Mining on-line sources for definition knowledge

Horacio Saggion and **Robert Gaizauskas**
Department of Computer Science - University of Sheffield
211 Portobello Street - Sheffield, England, UK, S1 4DP
Tel: +44-114-222-1947
Fax: +44-114-222-1810
{saggion,robertg}@dcs.shef.ac.uk

Abstract

Finding definitions in huge text collections 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. We have developed a method that uses already available external sources to gather knowledge about the "definiendum" before trying to define it using the given text collection. This knowledge consists of lists of relevant secondary terms that frequently co-occur with the definiendum in definition-bearing passages or "definiens". External sources used to gather secondary terms are an on-line enyclopedia, a lexical database and the Web. These secondary terms together with the definiendum are used to select passages from the text collection performing information retrieval. Further linguistic analysis is carried out on each passage to extract definition strings from the passages using a number of criteria including the presence of main and secondary terms or definition patterns.

Introduction

Finding textual answers to open domain questions (Hirschman & Gaizauskas 2001) in huge text collections is a challenging problem. Since 1999, the National Institute of Standards and Technology (NIST) has been conducting a series of Question Answering (QA) evaluation workshops to advance the state-of-the-art in question answering of open-domain, closed-type questions. One of the texts collections used in the evaluations is the AQUAINT collection which consists of 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.

One particular type of question which was made the focus of its own subtask within the TREC2003 QA track was the definition question, for example "What is aspirin?" (`http://trec.nist.gov/data/qa.html`). It would be naive to think that we could answer this type of question relying exclusively in dictionaries or encyclopedias, because human knowledge is ever expanding, those sources are always incomplete. As a result, tools for finding definitions in unstructured textual sources are of great importance, either for ad hoc querying or for automatically constructing glossaries.

One possible strategy when looking for definitions is to rely on a system architecture that uses a document retrieval engine as an effective filter between a text collection and a definition extraction component. Using document retrieval techniques to retrieve definition-bearing texts from general text collections is not straightforward, because definition questions provide very little information for finding relevant definition-bearing passages in the collection apart from the "definiendum". For example, a term like "aspirin" occurs in 1108 sentences in the AQUAINT collection and most of these sentences are not "aspirin" definition-bearing sentences. However, if we know from external sources that a term like "analgesic" usually occurs with "aspirin" in definition context, and we use both "aspirin" and "analgesic" as search terms for definition-bearing passages, then the search space is reduced to only 8 sentences in the AQUAINT collection.

In order to find good definitions, apart from reducing the search space to only a few good passages, it is useful to have a collection of definition patterns (i.e., "DEFINIENDUM is a", "DEFINIENDUM consists of", etc.) that implement filters for extracting "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 worse, patterns are usually ambiguous, matching non-definitional contexts as well definitional ones. For example, "aspirin is a" will match "aspirin is a great choice for active people" as well as "aspirin is a weak monotripic acid".

We have developed a method for identifying answers to definition questions by using secondary terms to restrict the search for definition-bearing passages together with a limited number of "seed" definition patterns. The secondary terms, which are mined from on-line sources, are words that co-occur with the definiendum in definition contexts outside the target collection, and as such they are used to refine the search for definition-bearing passages in the target collection. After reducing the search space to a few passages, definition patterns and filters involving the secondary terms are used to select the best non-redundant answers.

This paper discusses the acquisition "on-the-fly" of

knowledge necessary to find definition-bearing passages and the system that extract definiens from the retrieved passages.

Acquisition of knowledge from on-line sources

Obvious places for mining definition knowledge are dictionaries or encyclopedias. These resources are particularly appropriate when trying to create a library of general definition patterns for any term. In our case, however, we aim to obtain knowledge about particular terms that may or may not appear in these resources. In such cases, the redundancy of the Web (Brill *et al.* 2001) can be explored. We rely on the WordNet lexical database (Miller 1995), the site of Encyclopedia Britannica (`http://www.britannica.com`), and the Web for finding definition-bearing passages for the term sought. We use the Google API (`http://www.google.com/apis`) to fetch documents from Web pages.

