Analysis and Modeling of Effective Passage Retrieval Mechanisms in Question Answering Systems

Ortiz-Arroyo, Daniel

Published in:
3rd International Conference on Electrical and Electronics Engineering, 2006

Publication date:
2006

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Ortiz-Arroyo, D. (2006). Analysis and Modeling of Effective Passage Retrieval Mechanisms in Question Answering Systems. In *3rd International Conference on Electrical and Electronics Engineering, 2006* Electrical Engineering/Electronics, Computer, Communications and Information Technology Association.

Analysis and Modeling of Effective Passage Retrieval Mechanisms in QAS

Hans Ulrich Christensen and Daniel Ortiz-Arroyo
Computer Science Department
Aalborg University Esbjerg
Niels Bohrs Vej 8, 6700 Denmark
E-mail: huc1405@student.aaue.dk, do@cs.aaue.dk

***Abstract*— Effective mechanisms for passage retrieval have a direct impact in the performance of *Question Answering Systems (QAS)*. In this paper we investigate the application of fuzzy logic models in *Passage Retrieval (PR)*, and describe FuzzyPR, our passage retrieval system that applies these models. Additionally, a performance analysis of diverse mechanisms employed in passage retrieval for open domain QAS is presented. Preliminary performance results of FuzzyPR, when applied to retrieving answering passages in the TREC corpora, are provided.**

Keywords: Information Retrieval, Question Answering Systems, Passage Retrieval, Fuzzy Logic

I. Introduction and Related Work

A *Question Answering System (QAS)* is one type of information retrieval (IR) system that attempts to find exact answers to user's questions expressed in natural language. In an *Open-Domain Question Answering System (ODQAS)*, questions are not restricted to certain topics and answers have to be found in an unstructured document collection. *Passage Retrieval (PR)*, one component of a QAS, extracts text segments from a group of retrieved documents and ranks these passages in decreasing order of computed likelihood for containing the correct answer. Typically, such text segments are referred to as *candidate passages*.

Applying fuzzy logic in IR systems —especially similarity measures— is a promising approach since this mathematical framework models naturally the uncertainty and vagueness involved in the process of retrieving information. Examples of successful applications of fuzzy logic in IR include Larsen's Query Answering System [1] for libraries and Szczepaniak and Gil's experimental IR system for retrieving documents written in Polish [2]. Kong et al. [4] explored the use of fuzzy aggregation operators in a passage-based retrieval system for documents, where the relevance of a document is re-calculated taking into account the retrieved passages. Brøndsted et al. describe in [3] a fuzzy logic based implementation of a document retrieval system that employs *concept clusters* and statistical query term expansion in a closed domain spoken QA system for mobile devices.

This paper investigates the application of fuzzy logic based models for PR in ODQAS. A performance analysis of some proposed mechanisms for PR system is presented jointly with the preliminary performance results of FuzzyPR, our PR system. The paper is organized as follows. Section II describes and analyzes the main component mechanisms of a passage retrieval system. Section III briefly describes FuzzyPR and presents its preliminary performance results. Finally, Section IV presents some conclusions and future work.

II. Analysis of Main Component Mechanisms in a Passage Retrieval System

The intuition "*frequently, an answer to a (factoid) question can be found as a reformulation of the same question*" has been applied successfully in passage retrieval systems for QA [5], [6]. An example of the application of this *reformulation intuition* approach in PR is the question "How much is the international space station expected to cost?"" of QA@TREC 11 (QID: 1645) [1]. The answering passage contains the snippet: "(...)United States and Russia, are working together to build the SPACE STATION, which is EXPECTED TO COST between $40 billion and $60 billion.(...)". Successful applications of this intuition include Gómes-Soriano et al.'s [5] *n*-gram based passage retrieval system for QAS, where passages containing larger sequences of terms of the questions are ranked higher and Brill et al.'s [6] Web QAS, which poses queries constructed as permutations of the terms employed in the question. In the following subsections we briefly analyze three of the main modeling components that are used in our fuzzy logic modeling implementation of the reformulation intuition. Further details can be found in [7].

