EVENT-BASED TEXTUAL DOCUMENT RETRIEVAL BY USING SEMANTIC ROLE LABELING AND COREFERENCE RESOLUTION

[1, 2]Chia-Hung Lin, [1]Chia-Wei Yen, [1,]*Jen-Shin Hong, and [2]Samuel Cruz-Lara
[1]Department of Computer Science and Information Engineering, National ChiNan University, Taiwan
[2]LORIA / INRIA Lorraine, Nancy-Université, France
**Corresponding Author: E-mail jshong@ncnu.edu.tw*

ABSTRACT

Conventional keyword-based indexing and retrieval techniques for textual documents lack of precision when a long query string is employed in order to discover documents containing a specific “event”, such as “Einstein discovered relativity”. This paper proposes a framework to resolve such a problem. In our proposal, we apply semantic role labeling and coreference techniques in order to parse each sentence within textual documents into three elements: subject, object and predicates. These elements can subsequently be used for indexing and retrieval. Our primitive evaluation experiments have shown that this promising methodology raises the retrieval precision if we compared it to conventional literal term-matching techniques.

KEYWORDS

Event-based textual document retrieval, semantic role labeling, coreference resolution.

1. INTRODUCTION

Textual document indexing and retrieval from digital libraries have been extensively studied for decades. In the literature, there has been a broad variety of approaches proposing end users to retrieve text-based documents. In order to retrieve textual documents, the most popular approach is to rely on features existing within the document content, particularly the “terms”. Conventionally, there are three major categories for text-based document retrieval: the Boolean Model, the Vector Space Model, and the Probability Model.

Within the Boolean Model (Salton et al. 1983), documents and queries are represented as sets of indexed terms which are simply all words appearing within the text of a document in the collection. A query is next specified as boolean expressions with “and”, “or”, and “not” operations. Index terms are considered to be either present or absent within a document and to provide equal evidence with respect to information needs. Term frequency counts within the documents are not considered. Also, no notation of a partial match and ranking to the query condition are provided. Such an approach has been widely used in early Web search engines, such as “AltaVista”, “Excite”, and “WebCrawler”, where mostly on-page data (text and formatting) are parsed and indexed so that conventional “literal term-matching” can be used to access the documents.

Within the Vector Space Model (Salton et al. 1983; Wong et al. 1985), documents and queries are represented as vectors in a high dimensional space in which each element of the vector represents the frequency of the word in the document or the query. Such an approach supports term-weighting and partial matching. One potential problem here is to find a good set of basis vectors, as well as a good weighting scheme for terms.

Finally, within the Probabilistic Model (Fuhr 1992), documents and queries are represented by using a probability-based model. Retrieval is modeled as a classification process. There are two classes for each query: the relevant or non-relevant documents. All documents are ranked by using a decreasing order of probability of relevance. The challenge here is to estimate the characteristics of the relevant class/set without any training data, in the form of user-identified examples of relevant documents.

To further improve the content-based ranking of the search engine, various recent web search techniques use off-page, web-specific data as link analysis, anchor-text, and click-through data. Such approaches support both informational and navigational queries. At this stage, all major engines use all these types of data.

Overall speaking, these document retrieval techniques are based on word and/or phrase analysis of the texts in the contents, links, and/or anchor texts of a document. In principle, the statistical analysis of a term (word or phrase) frequency captures the importance of the term within a document. Retrieval is based on queries consisting of index terms. They tend to assume mutual independence of the indexed terms and the semantics of the documents. The information a user needs can be expressed through combinations of such terms, so relevance can be interpreted in terms of indexed terms and semantics. However, within this approach, semantic is often lost when expressed through sets of words. A clear example is that, quite often, a user would be interested in a query like: find all documents describing the event, “Einstein, won, Nobel Prize”. An intuitive query formulation from a native user to a Web search engine would be most likely a query “with all the words” of *Einstein, won, Nobel Prize,* which unfortunately retrieve many irrelevant documents. These irrelevant documents may contain the three terms, but they will frequently be dispersed throughout the documents in totally different context. A more sophisticated user could possibly formulate a query “*with the exact phrase*” of “*Einstein won a Nobel Prize,*” retrieves documents containing the exact phrase that has been literally formulated as the query string. In such a case, the grammatical structure should be perfectly matched to the query string. However, many relevant documents, such as within a document containing two neighboring sentences describing the desired event, would be missing. For example, in the above query scenario, a document containing two sentences, such as “*Einstein was born in Munich*“ and “*He won a Nobel Prize in 1921*”, which clearly fit to the user’s need but they would not be able to be retrieved with the “exact phrase” query.