The process (illustrated in Figure 1) consists of the following steps:

- Definiendum extraction: the term sought is identified in the question;
- Pattern generation: a number of definition patterns containing the term sought are generated;
- Obtaining texts for mining: texts containing the patterns are obtained from the web; additional texts are obtained from WordNet;
- Term extraction: words that co-occur with the term sought in sentences matching any of the patterns are extracted.

Each of these steps is described in detail in the following subsections. Linguistic analysis is performed using Gate (Cunningham *et al.* 2002) and LaSIE (Humphreys *et al.* 1998) components.

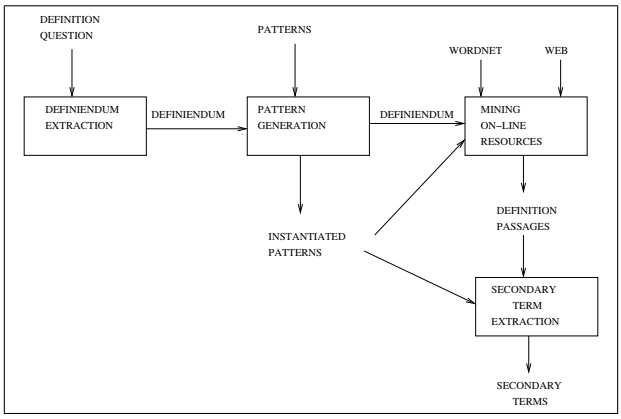

Figure 1: Knowledge extraction processes

Definiendum extraction

Since the definiendum can be a complex linguistic unit (e.g., "What is Aum Shirikyo?"), we rely on named entity recognition, part-of-speech tagging, and parsing in order to extract it from the question. Named entity recognition and classification is based on gazetteer list and regular expression matching. The result of this analysis together with the part-of-speech information is passed to a chart parser that uses a feature-based context-free grammar. The result is a syntactic representation of the question, that may be a set of unconnected chunks, from which we extract the definiendum (the rightmost noun phrase in the parse tree). Note that in this research we are not considering the problem of finding definitions for a definiendum expressed as a verb (e.g., "What is to parse?").

Pattern generation

A repository of 50 "seed" definition patterns containing a variable for the definiendum has been manually assembled through analysis of texts. Some patterns are specific of "What is X?" questions while others are specific of "Who is X?" questions. These patterns were inspired by dynamic linguistic and conceptual patterns used in text summarization by Saggion & Lapalme (2002) to locate useful text fragments defining specific terms. However, compared with that approach, the patterns used in this work are rather superficial, containing only lexical information. The idea of using reliable seed patterns has also been exploited by Hearst (1992) to identify lexical semantic relations such as hyperonymy, in order to construct ontologies from corpora.

The patterns are used by instantiating them for a given definiendum on-the-fly and stored for further linguistic processing. In Table 1 we show some definition patterns, their instantiation after the definiendum is known, and passages containing an exact pattern match.

Mining on-line sources

A collection of texts is obtained from on-line resources:

- WordNet: we extract all glosses and all hypernyms (super-ordinate words) associated with the definiendum; this is done by using a WordNet interface implemented in Gate.
- Britannica: we retrieve all pages containing the definiendum from `http://www.britannica.com` (because we do not subscribe to this resource, we have only access to incomplete information from this site). We use the Google API for this purpose submitting the search term and restricting the search to this particular site.
- Web: for each instantiated pattern, we fetch English HTML pages containing the pattern (e.g., "aspirin is a") relying on the Google Search exact match facility. Note that the same Web page can match different patterns, we keep one version of each, however.

  All texts found are stored for further processing.

