Exploring the Application of Fuzzy Logic and Data Fusion Mechanisms in QAS

Daniel Ortiz-Arroyo and Hans Ulrich Christensen

Computer Science Department
Aalborg University Esbjerg
Niels Bohrs Vej 8, 6700 Denmark
do@cs.aaue.dk, huc1405@student.aaue.dk

Abstract. In this paper we explore the application of fuzzy logic and data fusion techniques to improve the performance of passage retrieval in open domain *Question Answering Systems (QAS)*. Our experiments show that our proposed mechanisms provide significant performance improvements when compared to other similar systems.

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

1 Introduction

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 *Open-Domain Question Answering Systems (ODQAS)*, answers to questions have to be found within an unstructured document collection containing different topics. *Passage Retrieval (PR)* is one component of a QAS that extracts text segments from a group of retrieved documents and ranks these passages in decreasing order of computed likelihood for containing the correct answer to a question. The overall performance of a QAS is determined, in large part, by the performance of its PR system.

Data Fusion is the combined ranking performed by a variety of IR systems on a document's relevance to a user's information need. When applied to QAS, the goal of the fusing process is to improve performance by combining the relevance scores obtained by a diversity of PR systems participating in an ensemble.

This paper describes an efficient language-independent, fuzzy logic-based model PR system, together with data fusion mechanisms for QAS. The paper is organized as follows. Section 2 briefly describes related work on passage retrieval systems and data fusion. Section 3 describes the main component mechanisms of the fuzzy logic based PR system and its performance results. Section 4 briefly describes the data fusion methods employed and presents the final performance results obtained by our system. Finally, Section 5 presents some conclusions and future work.

F. Masulli, S. Mitra, and G. Pasi (Eds.): WILF 2007, LNAI 4578, pp. 102–109, 2007.

2 Related Work

JIRS is a PR system based on a n-gram model introduced by Gómoz-Soriano et al. in [1] that was adapted to the special needs of QA. JIRS supports two extensions to the basic n-gram matching mechanism (called *Simple Model*): *Term Weights* model and the *Distance Model* that include both term weights and a distance measure. JIRS ranks higher passages containing larger sequences of the terms contained in the questions. A related work is Web QA system [2] that builds queries constructed as permutations of the terms employed in the questions.

Several studies have investigated the application of Data Fusion to QA systems achieving in general promising results e.g. [3] reported a consistent improvements in terms of precision as high as 20%. However, few have investigated the potentially beneficial application of Data Fusion to the task of PR within a QAS. One example is [4] where a consistent significant improvement in Coverage@n is achived on the TREC11 collection. However, the machine learning techniques employed in [4] require an extra training step to learn the features of answering passages. Finally, Tellex et. al. experimentally fused three PR systems achieving a slight increase in performance in terms of MRR [5] using a simple voting mechanism.

Other studies on the application of Data Fusion to document retrieval (e.g. [6] and [7]) have reported important improvements in performance but on ad-hoc document retrieval systems and not specifically within PR for QAS.

3 A Fuzzy Logic-Based PR System

In QAS *the question reformulation intuition* stated as: "*a passage p is relevant to the user's question q if many question terms or variations of these question terms occur in close proximity*" is a commonly used technique to retrieve answering passages to a question. Fuzzy Logic is especially suited to model this intuition. The feature *"many (important) question terms"* can be modeled by the fuzzy subset: *The degree to which candidate passages contain all question terms. "Close proximity"* can be modeled by the fuzzy subset: *The degree to which the question terms contained in a candidate passage are juxtaposed* i.e. the more distributed the terms are, the lower the degree of proximity will be. The third vague concept that can be used to model the reformulation intuition is *term matching*. The fuzzy logic interpretation of binary term similarity is the fuzzy subset: *The degree to which two terms are identical.*

The reformulation intuition was modeled and implemented within *FuzzyPR*. *FuzzyPR* consists of two components: 1) a question–passage similarity measure module and 2) a passage identification and extraction mechanism adapted to the special needs of QAS. *FuzzyPR* uses a similarity measure based on the fuzzy logic interpretation of the *reformulation intuition* described by Equation 1.

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

The similarity measure combines lexical and statistical data extracted at *term-level* into two fuzzy measures: $\mu_f(p,q)$ the weighted fraction of question terms q occurring in the passage p and $\mu_p(p,q)$ the proximity of question terms q within the passage. $\mu_f(p,q)$ and $\mu_p(p,q)$ are defined in equations 2 and 3.