For such type of “event-based queries”, in order to achieve a more accurate analysis, the underlying data representation should capture the semantics of the texts. In this paper, we propose an indexing and retrieval framework for this kind of event-based queries. In the following section, we present a methodology based on state-of-the-art for natural language processing technologies, namely, the semantic role labeling and the coreference techniques.

2. METHODOLOGY

In the following, we will first briefly describe the essential technologies applied. Next, we will elaborate the underlying principles and the detailed operations for each stage of the proposed methodology.

2.1 Semantic Role Labeling (SRL) Techniques

Roughly speaking, in a sentence, a verb (predicate) indicates an event. A semantic role is the relationship that a syntactic argument has with the verb. One of the most commonly-used schemes for specifying the semantic roles are proposed to construct a large-scale corpus - the PropBank (Kingsbury and Palmer 2002; Palmer 2005). In PropBank, the arguments of a verb are labeled sequentially from ARG0 to ARG5, where ARG0 is usually the subject of a transitive verb; ARG1, is the direct object, etc. A variety of adjunctive arguments, such as ARGM-LOC, for locatives, and ARGM-TMP, for time, are also tagged. As an illustrative example, the semantic roles for the sentence “I saw a girl in the park in the morning” based on the PropBank style markup are given as:

[$_{\text{ARG0}}$ I] [$_{\text{Target}}$ saw] [$_{\text{ARG1}}$ a girl] [$_{\text{ARGM-Loc}}$ in the park] [$_{\text{ARGM-Tmp}}$ in the morning]

Semantic Role Labeling techniques automatically identify the semantic roles of a sentence. In the literature, there are several studies proposing different methodologies for such purposes, for example, (Gildea and Jurafsky 2002; Pradhan et al. 2004; Koomen et al. 2005), etc. These methodologies have obtained accurate results about 80% on ARG0, ARG1, and 70% on ARGM-LOC, ARGM-TMP, for a set of sample data coming from the Wall Street Journal (Pradhan et al. 2004). In (Lin et al. 2007), we have applied these SRL techniques to detect event-based knowledge in digital image descriptions with satisfactory precision.

2.2 Coreference Resolution

In linguistics, coreference occurs when different expressions in a sentence or contextual sentences refer to the same entity in the real world. Two expressions (noun phrases or pronouns) are said to be co-referring to each other if both of them resolve to a unique entity (i.e., the referent) unambiguously. For example, in the sentences “Leonardo da Vinci was one of the greatest painters of the Italian Renaissance. He left only a handful of completed paintings”, the “Leonardo da Vinci” and “he” are most likely coreferent. In a typical documents, quite often a complete event is expressed within surrounding contextual sentences. The coreference needs to be resolved automatically to identify which entity a noun phrase or pronoun actually refers to.

Coreference resolution is the task of resolving noun phrases or pronouns to the entities that they refer to. This has been an active research topic in natural language processing for decades. The coreference resolution techniques are widely used in areas such as named entity extraction, question answering, machine translation and so on. In the literature, quite a number of methodologies have been proposed for solving the coreference resolution. Most early attempts heavily rely on linguistic and domain knowledge (e.g., Hobbs, 1986). On the other hand, many recent approaches apply various machine learning techniques with sophisticated parsers and taggers (e.g., Ng and Cardie 2002; Kehler et al. 2004; Ponzetto and Strube 2006). Readers can refer to (Elango 2007) for an extensive survey on relevant studies.