A. Automatic Detection of Term Variations

A QAS requires an automatic mean to detect *term variations* occuring in documents and questions written in natural language. Term variations are lexical differences, in terms of meaning and spelling, between a word of the question typed by a user and an equivalent word contained in a document in the corpus. Reasons for term variations include grammatical inflection and spelling mistakes. These vocabulary mismatches have a negative impact on the effectiveness of an IR system when it is not able to recognize them. Two main features are needed in a mechanism to handle term variations effectively: 1) language-independence and 2) effectiveness, measured in terms of tolerance toward common misspellings and grammatical inflections (interpreted as a type of misspelling). Contrary

[1] TREC's Question Answering collections are available from: `http://trec.nist.gov/data/qa.html`

to Boolean algorithms for term matching, *fuzzy term similarity algorithms* determine the degree of similarity between two strings. Reflexivity and symmetry are desired properties of these algorithms. To select the most adequate fuzzy term similarity algorithm for FuzzyPR, we performed a comparative evaluation on the effectiveness of six different algorithms when set to calculate the similarity between 300 English homophone [2] pairs. The average of the similarity computations yields the score of the fuzzy term matching algorithm.

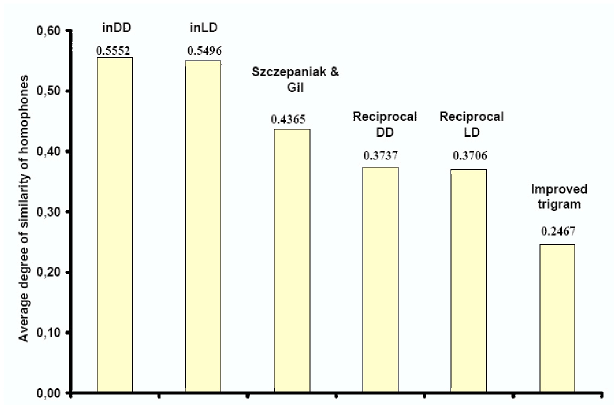

Fig. 1: Comparative evaluation of 6 fuzzy terms similarity algorithms

As illustrated in Fig. 1, the inverse normalized Damerau-Levenshtein (InDD) [7] distance performed best, giving an average homophone pair similarity rate of 0.5552.

B. Proximity of Question Terms Occurring in the Passage

The proximity of question terms occurring within a passage indicates the probable location of an answer in a document. *Fuzzy proximity measures* calculate the degree of proximity within a document of two or more question terms, based on the following two intuitions:

- If all matching document terms are juxtaposed then the measure yields 1, and;
- The farther away the matching document terms occur, the lower the degree of proximity.

We evaluated three different fuzzy proximity measures as to their ability in finding answering passages for the first 50 questions of TREC 11's question set using the AQUAINT corpus.

TABLE I: MRR of 3 fuzzy proximity measures

Fuzzy proximity measure	Mean Reciprocal Rank
Span Size Ratio	0.2933
Fuzzy Proximity Measure w. $k = 70$	0.3363
Extended Distance Factor w. $s = 0.1$ and $\alpha = 0.75$	0.3137

[2]A *homophone pair* is two terms pronounced the same but differing in meaning and spelling, thus reflecting misspellings and typos. Examples include ”advice vs. advise” and ”cite vs. site”.

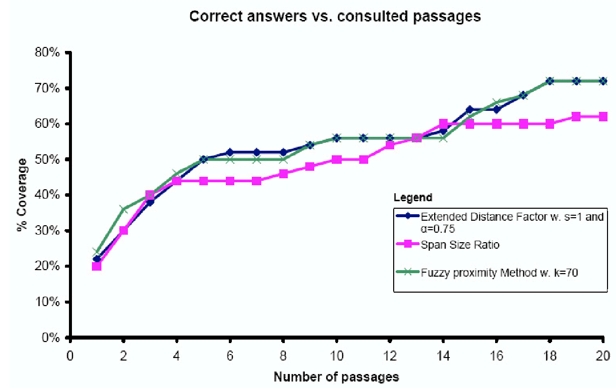

Fig. 2: The Coverage of the 3 fuzzy proximity measures

Fig. 2 and table I show that Mercier and Beigbeder's *Fuzzy Proximity Measure* [8] achieves better performance in terms of both *Coverage* and *MRR*. Coverage is defined as the proportion of questions for which an answer can be found within the n top-ranked passages: $cov(Q, D, n) \equiv \frac{|\{q \in Q | R_{D,q,n} \cap A_{D,q} \neq \emptyset\}|}{|Q|}$. MRR is the average of the reciprocal value of the first hit to each question within the top 5 candidate passages: $MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} RR_i$, $RR_i = \frac{1}{r_i}$ if $r_i \leq 5$ or 0 otherwise, where Q is the set of questions and r_i the rank of the first answering passage in response to a particular question.

