed from on-line sources or from the set of identified kernel facts. QUALIFIER (Yang et al., 2003) obtained the second best performance using a data-driven approach to definitional QA. The system uses linguistic tools such as fine-grained named entity recognition and coreference resolution. WordNet and the Web are used to expand the original definition question to bridge the semantic gap between query space and document space. Given a set of documents retrieved from AQUAINT after query expansion, extractive techniques similar to those used in text summarization are applied. The basic measure used to score sentences is a logarithmic sum of a variant of the tf*idf measure for each word. This metric scores a word proportional to the number of times it appears in sentences "containing" the definiendum and inversely proportional to the number of times it appears in sentences that do not contain the definiendum. Scores for words are computed from two sources: AQUAINT sentences and Web sentences. Sentence scores are first computed using word scores obtained from AQUAINT and Web and then these scores are combined in a linear way to obtain the sentence final value. Once all sentences have been evaluated and ranked, an iterative redundancy removal technique is applied to discard definitional sentences already in the answer set.

## 6. Conclusions and Future Work

With the massive availability of on-line text, tools for finding definitions in unstructured textual sources are of great importance either for ad hoc querying or for automatically constructing glossaries. We have described a method that contributes a viable and practical solution for definitional QA, because it relies on available on-line resources and on simple natural language techniques. Analysis of the results obtained in the recent TREC QA 2003 competition indicate a number of desirable improvements including the need to examine as many candidate documents as possible and a pattern relaxation technique to overcome the well known recall problem.

## 7. References


G. Barnbrook. 2002. *Defining Language. A local grammar of definition sentences*. John Benjamins Publlishing Company.

J. Chaurand and F. Mazière, editors. 1990. *La définition*. Langue et Langage. Larousse.

S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, and J. Bensley. 2003. Answer Mining by Combining Extraction Techniques with Abductive Reasoning. In *Proceedings of TREC-2003*.

M.A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of COLING'92*, Nantes.

M. Kouylekov, B. Magnini, M. Negri, and H. Tanev. 2003. ITC-irst at TREC-2003: the DIOGENE QA system. In *Proceedings of TREC-2003*.

George A. Miller. 1995. WordNet: A Lexical Database. *Communications of the ACM*, 38(11):39–41, November.

J. Pearson. 1998. *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. Jhon Benjamins Publishing Company.

S. Robertson and S. Walker. 1999. Okapi/Keenbow at TREC-8. In *Proceedings of the 8th Text REtrieval Conference*.

H. Saggion and R. Gaizauskas. 2004. Mining on-line sources for definition knowledge. In *Proceedings of FLAIRS 2004*, Orlando, Florida. AAAI.

H. Saggion and G. Lapalme. 2002. Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics*, 28(4):pp497–526.

Y. Wilks, B.M. Slator, and L. Guthrie. 1995. *Electric Words*. MIT Press.

J. Xu, A. Licuanan, and R. Weischedel. 2003. TREC2003 QA at BBN: Answering Definitional Questions. In *Proceedings of TREC-2003*.

H. Yang, H. Cui, M. Maslennikov, L. Qiu, M.-Y. Kan, and T.-S. Chua. 2003. QUALIFIER in TREC-12 QA Main Task. In *Proceedings of TREC-2003*.