Secondary term extraction

Secondary terms are obtained from the texts found in the previous step by performing the following processes using Gate: tokenisation; sentence splitting; part-of-speech tagging; pattern matching of the instantiated definition patterns. Three different methods are used depending on the source of the information:

Pattern		Passage	
uninstantiated	instantiated	relevant	non-relevant
TERM is a	aspirin is a	**Aspirin is a** weak monotripic acid	**Aspirin is a** great choice for active people
such as TERM	such as aspirin	blood-thinners **such as aspirin** ...	Look for travel size items **such as aspirin** and...
like TERM	like aspirin	non-steroidal anti-inflammatory drugs **like aspirin** ...	A clown is **like aspirin**, only he works twice as fast

Table 1: The first column contains uninstantiated definition patterns. The second column contains the pattern instantiated with the term "aspirin". The third column contains definition-bearing passages matching the pattern. The last column contains non definition-bearing passages matching the pattern.

- For each sentence in a WordNet gloss we extract nouns, verbs, and adjectives and store them in a table. We also extract and store hypernyms of the term sought.
- For each sentence in the Britannica documents, we extract nouns, verbs, and adjectives only if the sentence contains the definiendum. We record in a table the words and their frequency of occurrence.
- For each sentence from other Web sources: if the sentence contains a definition pattern, then we extract nouns, verbs, and adjectives and we record them in a table with their frequency of occurrence.

Terms in the Britannica and Web tables are sorted in decreasing order of their frequencies.

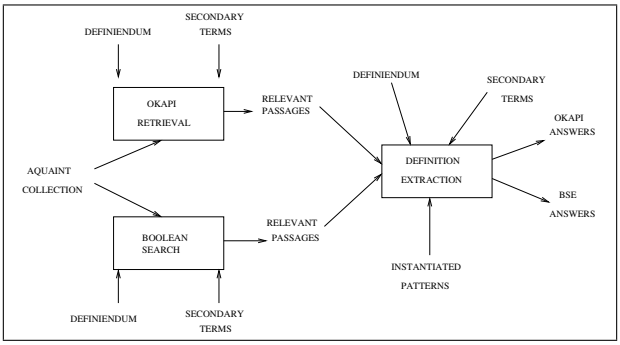

Figure 2: Extraction system

Table 2 shows secondary terms found in three different sources for the terms "aspirin" and "Aum Shirikyo" (a Japanese sect).

A list of terms to be used for query expansion (with a maximum of n elements) is created in the following way: first, all secondary terms from WordNet (m terms) are included in the list; after that, a maximum of $(n-m)/2$ Britannica terms with frequency greater than one are added to the list; finally Web terms with frequency greater than one are appended to the list until the maximum of n is reached. The order in the list reflects the degree of confidence we have in the sources, and also the fact that the more a term is used in a particular "definition" context the more we believe the term is associated with the definiendum.

Answer extraction system

The answer extraction system, shown in Figure 2, consists of a document retrieval component using a query derived from the definition question and the secondary terms, followed by a passage analysis and extraction module. We have used two different systems for document retrieval: Okapi (Robertson & Walker 1999), a probabilistic-type retrieval of textual material that uses inverted indexes, and a boolean search engine that indexes sentences.

Query generation and passage retrieval

The query for Okapi is composed of the list of all tokens contained in the question and the list of secondary terms found during the knowledge acquisition step. This query is submitted to Okapi in order to obtain the 20 most relevant passages from the AQUAINT collection. We have parameterized Okapi so that passages may contain between one and three paragraphs.

In the case of the boolean search an iterative procedure is used. Suppose the term sought is composed of tokens $\{m_i\}$ $(1 \leq i \leq k)$ and the list of secondary terms consists of tokens $\{s_j\}$ $(0 \leq j \leq l)$, then during iteration p the following boolean search is tried (term conjunction is represented by & and term disjunction by |):

$$(m_1 \& m_2 \& ... \& m_k \& (s_1 | s_2 | ... | s_p))$$

If the number of passages returned in iteration p is greater than 20, then the searching procedure stops. This procedure on the one hand limits the scope of the main term at each step by conjoining it with a secondary term and on the other hand broadens its scope by disjoining it with additional secondary terms. In a sense this approach is similar to the narrowing and broadening techniques used in the MURAX system (Kupiec 1993). In our approach, the order in which the secondary terms are used to expand the query reflects their association with the main term, thus this search procedure is expected to retrieve good passages. In this document retrieval mode passages consist of single sentences.