C. Weighted Fraction of Question Terms Occurring in the Passage

Additionally to finding term variations and measuring the proximity of terms, a model of the *reformulation intuition* must include the weighted fraction of question terms that occur in a passage. Larsen et al.[1] and Lauritsen[9] applied successfully a simple weighted minimum measure described by: $\min_{\forall t \in q}(\max(1 - v, a))$, where t is a term in the question q, $v \in [0, 1]$ is an importance weight, and $a \in [0, 1]$ is the degree to which t occurs in the passage. Unfortunately, this measure is too strict i.e. if just a single question term is not found in the passage, the weighted minimum yields 0. A possible solution to this problem is to use an AND-like importance-weighted Averaging Operator, where the degree of ANDness is in the interval $]0.5, 1]$. Two classes of importance-weighted Averaging Operators exist [10]:

- *Multiplicative importance weighted Averaging Operators*, where importance weights are sum-normalized and membership degrees multiplied by the importance weights, and;
- *Implicative importance weighted Averaging Operators*, where importance weights are max-normalized and applied using fuzzy implication: $v \Rightarrow a$—i.e. a high weight implies a high degree of satisfaction.

We chose the class of implicative importance weighted Averaging Operators since this type of operators have the desirable property that importance weights have maximum influence on the result at both high and low ANDness.

Among the operators available, two were used in our experiments: *Andness Directed Implicative-Importance Weighting (AIWA)* and *Importance-Weighted Maximum-Entropy Ordered Weighted Averaging (IW-MEOWA)* [11]. Both operators satisfy the requirements of efficiency and effectiveness and have complementary strengths and weaknesses as shown in Table II.

TABLE II: Comparison of the properties of AIWA and IW-MEOWA

	AIWA	**IW-MEOWA**
Time complexity	$O(n)$	$O(nlog(n))$
Effectiveness	Defined for the complete range of ANDness.	Defined for the complete range of ANDness.
Other properties	*Decomposability*—it is possible to add or remove a value for aggregation without recalculating the average of all values.	The *maximum spreading of weights* ensures that the satisfaction of all criteria is taken into account.

III. Passage Retrieval System and Preliminary Performance Results

FuzzyPR employs a document retrieval system based on Apache Lucene, a popular open source vector space search engine, to efficiently retrieve documents in response to a question. Additionally, FuzzyPR includes a special PR mechanism that identifies and processes the passages included within the retrieved documents. The PR mechanism consists of two main components: 1) a question–passage similarity measure module and 2) a passage identification and extraction mechanism adapted to the special needs of QAS. The following subsections describe these components.

A. Similarity Measure

The similarity measure used in FuzzyPR is based on the fuzzy logic interpretation of the intuition: *"a passage is relevant to the question posed if many question terms or variations of these question terms occur in close proximity"*. Equation 1 describes the measure that models such intuition.

$$\mu_{rel}(p,q) = wMin\left((v_1, \mu_f(p,q)), (v_2, \mu_p(p,q))\right) \quad (1)$$

This similarity measure combines lexical and statistical data extracted at *term-level* into the two fuzzy measures: $\mu_f(p,q)$ the weighted fraction of question terms occurring in the passage and $\mu_p(p,q)$ the proximity of question terms within a passage. Aggregating these fuzzy measures using the weighted minimum gives the overall relevance score $wMin$, which is expressed as: $\min\left(\max(1-v_1, \mu_f(p,q)), \max(1-v_1, \mu_p(p,q))\right)$ with the importance weights $v_1 = 1$, $v_2 = 1$ and both the passage p and the question q represented as sets of terms: $\{t_{p_1}, t_{p_2}, ..., t_{p_n}\}$ and $\{t_{q_1}, t_{q_2}, ..., t_{q_n}\}$, respectively. $wMin$ aggregates $\mu_f(p,q)$ and $\mu_p(p,q)$ into a single fuzzy value: $\mu_{rel}(p,q)$. $\mu_{rel}(p,q)$ measures the relevance of a passage p as its degree of membership in the fuzzy subset of passages that provide a correct answer to the question q. It must be noticed that since $\mu_{rel}(p,q)$ relies only on the lexical and statistical data that can be extracted from questions and the corpus, it has the advantage of being *language independent*.