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

where h is the AIWA importance weighted averaging operator [8] with an AND-ness of $\alpha_f = 0.65$, t_{q_i} is a question term, $v_i^f = NIDF(t_{q_i}) = 1 - \frac{log(n_i)}{1+log(N)}$ [1], 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 normalized longest common subsequence (nLCS), i.e. $sat(p,t_{q_i}) = \max\limits_{\forall t_p \in p}\left(\mu_{sim}^{nLCS}(t_p,t_{q_i})\right)$, where $\mu_{sim}^{nLCS}(t_p,t_{q_i}) = \frac{|LCS(t_p,t_{q_i})|}{\max(|t_p|,|t_{q_i}|)}$, LCS being the longest common subsequence. Finally,

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

where $\mu_p(p,q)$ is a max-normalization of Mercier and Beigbeder's *fuzzy proximity* method [9] 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 were determined experimentally. Aggregating these two fuzzy measures using the weighted minimum gives the overall relevance score $wMin$, which is defined as:

$$wMin(v_1,v_2,\mu_f,\mu_p) = \min\left(\max(1-v_1,\mu_f(p,q)), \max(1-v_2,\mu_p(p,q))\right). \tag{4}$$

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 $\{t_{q_1}, t_{q_2}, ..., t_{q_m}\}$, respectively. $wMin$ aggregates $\mu_f(p,q)$ and $\mu_p(p,q)$ into a single fuzzy value $\mu_{rel}(p,q)$ as described by Equation 1. $\mu_{rel}(p,q)$ is the fuzzy subset of passages providing a correct answer to the question q, where p is a specific passage. $\mu_{rel}(p,q)$ has the advantage of being *language-independent*.

FuzzyPR also employs a fuzzified variation of the concept *arbitrary passages*[2]. Details on the membership function employed to describe an arbitrary passage can be found in [10].

We measured the effectiveness of *FuzzyPR* comparing its ability to find correct answers to questions with JIRS' PR system [1] and an adapted PR system that we have integrated within Lucene using two different document corpora. The adapted PR system allows Lucene to be used as the PR module in a QAS, employing a simple query expansion method that keeps removing the question term with the lowest IDF until ≥ 20 passages are retrieved from the index of

[1] NIDF is an abbreviation of normalized inverse document frequency.

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

Table 1. MRRs obtained with TREC12's and CLEF04's QA test data

PR system / QA test data	**TREC12**	**%Impr.**	**CLEF04**	**%Impr.**
FuzzyPR	0.3394	-	0.3726	-
JIRS Distance Model	0.3180	6.73%	0.3721	0.13%
JIRS Simple Model	0.2724	24.60%	0.3771	−1.19%
Lucene	0.2910	16.63%	0.3399	9.62%

Table 2. The PR systems' coverages tested with (a) TREC12 and (b) CLEF04 data

(a)

	FuzzyPR	Lucene	JIRS_SM	JIRS_DM
1	0.250	0.224 (11.8%)	0.222 (12.5%)	0.243 (2.7%)
2	0.358	0.305 (17.2%)	0.270 (32.7%)	0.320 (11.8%)
3	0.418	0.350 (19.5%)	0.299 (40.0%)	0.384 (9.1%)
4	0.450	0.371 (21.3%)	0.347 (29.8%)	0.421 (7.0%)
5	0.487	0.403 (20.9%)	0.370 (31.4%)	0.450 (8.2%)
6	0.518	0.424 (22.4%)	0.405 (28.1%)	0.479 (8.3%)
7	0.542	0.434 (24.9%)	0.431 (25.7%)	0.492 (10.2%)
8	0.568	0.453 (25.6%)	0.447 (27.1%)	0.508 (11.9%)
9	0.582	0.479 (21.4%)	0.479 (21.5%)	0.532 (9.4%)
10	0.595	0.495 (20.2%)	0.489 (21.5%)	0.548 (8.6%)
11	0.611	0.505 (20.8%)	0.495 (23.4%)	0.558 (9.4%)
12	0.616	0.524 (17.6%)	0.505 (21.9%)	0.569 (8.3%)
13	0.621	0.529 (17.4%)	0.521 (19.2%)	0.579 (7.2%)
14	0.624	0.537 (16.2%)	0.527 (18.5%)	0.590 (5.7%)
15	0.624	0.547 (13.9%)	0.529 (17.9%)	0.595 (4.8%)
16	0.626	0.550 (13.9%)	0.532 (17.8%)	0.603 (3.8%)
17	0.632	0.558 (13.2%)	0.548 (15.3%)	0.609 (3.8%)
18	0.637	0.561 (13.6%)	0.556 (14.6%)	0.611 (4.2%)
19	0.637	0.561 (13.6%)	0.564 (13.0%)	0.616 (3.3%)
20	0.645	0.563 (14.5%)	0.571 (12.8%)	0.619 (4.2%)