Passage analysis and definition extraction

More than one definition can be extracted in this task not only because different fragments would cover different aspects (essential vs. non-essential) of the definiendum (i.e.,

Definiendum	WordNet	Britannica	Web
Aum Shirikyo	*nothing found*	*nothing found*	group; groups; cult; religious; japanese; stages; terrorist; early; doomsday; etc.
aspirin	analgesic; anti-inflammatory; antipyretic; drug, etc.	inhibit; prostaglandin; ketoprofen; synthesis; etc.	drugs; drug; blood; ibuprofen; medication; pain; etc.

Table 2: The first column contains the term sought. The second column shows terms associated with the definiendum found in WordNet. The third column contains terms associated with the definiendum found in the Britannica site. The last column contains the terms found on the Web.

aspirin is a drug; aspirin is a blood thinner) but also because the definiendum can be ambiguous (i.e., "battery" has 7 definitions in WordNet). 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 instantiated definition "seed" 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 (this heuristic has been decided upon after several experiments on training data).

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 "Andrew Carnegie" extracted from the sentence:

> *In a question-and-answer session after the panel discussion, Clinton cited philanthropists from an earlier era such as Andrew Carnegie, J.P. Morgan, and John D. Rockefeller, who, he said, "great monuments to our culture – the great museums, the great public, the great libraries.*"

is

> *philanthropists from an earlier era such as Andrew Carnegie, J.P. Morgan, and John D. Rockefeller, who, he said, "great monuments to our culture – the great museums, the great public, the great libraries.*

Note that the sentence contains the key secondary term "philanthropist" that constitutes a required piece of information for the definition.

One of the problems faced when extracting definitions from huge text collections is that of redundancy: in newspaper collections the same piece of information appears many times either in the same or different form. We address this problem by measuring the similarity of a candidate definition to a previously extracted definition from the collection in a way similar but simpler to maximal marginal relevance (Carbonell & Goldstein 1998) trying to cover many different aspects of the definiendum. A vector representation of the extracted candidate definition, consisting of its terms and term frequencies, is created. If the current answer candidate is too similar to any of the previous extracted answers, it is ignored. The similarity score is the cosine between the two vector representations, and two vectors v_1 and v_2 are considered similar if $cosine(v_1, v_2) > threshold$ (the $threshold$ was set after a number of experiments on a training set). Semantic similarity is not considered in this approach.

Evaluation

The system described here participated in the TREC QA 2003 competition that included a definition subtask. The subtask required finding answers for 50 definition questions (two examples are shown in Table 3). The set consisted of 30 "Who" definition questions and 20 "What" definition questions. NIST 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 (See Figure 3 for formulas). These scores are combined in the F-score measure with recall five times as important as precision. We obtained a combined F-score of 0.171 relying on the boolean search document retrieval system and an F-score of 0.236 when using the Okapi document retrieval system. The F-score of the systems that participated in the competition (54 runs) is 0.555 (best), 0.192 (median), 0.000 (worst); placing our definition system either above or very close to the median.

Number	Question
1901	Who is Aaron Copland?
1905	What is a golden parachute?