Using the results of the performance analysis described in Section II, $\mu_f(p,q)$ and $\mu_p(p,q)$ are defined in equations 2 and 3.

$$\mu_f(p,q) = h_{\alpha_f}\left((v_1^f, sat(t_{q_1}, p)) \ldots (v_n^f, sat(t_{q_n}, p))\right) \quad (2)$$

where h is the AIWA importance weighted averaging operator [11] with an ANDness of $\alpha_f = 0.7$, t_{q_i} is a question term, $v_i^f = NIDF(t_{q_i}) = 1 - \frac{log(n_i)}{1+log(N)}$, n=frequency of t_{q_i} in Ω the set of documents, $N = |\Omega|$. $sat(p, t_{q_i})$ measures the degree to which p contains t_{q_i} using the inverse normalized Damerau-Levenshtein Distance i.e. $sat(p, t_{q_i}) = \max_{\forall t_p \in p}\left(\mu_{sim}^{inDD}(t_p, t_{q_i})\right)$, where $\mu_{sim}^{inDD}(t_p, t_{q_i}) = 1 - \frac{DD(t_p, t_{q_i})}{\max(|t_p|, |t_{q_i}|)}$, DD being the Damerau-Levenshtein Distance. Finally,

$$\mu_p(p,q) = \frac{s(p,q)}{\max\limits_{\forall p_i \in \Omega} s(p_i, q)} \quad (3)$$

where $\mu_p(p,q)$ is a max-normalization of Mercier and Beigbeder's *fuzzy proximity* method [8] described by $s(p,q) = \int_1^n \mu_t^p(x)dx$, $t \in q$ with the term influence function $\mu_t^p(x) = \max\limits_{i \in Occ(t,p)}\left(\max\left(\frac{k-|x-i|}{k}, 0\right)\right)$, where the parameter adjusting the support $k = 70$. The values of v_1, v_2, α_f and k are determined experimentally.

B. Mechanism for Passage Identification and Extraction

FuzzyPR employs a fuzzified variation of the concept *arbitrary passages* [3]. Arbitrary passages are modeled by its membership function in the ideal set of passage sizes as stated in equation 4.

$$\mu_{Ideal\ passage\ size} \begin{cases} 1 & \text{if } 0 \le x \le d \\ \frac{x-b}{d-b} & \text{if } d < x < b \\ 0 & \text{if } x \ge b \end{cases} \quad (4)$$

Parameters d and b are used to adjust the crisp support and the fuzzy support values respectively. Due to efficiency concerns, the membership function of the ideal passage size set is transformed into an equivalent symmetric membership function, where the center term of a passage is required to have a question term similarity greater than α and a NIDF greater than β. The justification for this restriction is the intuition that a passage that contains none or very few of the terms in the question is unlikely to provide an answer to the question.

[3]Arbitrary passages are defined as: *"any sequence of words of any length starting at any word in the document"*.

C. Preliminary Performance Results

We measured the effectiveness of FuzzyPR, by comparing its ability to find correct answers to questions in a document corpora against an adapted PR system that we have integrated within Lucene. This adapted PR system implements an index of 3 sentence passages with 1 sentence overlapping. Llopis et al. in [12] report that this approach achieves good results. The PR system allows Lucene to be used as QAS by employing a simple query expansion method. In this method the question term with the lowest IDF is removed until ≥ 20 passages are retrieved from the index of 3 sentence passages. As test data we used TREC-12's set of 495 questions and the corpus called AQUAINT, which consists of approximately 1 million documents of English news text. To check automatically for correct answers to questions, using Ken Litkowsky's regular expression patterns[4], the question set was reduced to 380, since 115 questions do not have a pattern. As evaluation metrics we used *Coverage*, and the *Mean Reciprocal Rank (MRR)*. As is done in the JIRS system [5], we measure coverage on the first top 20 passages.

Fig. 3 and table III show that FuzzyPR consistently outperforms the vector space PR system in terms of coverage (by a margin of 5-10%) and MMR, independently of the number of top-ranked passages consulted.