(b)

	FuzzyPR	Lucene	JIRS_SM	JIRS_DM
1	0.283	0.272 (4.1%)	0.322 (−12.1%)	0.300 (−5.6%)
2	0.378	0.372 (1.5%)	0.389 (−2.9%)	0.372 (1.5%)
3	0.439	0.394 (11.3%)	0.411 (6.8%)	0.444 (−1.2%)
4	0.494	0.422 (17.1%)	0.450 (9.9%)	0.483 (2.3%)
5	0.533	0.439 (21.5%)	0.472 (12.9%)	0.494 (7.9%)
6	0.556	0.456 (21.9%)	0.494 (12.4%)	0.528 (5.3%)
7	0.561	0.472 (18.8%)	0.522 (7.4%)	0.544 (3.1%)
8	0.572	0.472 (21.2%)	0.528 (8.4%)	0.567 (1.0%)
9	0.572	0.483 (18.4%)	0.533 (7.3%)	0.572 (0.0%)
10	0.594	0.489 (21.6%)	0.561 (5.9%)	0.583 (1.9%)
11	0.600	0.489 (22.7%)	0.561 (6.9%)	0.583 (2.9%)
12	0.617	0.489 (26.1%)	0.567 (8.8%)	0.594 (3.8%)
13	0.622	0.489 (27.3%)	0.567 (9.8%)	0.600 (3.7%)
14	0.628	0.500 (25.6%)	0.578 (8.7%)	0.606 (3.7%)
15	0.628	0.506 (24.2%)	0.578 (8.7%)	0.617 (1.8%)
16	0.639	0.506 (26.4%)	0.578 (10.6%)	0.617 (3.6%)
17	0.639	0.506 (26.4%)	0.578 (10.6%)	0.617 (3.6%)
18	0.639	0.517 (23.7%)	0.578 (10.6%)	0.622 (2.7%)
19	0.644	0.522 (23.4%)	0.583 (10.5%)	0.628 (2.6%)
20	0.650	0.533 (21.9%)	0.583 (11.4%)	0.633 (2.6%)

3 sentence passages. Both, the PR system and JIRS implement an index of 3 sentence passages with 1 sentence overlapping since as it is reported in [11] this approach achieves good results.

As test data we used TREC12's set of 495 questions and the corpus called AQUAINT consisting of 1,033,461 documents of English news text together with CLEF04's 180 question and the AgenciaEFE corpus of 454,045 Spanish newswire documents. To check for correct answers automatically we used Ken Litkowsky's regular expression patterns of correct answers for TREC12 and the patterns supplied with JIRS.[3] As evaluation metrics we used *Mean Reciprocal Rank (MRR)* and *coverage*. Finally, the TREC12 question set was reduced to 380, since 115 questions do not have a recognizable pattern.

In Table 3, a parenthesized value is *FuzzyPR*'s performance improvement, expressed as a percentage, compared to other PR systems. Tables 3 and 2(b) show that *FuzzyPR* consistently performs better than Lucene's vector space PR system independently of the number of top-ranked passages consulted tested with both TREC12 and CLEF04 QA test data. MRR is improved at least 9.62% and coverage@20 at least 14.5%. Our results also show that FuzzyPR performs better than JIRS_SM and JIRS_DM (simple and distance model respectively) on the

[3] Patterns of correct answers to CLEF QA test data are available from JIRS' web site `http://jirs.dsic.upv.es/`

TREC copora, but slightly worst than JIRS_SM on the CLEF corpora. One explanation for this is that answers sometimes do not conform to the reformulation intuition.

4 Data Fusion Methods

In this section we present a brief description of the data fusion methods employed in our experiments[4]. The *Condorcet-fuse* method is a generalization of the Condorcet election process, where the winner of an election is the candidate that beats or ties with every other candidate in a pair-wise comparison, such that the result is a ranked list of documents rather than a single winner. *Borda-Fuse*, introduced in [12], is an adaptation of the Borda Count election process, where voters give candidates a certain amount of points and the winner is the one who makes more points. Tellex et al. [5] propose a method that combines passages rank and a simple vote: the total number of passages retrieved by all component PR systems with a specific document ID fused into a relevance score. Based on the observation that frequently when Tellex et al.'s Fusion method boosted low ranked passages, those passages in fact were non-relevant, we propose a new Fusion method called *Tellex Modified*, where the union of top m passages retrieved by all component PR systems is re-ranked. Fox and Shaw [13] introduce and evaluate the 6 simple Fusion methods depicted in Table 3.