Table 3: Sample questions for TREC QA 2003

Failure analysis

In 5 cases (10%), the definiendum extraction component extracted an incorrect term, for example for the question "Who was Abraham in the Old Testament" the system extracted the definiendum "Abraham in the Old Testament" instead of "Abraham". This seriously affects the performance of the extraction system because of the particular way in which patterns are created with the definiendum.

$$NR = \frac{\text{number of essential nuggets returned}}{\text{number of essential nuggets}}$$

$$\text{allowance} = 100 * \text{number of essential or non-essential nuggets returned}$$

$$\text{length} = \text{number of non-white-space characters in the answer set}$$

$$NP = 1 \text{: if length} < \text{allowance}$$

$$NP = 1 - \frac{\text{length-allowance}}{\text{length}} \text{: if length} \geq \text{allowance}$$

$$F = \frac{26 * NP * NR}{25 * NP + NR}$$

Figure 3: Scoring formulas

The submission that used the Okapi document retrieval system contained response sets for 28 of the 50 questions (56%); for the other 22 the algorithm did not identify any definition-bearing sentences in the 20 passages examined. From the 28 response sets, 23 (82%) contained at least one definition nugget and all of them contained at least one essential nugget. The system answered 12 "Who" questions (40% of this type) and 11 "What" questions (55% of this type).

The submission that used the boolean document retrieval system also retrieved 28 answer sets for 50 questions (56%), though not for the same questions. From these, 20 (71%) contained at least one definition nugget, and 19 of them (95%) contained at least one essential nugget. The system answered 11 "Who" questions (36% of this type) and 9 "What" questions (45% of this type).

Knowledge acquisition analysis

We turn now to the analysis of the knowledge acquisition step of the definition system. WordNet provided relevant secondary terms in only 4 cases (8%) (by "relevant" here we mean that the term occurred in, or was judged to be closely related to, one of the gold standard nuggets supplied by NIST). The Britannica website helped with relevant secondary terms in only 5 cases (10%), however in two of these cases the terms have only one occurrence and hence were not considered for query expansion. The Web helped find relevant secondary terms in 39 cases (78%). In cases where the knowledge acquisition system failed to find relevant terms or found only a few irrelevant terms, the answer extraction system also failed. This is due to the filters that are imposed on sentences in order to be considered for extraction. We discovered that, when extracting secondary terms from sentences, we omitted the extraction of proper nouns. This has great impact on the discovery of relevant secondary terms not only for defining people (i.e., "Hillary Clinton" is a good secondary term for "Bill Clinton") but also for defining common things (i.e., "Bayer" is a good secondary term for "aspirin"). In the case of "Who" definition questions, the patterns we used had the limitation of only considering one version of the definiendum, but in this particular case it would be useful to consider name aliases or in the case of persons to consider surnames (for example in the case of "Alberto Tomba" to have "Alberto Tomba who is" and also "Tomba who is"). Finally, the patterns as they are implemented in the current version of the system allow for little variation (i.e., "Aaron Coplan, who is..." will not match "Aaron Coplan who is" because of the comma), this has an impact on both the knowledge acquisition and extraction steps.

Lenient analysis

In the light of nuggets identified by NIST analysts, a more "lenient" and optimistic analysis can be done of the answers provided by (and omissions of) our system.

The Okapi-based submission returned no answers in 22 cases, after close examination of the documents returned by the Okapi retrieval system, we can conclude that in 15 cases (68%) the documents contain some of the nuggets required to answer the question. The system failed to return answers because the definition patterns and filters were far too restrictive to cover these definitions (recall problem), a problem that we are presently addressing by investigating ways of relaxing the criteria for selecting definition-bearing sentences. A similar analysis can be done for the boolean search engine submission.

Related work

Related to our approach is the DefScriber system (Blair-Goldensohn, McKeown, & Schlaikjer 2003) that combines world knowledge in the form of definitional predicates (genus, species, and non-specific) and data-driven statistical techniques. World knowledge about the predicates is created relying on machine learning over annotated data. Data-driven techniques including the vector space model and cosine distance are used to exploit the redundancy of information about the "definiendum" in non-definitional Web texts. We differ from this approach in two respects. First, we do not rely on annotated data to create the required knowledge; secondly, we mine information in definition-bearing passages to ensure that only terms relevant for definitions are found. Nevertheless, we employ similar statistical techniques to identify similarities across sentences.