2.3 Processes

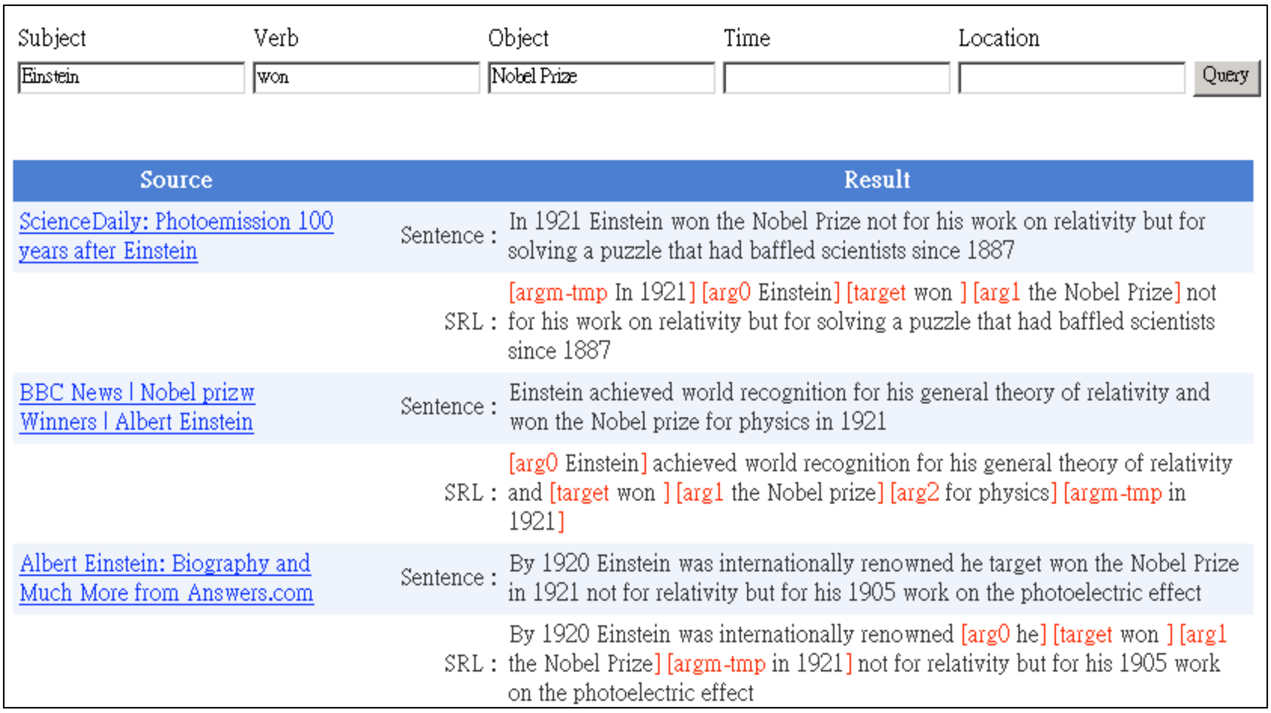
The overall process includes two major stages:

1. **Recognition of the semantic roles using SRL and coreference resolution tools**

In the first stage, the sentences in the documents are processed using SRL tools to identify the semantic roles of each sentence. For sentences with subject or object expressed by a pronoun, the exact entity referred is obtained using the coreference resolution techniques. The resolved semantic roles are indexed in the database for field-based query.

2. **Retrieval based on conjunction of subject, object, and predicate**

An event-based query input interface is provided (see Figure 1) in order to allow end users to query the images in an event-based scenario. The interface guides the users to decompose a query string of an event into five semantic roles, namely, subject, verb, object, time, and location.



