Importance of Pronominal Anaphora resolution in Question Answering systems

José L. Vicedo and **Antonio Ferrández**
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apartado 99. 03080 Alicante, Spain
{vicedo,antonio}@dlsi.ua.es

Abstract

The main aim of this paper is to analyse the effects of applying pronominal anaphora resolution to Question Answering (QA) systems. For this task a complete QA system has been implemented. System evaluation measures performance improvements obtained when information that is referenced anaphorically in documents is not ignored.

1 Introduction

Open domain QA systems are defined as tools capable of extracting the answer to user queries directly from unrestricted domain documents. Or at least, systems that can extract text snippets from texts, from whose content it is possible to infer the answer to a specific question. In both cases, these systems try to reduce the amount of time users spend to locate a concrete information.

This work is intended to achieve two principal objectives. First, we analyse several document collections to determine the level of information referenced pronominally in them. This study gives us an overview about the amount of information that is discarded when these references are not solved. As second objective, we try to measure improvements of solving this kind of references in QA systems. With this purpose in mind, a full QA system has been implemented. Benefits obtained by solving pronominal references are measured by comparing system performance with and without taking into account information referenced pronominally. Evaluation shows that solving these references improves QA performance.

In the following section, the state-of-the-art of open domain QA systems will be summarised. Afterwards, importance of pronominal references in documents is analysed. Next, our approach and system components are described. Finally, evaluation results are presented and discussed.

2 Background

Interest in open domain QA systems is quite recent. We had little information about this kind of systems until the First Question Answering Track was held in last TREC conference (TRE, 1999). In this conference, nearly twenty different systems were evaluated with very different success rates. We can classify current approaches into two groups: *text-snippet extraction* systems and *noun-phrase extraction* systems.

Text-snippet extraction approaches are based on locating and extracting the most relevant sentences or paragraphs to the query by supposing that this text will contain the correct answer to the query. This approach has been the most commonly used by participants in last TREC QA Track. Examples of these systems are (Moldovan et al., 1999) (Singhal et al., 1999) (Prager et al., 1999) (Takaki, 1999) (Hull, 1999) (Cormack et al., 1999).

After reviewing these approaches, we can notice that there is a general agreement about the importance of several Natural Language Processing (NLP) techniques for QA task. Pos-tagging, parsing and Name En-

tity recognition are used by most of the systems. However, few systems apply other NLP techniques. Particularly, only four systems model some coreference relations between entities in the query and documents (Morton, 1999)(Breck et al., 1999) (Oard et al., 1999) (Humphreys et al., 1999). As example, Morton approach models identity, definite noun-phrases and non-possessive third person pronouns. Nevertheless, benefits of applying these coreference techniques have not been analysed and measured separately.

The second group includes noun-phrase extraction systems. These approaches try to find the precise information requested by questions whose answer is defined typically by a noun phrase.

MURAX is one of these systems (Kupiec, 1999). It can use information from different sentences, paragraphs and even different documents to determine the answer (the most relevant noun-phrase) to the question. However, this system does not take into account the information referenced pronominally in documents. Simply, it is ignored.

With our system, we want to determine the benefits of applying pronominal anaphora resolution techniques to QA systems. Therefore, we apply the developed computational system, Slot Unification Parser for Anaphora resolution (SUPAR) over documents and queries (Ferrández et al., 1999). SUPAR's architecture consists of three independent modules: lexical analysis, syntactic analysis, and a resolution module for natural language processing problems, such as pronominal anaphora.

For evaluation, a standard based IR system and a sentence-extraction QA system have been implemented. Both are based on Salton approach (1989). After IR system retrieves relevant documents, our QA system processes these documents with and without solving pronominal references in order to compare final performance.

As results will show, pronominal anaphora resolution improves greatly QA systems performance. So, we think that this NLP technique should be considered as part of any open domain QA system.

3 Importance of pronominal information in documents

Trying to measure the importance of information referenced pronominally in documents, we have analysed several text collections used for QA task in TREC-8 Conference as well as others used frequently for IR system testing. These collections were the following: Los Angeles Times (LAT), Federal Register (FR), Financial Times (FT), Federal Bureau Information Service (FBIS), TIME, CRANFIELD, CISI, CACM, MED and LISA. This analysis consists on determining the amount and type of pronouns used, as well as the number of sentences containing pronouns in each of them. As average measure of pronouns used in a collection, we use the ratio between the quantity of pronouns and the number of sentences containing pronouns. This measure approximates the level of information that is ignored if these references are not solved. Figure 1 shows the results obtained in this analysis.