Also related to our research is the work by (Fleischman, Hovy, & Echihabi 2003) who create pairs of concept-instances such as "Bill Clinton-president" mining newspaper articles and web documents. These pairs constitue pre-available knowledge that can be used to answer "Who is ...?" questions. They use lexico-syntactic patterns learnt from annotated data to identify such pairs. We differ in that we learn knowledge in a targeted on-line way, without relying on annotated data. The knowledge acquired can be used to answer two different types of definition questions "Who" and "What".

Our approach for the identification of answer-bearing passages is closely related to work done in corpus linguistics for the identification of terminological definitions (Pearson 1998; Sierra *et al.* 2003).

Conclusions and future work

In this paper, we have presented a method for the acquisition of knowledge for answering definitions. It delays the knowledge acquisition step until the definiendum is known and uses definition patterns to mine on-line resources in order to obtain relevant secondary terms for the definiendum.

This research contributes a viable and practical solution for definitional QA, because it relies on available on-line resources and on simple natural language techniques. The problem of definition is far more complex than the one addressed here, however. Theories and methods about definitions should be examined in more detail, as for example is done in Chaurand & Mazière (1990); this would not only shed light on a typology of definitions, contributing good lexico-syntactic patterns for identification of such constructs but would also provide the basis for exploring the linguistic characteristics that contribute to good textual definitions, a particularly important issue in "text-to-text" natural language generation.

Analysis of the results obtained during evaluation indicate many avenues for further improvements:

- more informed measures for deciding on the relevance of each definition pattern, perhaps similar to those used by Ravichandran & Hovy (2002), are required;
- extracted secondary terms could be ranked, perhaps using their IDF values, in order to help to eliminate inappropriate definition pattern matches ("aspirin is a great choice for active people").
- in order to extract better answer strings, a syntactic-based technique that prunes a parse tree could be implemented;
- in order to cross the sentence barrier, coreference information could be used in the extraction patterns;
- in order to improve recall, a relaxation of the definition criteria currently implemented or development of a back-off strategy could be considered.
- in order to cover different essential vs. non-essential characteristics of the definiendum and give structure to the list of secondary terms, clustering techniques could be explored.

Acknowledgments

This research is funded by UK EPSRC under grant number GR/R91465/01.

References

Blair-Goldensohn, S.; McKeown, K.; and Schlaikjer, A. 2003. A hybrid approach for answering definitional questions (demo). In *Processdings of the 26th ACM SIGIR Conference*. Toronto, Canada: ACM.

Brill, E.; Lin, J.; Banko, M.; Dumais, S.; and Ng, A. 2001. Data-Intensive Question Answering. In *Proceedings of the Tenth Text REtrieval Conference*.

Carbonell, J. G., and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, 335–336.

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

Cunningham, H.; Maynard, D.; Bontcheva, K.; and Tablan, V. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Fleischman, M.; Hovy, E.; and Echihabi, A. 2003. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of the ACL 2003*, 1–7. ACL.

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

Hirschman, L., and Gaizauskas, R. 2001. Natural Language Question Answering: The View From Here. *Natural Language Engineering* 7(4).

Humphreys, K.; Gaizauskas, R.; Azzam, S.; Huyck, C.; Mitchell, B.; Cunningham, H.; and Wilks, Y. 1998. Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Kupiec, J. 1993. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Research and Development in Information Retrieval*, 181–190.

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

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

Ravichandran, D., and Hovy, E. 2002. Learning Surface Text Patterns for a Question Answering System. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 41–47.

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

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

Sierra, G.; Medina, A.; Alarcón, R.; and Aguilar, C. 2003. Towards the Extraction of Conceptual Information From Corpora. In Archer, D.; Rayson, P.; Wilson, A.; and McEnery, T., eds., *Proceedings of the Corpus Linguistics 2003 Conference*, 691–697. University Centre for Computer Corpus Research on Language.