Table 3. The six Fusion Methods introduced by Fox and Shaw (adapted from [13])

CombMAX	$r_f(d_i) = \max\limits_{\forall s_j \in S} \left(r_{s_j}(d_i)\right)$
CombMIN	$r_f(d_i) = \min\limits_{\forall s_j \in S} \left(r_{s_j}(d_i)\right)$
CombSUM	$r_f(d_i) = \sum\limits_{\forall s_j \in S} \left(r_{s_j}(d_i)\right)$
CombANZ	$r_f(d_i) = CombSUM/t$
CombMNZ	$r_f(d_i) = CombSUM * t$
CombMED	The median of a document's similarities

In table 3, $r_f(d_i)$ is the fused relevance score (similarity) of the document d_i, $r_{s_j}(d_i)$ document d_i's similarity at the IR system $s_j \in S$, the set of IR systems to be fused, and t the number of IR systems retrieving d_i.

Borda-fuse can be extended to a weighted variant: *Weighted Borda-fuse* by multiplying the points, which a PR system S_i assigns to a candidate passage with an overall system weight α_i [12].

In *weighted Condorcet-fuse*, Condorcet-fuse is extended to take importance weights into account, where each component PR system provides an importance weighted vote. These importance weights are used in binary candidate elections,

[4] A complete description can be found in [10].

where the sum of weights rather than votes is compared, giving preference to the highest sum.

The *Linear combination* (LC) Data Fusion method combines the relevance scores and training data of two or more component IR systems into a combined relevance score per document [14]. In LC, training data are used for calculating *importance weights* based on standard IR metrics, thus reflecting the overall ability of the system to provide relevant documents. The aggregated relevance score of a document is calculated in equation 5 using individual relevance scores and performance weights.

$$s_{LC}(d) = \sum_{\forall s_i \in S} \alpha * s_i(d) \tag{5}$$

where $S_{LC}(d)$ is the fused relevance score assigned to the document d, s_i is the ith system of the set S of PR systems whose relevance score will be combined, and a_i the importance weight assigned to the ith PR systems. In LC, if an IR system does not retrieve a particular document, then the IR system is assumed to consider it non-relevant by assigning a relevance score of 0. Additionally to the 9 Data Fusion methods previously described, we applied in our experiments subclass weighting to *weighted Condorcet-Fuse*, *weighted Borda-Fuse*, *LC* and *weighted Maximum Entropy OWA (MEOWA)*, comprising a total of 13 different methods.

Since IR systems use different scales for relevance scores it is necessary to normalize them. For this task and based on the evaluation of 7 different performance weights, we selected max-normalized MRR (nMRR). Lastly, we found it necessary to exclude the Condorcet-fuse method with question type weights because it consistently worsened overall performance. In our experiments we used the following 8 component PR systems: JIRS [15] using both the Simple Model and the Distance Model, FuzzyPR, FuzzyPRS+LucenePRS, LucenePRS, Swish-e, Terrier PL2 [16] using In expC2 probabilistic model, and Zettair.

As test data we used TREC12's set of 495 questions and the corpus called AQUAINT consisting of $1,033,461$ documents of English news text and CLEF04's 180 question and the AgenciaEFE corpus of $454,045$ Spanish newswire documents. We used Ken Litkowsky's regular expression patterns of correct answers to check answers automatically for TREC12 and for CLEF4 we used the pattern supplied with JIRS[5] The TREC12 question set was reduced to 380, since 115 questions do not have a recognizable pattern. As evaluation metrics we used MRR, Coverage, and Redundancy.

Table 4 shows the results obtained using from 2 up to 6 of the best performing PR mechanisms combined. These results show that the Data Fusion methods were able to improve performance measured as MRR by a maximum of 6.43% and Coverage@20 by 11.39%. This result was obtained fusing 4 of the best performing PR system with the *Tellex Modified* fusion method. *Tellex Modified*

[5] Patterns of correct answers to CLEF QA test data are available from JIRS' website: `http://jirs.dsic.upv.es/`

Table 4. The MRR and Coverage@20 of Tellex Modified compared to the 2nd best Fusion methods tested with TREC12 and CLEF04 QA test data

(a) TREC12 QA test data

Performance metric	MRR					Coverage@20				
No. of PR4QA systems combined	2	3	4	5	6	2	3	4	5	6
Avg. of best PR4QA systems	0.300	0.309	0.316	0.321	0.324	0.590	0.607	0.617	0.623	0.627
Tellex Modified (best)	0.317	0.329	0.336	0.341	0.344	0.642	0.675	0.687	0.694	0.696
Relative performance in %	5.54%	6.41%	6.43%	6.33%	6.25%	8.77%	11.32%	11.39%	11.30%	10.91%
LC	0.300	0.316	0.327	0.332	0.335	0.613	0.636	0.651	0.659	0.664
Relative performance in %	0.00%	2.15%	3.46%	3.55%	3.43%	3.85%	4.79%	5.51%	5.66%	5.89%