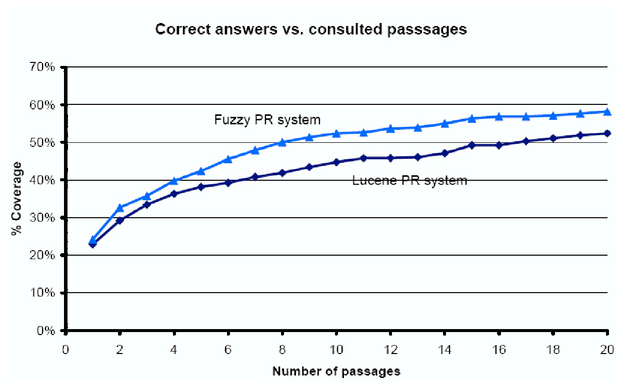

Fig. 3: The coverage of Lucene PR and FuzzyPR tested with TREC 12's QA test data

TABLE III: MRR obtained with TREC-12's QA test data

Passage retrieval system	MRR
Lucene PR	0.2855
FuzzyPR	0.3099

IV. Conclusions and Future Work

In this paper we presented FuzzyPR, a novel PR system that implements fuzzy logic models for passage retrieval. The main component mechanisms included in FuzzyPR are: 1) *automatic detection of term variations*, 2) *proximity of question terms*, and 3) *fraction of question terms occurring in the passage*. Using these components we created a fuzzy logic model based interpretation of the *reformulation intuition*. FuzzyPR has three main advantages: 1) its *passage identification and extraction methods* enables it to retrieve candidate passages from documents at retrieval retrieval time thus avoiding the time-consuming indexing process[5] 2) its language independence property, and 3) its capability for handling term variations due to spelling errors and grammatical reflections.

Our preliminary evaluation shows that FuzzyPR achieves a consistently higher coverage and MRR than a PR system adapted within Lucene. To test its performance on another language, FuzzyPR will be further evaluated with the Spanish CLEF corpus. We also plan to perform a comparative evaluation of FuzzyPR against JIRS [5]. Finally, we will explore the effect in performance of combining different fuzzy logic based PR mechanisms with machine learning techniques.

References

[1] H. L. Larsen, T. Andreasen, and H. Christiansen. Knowledge discovery for flexible querying. *In Christiansen, H., Andreasen, T., Larsen, H.L., Eds., Proceedings of the International Conference on Flexible Query Answering Systems*, Lecture Notes in Artificial Intelligence, Springer Verlag, Berlin, May 1998. Roskilde.

[2] P. Szczepaniak and M. Gil. Practical evaluation of textual fuzzy similarity as a tool for information retrieval. *AWIC 2003*, pages 250–257, 2003.

[3] T. Brøndsted, H. L. Larsen, L. Larsen, B. Lindberg, D. Ortiz-Arroyo, Z. Tan, and H. Xu. Mobile information access with spoken query answering. *In proceedings of the Workshop on Applied Spoken Language Interaction in Distributed Environments, Aalborg University*, November 2005.

[4] K. Kong, R. Luk, K. Ho, and F. Chung. Passage-based retrieval using parameterized fuzzy set operators. *ACM SIGIR Workshop on Mathematical/Formal. Methods for Information Retrieval*, 2004.

[5] J. Gómez-Soriano, M. Montes y Gómez, E. Arnal, L. Villase nor Pineda, and P. Rosso. Language independent passage retrieval for question answering. *Fourth Mexican International Conference on Artificial Intelligence MICAI 2005*, Lecture Notes in Artificial Intelligence, Springer-Verlag, November 2005. Monterrey, Mexico.

[6] Eric Brill, Jimmy Lin, Michele Banko, Susan Dumais, and Andrew Ng. Data-intensive question answering. *In Proceedings of the Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, Maryland*, pages 443–462, November 2001.

[7] H. U. Christensen. Exploring the use of fuzzy logic in passage retrieval for question answering, March 2006. Preliminary Master Thesis report available from: `http://hufnc.1go.dk/MidtermReport.pdf`.

[8] M. Beigbeder and A. Mercier. An information retrieval model using the fuzzy proximity degree of term occurrences. *Proceedings of the 2005 ACM symposium on Applied computing*, March 2005.

[9] S. Lauritsen. Fuzzylogik i domænespecifik informationsøgning (written in danish), 2006. Master Thesis in Computer Science at Aalborg University Esbjerg.

[10] J. J. Dujmovic and H. L. Larsen. Properties and modeling of partial conjunction/disjunction. *In B. De Baets et al., Ed, Current issues in data and knowledge engineering*, 18:215–224, 2004.

[11] H. L. Larsen. Efficient andness-directed importance weighted averaging operators. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, pages 67–82, 2003.

[12] F. Llopis, A. Ferrández, and J. Luis Vicedo. Text segmentation for efficient information retrieval. *In Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002), Mexico City, Mexico*, Lecture Notes in Computer Science, Springer-Verlag:373–380, 2002.

[4]Ken Litkowsky's patterns are available from the TREC website: `http://trec.nist.gov`.

[5]An unoptimized method in Java for segmenting and indexing the AQUAINT corpus took 4 hours on an AMD64 3400+ w. 2 GB RAM and RAID 0.