Subject	Verb	Object	Time	Location
Einstein	won	Nobel Prize		

Source		Result
ScienceDaily: Photoemission 100 years after Einstein	Sentence :	In 1921 Einstein won the Nobel Prize not for his work on relativity but for solving a puzzle that had baffled scientists since 1887
	SRL :	[argm-tmp In 1921] [arg0 Einstein] [target won] [arg1 the Nobel Prize] not for his work on relativity but for solving a puzzle that had baffled scientists since 1887
BBC News | Nobel prizw Winners | Albert Einstein	Sentence :	Einstein achieved world recognition for his general theory of relativity and won the Nobel prize for physics in 1921
	SRL :	[arg0 Einstein] achieved world recognition for his general theory of relativity and [target won] [arg1 the Nobel prize] [arg2 for physics] [argm-tmp in 1921]
Albert Einstein: Biography and Much More from Answers.com	Sentence :	By 1920 Einstein was internationally renowned he target won the Nobel Prize in 1921 not for relativity but for his 1905 work on the photoelectric effect
	SRL :	By 1920 Einstein was internationally renowned [arg0 he] [target won] [arg1 the Nobel Prize] [argm-tmp in 1921] not for relativity but for his 1905 work on the photoelectric effect

Figure 1. The event-based query input interface

3. PRIMITIVE EVALUATION AND DISCUSSION

To verify the applicability of the proposed methodology in real-life applications, we have conducted a primitive evaluation experiment. We use an automatic semantic labeling engine - ASSERT (http://oak.colorado.edu/assert/) to parse the semantic roles of sentences in textual passages. For the coreference resolution, we applied the “Gate tool” (Dimitrov et al. 2002). The approaches of Gate can be found in [http://gate.ac.uk/]. An experiment was conducted to compare the performance of keyword-based approach, SRL-based approach, and coreference resolve SRL-based approach. First, we collect a large set of documents containing a keyword “Einstein”. The textual data were parsed to extract the event-related knowledge, sentence-by-sentence, using ASSERT and Gate. The parsed semantic roles for each sentence were managed in a database.

Two human subjects judged the relevance of each retrieved document to the event queried, based on whether a desired event is described in the document or not. Throughout the evaluations, we measured the retrieval effectiveness using the precision rate. The number of the relevant retrieved items is also given in order to provide an idea on the recall capabilities for the different approaches. In the experiments, four selected events, as listed in Table 1, were used in order to query the documents containing the term, “Einstein”. Table 1 shows the results for three different approaches.

These results indicate that the precision rates of the SRL-assisted query were very high, but the numbers of the retrieved relevant items were fewer than the keyword-based approach. This is natural, since there are many documents in which keywords appear in different sentences but describe irrelevant scenes. The queries based on “coreference resolved SRL-assisted” approaches slightly reduce the precision rate of the SRL-assisted queries. However, the correct items retrieved increased slightly. It indicates that the coreference resolution can solve situations where an event is implicitly dispersed in a number of neighboring sentences. With such settings, the precision of the keyword-based approach was significantly lower than that of the SRL-assisted approach.

Overall speaking, these results indicate that the proposed coreference-solved SRL-based methodology has a good potential for the applications of event queries (see Table 1).

Table 1. Results of the primitive evaluation experiments

Query Events			Keyword-based query		SRL-assisted query		Coreference-resolved SRL-assisted query	
Subject	Verb	Object	Precision	Correct documents retrieved	Precision	Correct documents retrieved	Precision	Correct documents retrieved
Einstein	Win	Nobel Prize	73% (61/84)	61	98% (49/50)	49	95% (55/58)	55
Einstein	Discover	Relativity	39% (12/31)	12	80% (4/5)	4	80% (4/5)	4
Einstein	Study	Math	34% (10/29)	10	100% (6/6)	6	100% (9/9)	9
Einstein	Teach	Math	18% (2/11)	2	100% (1/1)	1	100% (2/2)	2
			Average Precision 55%	Total correct document retrieved:85	Average Precision: 97%	Total correct documents retrieved:60	Average Precision: 95%	Total correct documents retrieved:70