As we can see, the amount and type of pronouns used in analysed collections vary depending on the subject the documents talk about. LAT, FBIS, TIME and FT collections are composed from news published in different newspapers. The ratio of pronominal reference used in this kind of documents is very high (from 35,96% to 55,20%). These documents contain a great number of pronominal references in third person (he, she, they, his, her, their) whose antecedents are mainly people's names. In this type of documents, pronominal anaphora resolution seems to be very necessary for a correct modelling of relations between entities. CISI and MED collections appear ranked next in decreasing ratio level order. These collections are composed by general comments about document managing, classification and indexing and documents extracted from medical journals respectively. Although the ratio presented by these collections (24,94% and 22,16%) is also high, the most important group of pronominal references used in these collections is formed by "it" and "its" pronouns. In this case,

TEXT COLLECTION	LAT	FBIS	TIME	FT	CISI	MED	CACM	LISA	FR	CRANFIELD
Pronoun type										
HE, SHE, THEY	38,59%	29,15%	31,20%	26,20%	15,38%	15,07%	8,59%	12,24%	13,31%	6,54%
HIS, HER, THEIR	25,84%	21,54%	35,01%	20,52%	22,96%	21,46%	15,69%	31,03%	20,70%	10,35%
IT, ITS	26,92%	39,60%	22,43%	46,68%	52,11%	57,41%	67,61%	47,86%	61,06%	79,76%
HIM, THEM	7,04%	7,08%	7,82%	4,44%	6,38%	3,96%	4,87%	6,30%	3,45%	1,60%
HIM, HER,IT(SELF), THEMSELVES	1,61%	2,63%	3,54%	2,17%	3,17%	2,10%	3,25%	2,57%	1,48%	1,75%
Pronouns in Sentences										
Containing 0 pronouns	44,80%	48,09%	51,37%	64,04%	75,06%	77,84%	79,06%	83,79%	84,92%	90,95%
Containing 1 pronoun	30,40%	31,37%	29,46%	23,07%	17,17%	15,02%	17,54%	13,01%	11,64%	8,10%
Containing 2 pronouns	14,94%	12,99%	12,26%	8,54%	5,27%	4,75%	2,79%	2,56%	2,57%	0,85%
Containing +2 pronouns	9,86%	7,55%	6,90%	4,34%	2,51%	2,39%	0,60%	0,64%	0,88%	0,09%
Ratio of pronominal reference	**55,20%**	**51,91%**	**48,63%**	**35,96%**	**24,94%**	**22,16%**	**20,94%**	**16,21%**	**15,08%**	**9,05%**

Figure 1: *Pronominal references in text collections*

antecedents of these pronominal references are mainly concepts represented typically by noun phrases. It seems again important solving these references for a correct modelling of relations between concepts expressed by noun-phrases. The lowest ratio results are presented by CRANFIELD collection with a 9,05%. The reason of this level of pronominal use is due to text contents. This collection is composed by extracts of very high technical subjects. Between the described percentages we find the CACM, LISA and FR collections. These collections are formed by abstracts and documents extracted from the Federal Register, from the CACM journal and from Library and Information Science Abstracts, respectively. As general behaviour, we can notice that as more technical document contents become, the pronouns "it" and "its" become the most appearing in documents and the ratio of pronominal references used decreases. Another observation can be extracted from this analysis. Distribution of pronouns within sentences is similar in all collections. Pronouns appear scattered through sentences containing one or two pronouns. Using more than two pronouns in the same sentence is quite infrequent.

After analysing these results an important question may arise. Is it worth enough to solve pronominal references in documents? It would seem reasonable to think that resolution of pronominal anaphora would only be accomplished when the ratio of pronominal occurrence exceeds a minimum level. However, we have to take into account that the cost of solving these references is proportional to the number of pronouns analysed and consequently, proportional to the amount of information a system will ignore if these references are not solved.

As results above state, it seems reasonable to solve pronominal references in queries and documents for QA tasks. At least, when the ratio of pronouns used in documents recommend it. Anyway, evaluation and later analysis (section 5) contribute with empirical data to conclude that applying pronominal anaphora resolution techniques improve QA systems performance.

4 Our Approach

Our system is made up of three modules. The first one is a standard IR system that retrieves relevant documents for queries. The second module will manage with anaphora resolution in both, queries and retrieved documents. For this purpose we use SUPAR computational system (section 4.1). And the third one is a sentence-extraction QA system that interacts with SUPAR module and ranks sentences from retrieved documents to locate the answer where the correct answer appears (section 4.2).

For the purpose of evaluation an IR system has been implemented. This system is based on the standard information retrieval

approach to document ranking described in Salton (1989). For QA task, the same approach has been used as baseline but using sentences as text unit. Each term in the query and documents is assigned an inverse document frequency (idf) score based on the same corpus. This measure is computed as:

$$idf(t) = log(\frac{N}{df(t)}) \quad (1)$$

where N is the total number of documents in the collection and $df(t)$ is the number of documents which contains term t. Query expansion consists of stemming terms using a version of the Porter stemmer. Document and sentence similarity to the query was computed using the cosine similarity measure. The LAT corpus has been selected as test collection due to his high level of pronominal references.

4.1 Solving pronominal anaphora

In this section, the NLP Slot Unification Parser for Anaphora Resolution (SUPAR) is briefly described (Ferrández et al., 1999; Ferrández et al., 1998). SUPAR's architecture consists of three independent modules that interact with one other. These modules are lexical analysis, syntactic analysis, and a resolution module for Natural Language Processing problems.

Lexical analysis module. This module takes each sentence to parse as input, along with a tool that provides the system with all the lexical information for each word of the sentence. This tool may be either a dictionary or a part-of-speech tagger. In addition, this module returns a list with all the necessary information for the remaining modules as output. SUPAR works sentence by sentence from the input text, but stores information from previous sentences, which it uses in other modules, (e.g. the list of antecedents of previous sentences for anaphora resolution).

Syntactic analysis module. This module takes as input the output of lexical analysis module and the syntactic information represented by means of grammatical formalism Slot Unification Grammar (SUG). It returns what is called slot structure, which stores all necessary information for following modules. One of the main advantages of this system is that it allows carrying out either partial or full parsing of the text.

Module of resolution of NLP problems. In this module, NLP problems (e.g. anaphora, extra-position, ellipsis or PP-attachment) are dealt with. It takes the slot structure (SS) that corresponds to the parsed sentence as input. The output is an SS in which all the anaphors have been resolved. In this paper, only pronominal anaphora resolution has been applied.

The kinds of knowledge that are going to be used in pronominal anaphora resolution in this paper are: pos-tagger, partial parsing, statistical knowledge, c-command and morphologic agreement as restrictions and several heuristics such as syntactic parallelism, preference for noun-phrases in same sentence as the pronoun preference for proper nouns.

We should remark that when we work with unrestricted texts (as it occurs in this paper) we do not use semantic knowledge (i.e. a tool such as WorNet). Presently, SUPAR resolves both Spanish and English pronominal anaphora with a success rate of 87% and 84% respectively.

SUPAR pronominal anaphora resolution differs from those based on restrictions and preferences, since the aim of our preferences is not to sort candidates, but rather to discard candidates. That is to say, preferences are considered in a similar way to restrictions, except when no candidate satisfies a preference, in which case no candidate is discarded. For example in sentence: *"Rob was asking us about John. I replied that Peter saw John yesterday. James also saw him."* After applying the restrictions, the following list of candidates is obtained for the pronoun *him*: [*John, Peter, Rob*], which are then sorted according to their proximity to the anaphora. If preference for candidates in same sentence as the anaphora is applied, then no candidate satisfies it, so the following preference is applied on the same list of candidates. Next, preference for candidates in the previous sentence is applied and the list is reduced to the following

candidates: [*John, Peter*]. If syntactic parallelism preference is then applied, only one candidate remains, [*John*], which will be the antecedent chosen.