4. CONCLUSION

Conventional keyword-based indexing and retrieval techniques for textual documents have poor precision when a long query string is employed in order to discover documents containing a specific "event". In this paper, we have proposed a promising methodology based on semantic role labeling and coreference techniques allowing to parse each sentence within textual documents into three elements: subject, object and predicate. These elements can subsequently be used in order to perform indexing and retrieval.

As we have shown, a primitive evaluation of this methodology has been performed. The corresponding results show that the precision rate of SRL-assisted event-based query appears to be much higher than the keyword-based approach in particular within some situations where an event is implicitly dispersed in a number of neighboring sentences.

ACKNOWLEDGEMENT

This work is supported by the National Science Council of Taiwan, under the grant numbered as NSC-94-2422-H-260-002.

REFERENCES

Dimitrov, M. et al, 2002. A Light-weight Approach to Coreference Resolution for Named Entities in Text. *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*. Lisbon.

Elango, P., 2007. Coreference Resolution: A Survey. available at: *http://www.cs.wisc.edu/~apirak/cs/cs838/pradheep-survey.pdf* . last visited July 15th 2007.

Fuhr, N., 1992. Probabilistic Models in Information Retrieval. *The Computer Journal*. Vol. 35, No. 3, pp. 243-255.

Gildea, D. and Jurafsky, D., 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*. Vol. 28, No. 3, pp. 245-288.

Hobbs, J. R., 1986. Resolving Pronoun References. *Readings in Natural Language Processing*. Morgan Kaufmann, Los Altos, California, pp. 339-352.

Kehler, A. et al, 2004. The (non)utility of Predicate-argument Frequencies for Pronoun Interpretation. *Proceeding of 2004 North American Chapter of the Association for Computational Linguistics Annual Meeting*. pp. 289-296.

Kingsbury, P. and Palmer, M., 2002. From Treebank to Propbank. *Proceedings of the LREC*.

Koomen, P. et al, 2005. Generalized Inference with Multiple Semantic Role Labeling Systems. *Proceedings of the 9th Conference on Computational Natural Language Learning*. CoNLL-2005, pp. 181-184.

Lin et al. 2007. Semantic Role Labeling Techniques for Event-based Knowledge Extraction from Free-text Descriptions for Art Images. *The Electronic Library*, v26, n1, scheduled to appear in April 2008.

Ng, V. and Cardie, C., 2002. Improving Machine Learning Approaches to Coreference Resolution. *In 40th Anniversary Meeting of the Association for Computational Linguistics*. ACL-02, pp. 104-111.

Pradhan, S. et al, 2004. Shallow Semantic Parsing Using Support Vector Machines. *In Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting*. pp. 233-240.

Ponzetto, S. P. and Strube, M., 2006. Semantic Role Labeling for Coreference Resolution. *In Companion Volume of the Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics*. pp. 143-146.

Robertson, S. E., 1977. The Probability Ranking Principle in IR. *Journal of Document*. Vol. 33, No. 4, pp. 294-304.

Salton, G. et al, 1983. Extended Boolean Information Retrieval. *Communications of the ACM*. Vol. 26, No. 11, pp. 1022-1036.

Salton, G. et al, 1975. A Vector Space Model for Information Retrieval. *Communications of the ACM*. Vol. 18, No. 11, pp. 613-620.

Wong, S. K. M. et al, 1985. Generalized Vector Space Model in Information Retrieval. *In ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'85, pp. 18-25.

Palmer, M., Kingsbury, P. and Gildea, D. (2005), *"The Proposition Bank: An Annotated Corpus of Semantic Roles". Computational Linguistics,* Vol. 31 No. 1, pp. 71-106.