Each kind of anaphora has its own set of restrictions and preferences, although they all follow the same general algorithm: first come the restrictions, after which the preferences are applied. For pronominal anaphora, the set of restrictions and preferences that apply are described in Figure 2.

```
Procedure SelectingAntecedent ( INPUT L: ListOfCandidates,
                                OUTPUT Solution: Antecedent )
    Apply restrictions to L with a result of L1
        Morphologic agreement
        C-command constraints
        Semantic consistency
    Case of:
        NumberOfElements (L1) = 1
            Solution = TheFirstOne (L1)
        NumberOfElements (L1) = 0
            Exophora or cataphora
        NumberOfElements (L1) > 1
            Apply preferences to L1 with a result of L2
                1) Candidates in the same sentence as anaphor.
                2) Candidates in the previous sentence
                3) Preference for proper nouns.
                4) Candidates in the same position as the anaphor
                with reference to the verb (before or after).
                5) Candidates with the same number of parsed
                constituents as the anaphora
                6) Candidates that have appeared with the verb of
                the anaphor more than once
                7) Preference for indefinite NPs.

            Case of:
                NumberOfElements (L2) = 1
                    Solution = TheFirstOne (L2)
                NumberOfElements (L2) > 1
                    Extract from L2 in L3 those candidates that have
                    been repeated most in the text

                    If NumberOfElements (L3) > 1
                        Extract from L3 in L4 those candidates that
                        have appeared most with the verb of the
                        anaphora

                        Solution = TheFirstOne (L4)
                    Else
                        Solution = TheFirstOne (L3)
                    EndIf
            EndCase
    EndCase
EndProcedure
```

Figure 2: *Pronominal anaphora resolution algorithm*

The following restrictions are first applied to the list of candidates: morphologic agreement, c-command constraints and semantic consistency. This list is sorted by proximity to the anaphor. Next, if after applying restrictions there is still more than one candidate, the preferences are then applied, in the order shown in this figure. This sequence of preferences (from *1* to *7*) stops when, after having applied a preference, only one candidate remains. If after applying preferences there is still more than one candidate, then the most repeated candidates[1] in the text are extracted from the list after applying preferences. After this is done, if there is still more than one candidate, then those candidates that have appeared most frequently with the verb of the anaphor are extracted from the previous list. Finally, if after having applied all the previous preferences, there is still more than one candidate left, the first candidate of the resulting list, (the closest one to the anaphor), is selected.

4.2 Anaphora resolution and QA

Our QA approach provides a second level of processing for relevant documents: Analysing matching documents and Sentence ranking.

Analysing Matching Documents. This step is applied over the best matching documents retrieved from the IR system. These documents are analysed by SUPAR module and pronominal references are solved. As result, each pronoun is associated with the noun phrase it refers to in the documents. Then, documents are split into sentences as basic text unit for QA purposes. This set of sentences is sent to the sentence ranking stage.

Sentence Ranking. Each term in the query is assigned a weight. This weight is the sum of inverse document frequency measure of terms based on its occurrence in the LAT collection described earlier. Each document sentence is weighted the same way. The only difference with baseline is that pronouns are given the weight of the entity they refer to. As we only want to analyse the effects of pronominal reference resolution, no more changes are introduced in weighting scheme. For sentence ranking, cosine similarity is used between query and document sentences.

5 Evaluation

For this evaluation, several people unacquainted with this work proposed 150 queries

[1]Here, we mean that firstly we obtain the maximum number of repetitions for an antecedent in the remaining list. After that, we extract from that list the antecedents that have this value of repetition.

whose correct answer appeared at least once into the analysed collection. These queries were also selected based on their expressing the user's information need clearly and their being likely answered in a single sentence.

First, relevant documents for each query were retrieved using the IR system described earlier. Only the best 50 matching documents were selected for QA evaluation. As the document containing the correct answer was included into the retrieved sets for only 93 queries (a 62% of the proposed queries), the remaining 57 queries were excluded for this evaluation.

Once retrieval of relevant document sets was accomplished for each query, the system applied anaphora resolution algorithm to these documents. Finally, sentence matching and ranking was accomplished as described in section 4.2 and the system presented a ranked list containing the 10 most relevant sentences to each query.

For a better understanding of evaluation results, queries were classified into three groups depending on the following characteristics:

- Group A. There are no pronominal references in the target sentence (sentence containing the correct answer).
- Group B. The information required as answer is referenced via pronominal anaphora in the target sentence.
- Group C. Any term in the query is referenced pronominally in the target sentence.

Group A was made up by 37 questions. Groups B and C contained 25 and 31 queries respectively. Figure 3 shows examples of queries classified into groups B and C.

Evaluation results are presented in Figure 4 as the number of target sentences appearing into the 10 most relevant sentences returned by the system for each query and also, the number of these sentences that are considered a correct answer. An answer is considered correct if it can be obtained by simply looking at the target sentence. Results

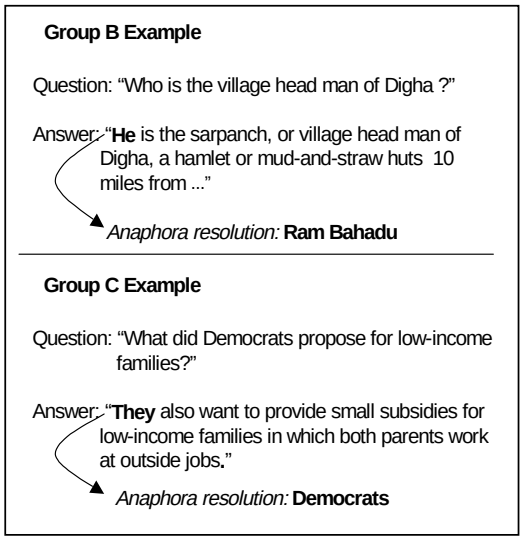

Figure 3: *Group B and C query examples*

are classified based on question type introduced above. The number of queries pertaining to each group appears in the second column. Third and fourth columns show baseline results (without solving anaphora). Fifth and sixth columns show results obtained when pronominal references have been solved.

Results show several aspects we have to take into account. Benefits obtained from applying pronominal anaphora resolution vary depending on question type. Results for group A and B queries show us that relevance to the query is the same as baseline system. So, it seems that pronominal anaphora resolution does not achieve any improvement. This is true only for group A questions. Although target sentences are ranked similarly, for group B questions, target sentences returned by baseline can not be considered as correct because we do not obtain the answer by simply looking at returned sentences. The correct answer is displayed only when pronominal anaphora is solved and pronominal references are substituted by the noun phrase they refer to. Only if pronominal references are solved, the user will not need to read more text to obtain the correct answer. For noun-phrase extraction QA systems the improvement is greater. If pronominal references are not solved, this information will

Answer Type	Number		Baseline				Anaphora solved			
			Target included		Correct answer		Target included		Correct answer	
A	37	(39,78%)	18	(48,65%)	18	(48,65%)	18	(48,65%)	18	(48,65%)
B	25	(26,88%)	12	(48,00%)	0	(0,00%)	12	(48,00%)	12	(48,00%)
C	31	(33,33%)	9	(29,03%)	9	(29,03%)	21	(67,74%)	21	(67,74%)
A+B+C	**93**	**(100,00%)**	**39**	**(41,94%)**	**27**	**(29,03%)**	**51**	**(54,84%)**	**51**	**(54,84%)**

Figure 4: *Evaluation results*

not be analysed and probably a wrong noun-phrase will be given as answer to the query.

Results improve again if we analyse group C queries performance. These queries have the following characteristic: some of the query terms were referenced via pronominal anaphora in the relevant sentence. When this situation occurs, target sentences are retrieved earlier in the final ranked list than in the baseline list. This improvement is because similarity increases between query and target sentence when pronouns are weighted with the same score as their referring terms. The percentage of target sentences obtained increases 38,71 points (from 29,03% to 67,74%).

Aggregate results presented in Figure 4 measure improvement obtained considering the system as a whole. General percentage of target sentences obtained increases 12,90 points (from 41,94% to 54,84%) and the level of correct answers returned by the system increases 25,81 points (from 29,03% to 54,84%).

At this point we need to consider the following question: Will these results be the same for any other question set? We have analysed test questions in order to determine if results obtained depend on question test set. We argue that a well-balanced query set would have a percentage of target sentences that contain pronouns (PTSC) similar to the pronominal reference ratio of the text collection that is being queried. Besides, we suppose that the probability of finding an answer in a sentence is the same for all sentences in the collection. Comparing LAT ratio of pronominal reference (55,20%) with the question test set PTSC we can measure how a question set can affect results. Our question set PTSC value is a 60,22%. We obtain as target sentences containing pronouns only a 5,02% more than expected when test queries are randomly selected. In order to obtain results according to a well-balanced question set, we discarded five questions from both groups B and C. Figure 5 shows that results for this well-balanced question set are similar to previous results. Aggregate results show that general percentage of target sentences increases 10,84 points when solving pronominal anaphora and the level of correct answers retrieved increases 22,89 points (instead of 12,90 and 25,81 obtained in previous evaluation respectively).

As results show, we can say that pronominal anaphora resolution improves QA systems performance in several aspects. First, precision increases when query terms are referenced anaphorically in the target sentence. Second, pronominal anaphora resolution reduces the amount of text a user has to read when the answer sentence is displayed and pronominal references are substituted with their coreferent noun phrases. And third, for noun phrase extraction QA systems it is essential to solve pronominal references if a good performance is pursued.

6 Conclusions and future research

The analysis of information referenced pronominally in documents has revealed to be important to tasks where high level of recall is required. We have analysed and measured the effects of applying pronominal anaphora resolution in QA systems. As results show, its application improves greatly QA performance and seems to be essential in some cases.

Three main areas of future work have appeared while investigation has been developed. First, IR system used for retrieving relevant documents has to be adapted for QA

Answer Type	Number		Baseline				Anaphora solved			
			Target included		Correct answer		Target included		Correct answer	
A	37	(39,78%)	18	(48,65%)	18	(48,65%)	18	(48,65%)	18	(48,65%)
B	20	(21,51%)	10	(50,00%)	0	(0,00%)	10	(50,00%)	10	(50,00%)
C	26	(27,96%)	9	(34,62%)	9	(34,62%)	18	(69,23%)	18	(69,23%)
A+B+C	**83**	**(89,25%)**	**37**	**(44,58%)**	**27**	**(32,53%)**	**46**	**(55,42%)**	**46**	**(55,42%)**

Figure 5: *Well-balanced question set results*

tasks. The IR used, obtained the document containing the target sentence only for 93 of the 150 proposed queries. Therefore, its precision needs to be improved. Second, anaphora resolution algorithm has to be extended to different types of anaphora such as definite descriptions, surface count, verbal phrase and one-anaphora. And third, sentence ranking approach has to be analysed to maximise the percentage of target sentences included into the 10 answer sentences presented by the system.

References

Eric Breck, John Burger, Lisa Ferro, David House, Marc Light, and Inderjeet Mani. 1999. A Sys Called Quanda. In *Eighth Text REtrieval Conference* (TRE, 1999).

Gordon V. Cormack, Charles L. A. Clarke, Christopher R. Palmer, and Derek I. E. Kisman. 1999. Fast Automatic Passage Ranking (MultiText Experiments for TREC-8). In *Eighth Text REtrieval Conference* (TRE, 1999).

Antonio Ferrández, Manuel Palomar, and Lidia Moreno. 1998. Anaphora resolution in unrestriced texts with partial parsing. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Lingustics COLING-ACL.*

Antonio Ferrández, Manuel Palomar, and Lidia Moreno. 1999. An empirical approach to Spanish anaphora resolution. *To appear in Machine Translation.*

David A. Hull. 1999. Xerox TREC-8 Question Answering Track Report. In *Eighth Text REtrieval Conference* (TRE, 1999).

Kevin Humphreys, Robert Gaizauskas, Mark Hepple, and Mark Sanderson. 1999. University of Sheffield TREC-8 Q&A System. In *Eighth Text REtrieval Conference* (TRE, 1999).

Julian Kupiec, 1999. *MURAX: Finding and Organising Answers from Text Search*, pages 311–331. Kluwer Academic, New York.

Dan Moldovan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Richard Goodrum, Roxana Gîrju, and Vasile Rus. 1999. LASSO: A Tool for Surfing the Answer Net. In *Eighth Text REtrieval Conference* (TRE, 1999).

Thomas S. Morton. 1999. Using Coreference in Question Answering. In *Eighth Text REtrieval Conference* (TRE, 1999).

Douglas W. Oard, Jianqiang Wang, Dekang Lin, and Ian Soboroff. 1999. TREC-8 Experiments at Maryland: CLIR, QA and Routing. In *Eighth Text REtrieval Conference* (TRE, 1999).

John Prager, Dragomir Radev, Eric Brown, Anni Coden, and Valerie Samn. 1999. The Use of Predictive Annotation for Question Answering. In *Eighth Text REtrieval Conference* (TRE, 1999).

Gerard A. Salton. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison Wesley, New York.

Amit Singhal, Steve Abney, Michiel Bacchiani, Michael Collins, Donald Hindle, and Fernando Pereira. 1999. ATT at TREC-8. In *Eighth Text REtrieval Conference* (TRE, 1999).

Toru Takaki. 1999. NTT DATA: Overview of system approach at TREC-8 ad-hoc and question answering. In *Eighth Text REtrieval Conference* (TRE, 1999).

TREC-8. 1999. *Eighth Text REtrieval Conference.*