(b) CLEF04 QA test data

Performance metric	MRR					Coverage@20				
No. of PR4QA systems combined	2	3	4	5	6	2	3	4	5	6
Avg. of best PR4QA systems	0.352	0.362	0.369	0.375	0.379	0.590	0.607	0.617	0.623	0.627
Tellex Modified (best)	0.357	0.367	0.371	0.376	0.379	0.622	0.658	0.673	0.681	0.685
Relative performance in %	1.28%	1.26%	0.59%	0.39%	−0.03%	5.45%	8.40%	9.07%	9.19%	9.10%
LC w. quest. class weights	0.350	0.362	0.370	0.377	0.385					
LC						0.615	0.642	0.655	0.664	0.671
Relative performance in %	−0.7%	−0.11%	0.12%	0.54%	1.60%	4.24%	5.83%	6.15%	6.52%	6.92%

required neither relevance scores of passages nor importance weights assigned to the fused PR systems.

5 Conclusions and Future Work

Our experiments show that *FuzzyPR* achieves higher MRR and coverage than other similar systems on the TREC corpora. Furthermore it performs better in terms of coverage than JIRS on the CLEF corpora at ranks 4 to 20 but also slightly worse than JIRS simple model in terms of MRR on the same collection. Additionally, we investigated the application of a total of 13 Data Fusion methods, eight of these utilizing importance weights and importance weight per subclass of questions. We found that our proposed modification to Tellex et. al.'s method is able to improve MRR by a maximum of 6.43% and Coverage@20 by 11.39% fusing 4 different PR systems. However, contrary to our initial expectations, we found that the use of importance weights and importance weights per subclass of questions did not provide any improvement in data fusion performance.

As future work we consider addressing some of the weaknesses we found in our approach, namely handling questions with answers not conforming to the reformulation intuition and investigating optimal ways to include relevance scores in the data fusion mechanisms.

References

1. Gómez-Soriano, J.: A passage retrieval system for multilingual question answering. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, pp. 443–450. Springer, Heidelberg (2005)
2. Brill, E., Lin, J., Banko, M., Dumais, S., Ng, A.: Data-intensive question answering. In: Proceedings of the Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, Maryland (November 2001) pp. 443–462 (2001)

3. H., et al.: Employing two question answering systems in trec-2005. In: proceedings of The Fourteenth Text REtrieval Conference (TREC 2005) (2005)
4. Unsunier, N., Amini, M., Gallinari, P.: Boosting weak ranking functions to enhance passage retrieval for question answering. In: IR4QA workshop of SIGIR 2004 (2004)
5. Tellex, S., Katz, B., Lin, J., Marton, G., Fernandes, A.: Quantitative evaluation of passage retrieval algorithms for question answering. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003) (July 2003)
6. Lee, J.H.: Combining multiple evidence from different properties of weighting schemes. In: proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (July 1995)
7. Montague, M.: Metasearch: Data fusion for document retrieval. PhD thesis, Dartmouth College (2002)
8. Larsen, H.L.: Efficient andness-directed importance weighted averaging operators. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, pp. 67–82 (2003)
9. Beigbeder, M., Mercier, A.: An information retrieval model using the fuzzy proximity degree of term occurrences. In: Proceedings of the 2005 ACM symposium on Applied computing (March 2005)
10. Christensen, H.U.: Exploring the use of fuzzy logic and data fusion techniques in passage retrieval for question answering. Master's thesis, Aalborg University Esbjerg (December 2006)
11. Llopis, F., Ferrandez, A., Vicedo, J.L.: Text segmentation for efficient information retrieval. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 373–380. Springer, Heidelberg (2002)
12. Aslam, J., Montague, M.: Models for metasearch. In: The 24th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '01), New Orleans, LA (2001)
13. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: The Second Text REtrieval Conference (TREC-2), Gaithersburg, MD, USA, pp. 243–249 (March 1994)
14. Vogt, C.C., Cottrell, G.W.: Fusion via a linear combination of scores. Information Retrieval 1(1), 151–173 (1999)
15. Gómez-Soriano, J.: y Gómez, M.M.: Jirs—the mother of all the passage retrieval systems for multilingual question answering?, `http://www.dsic.upv.es/workshops/euindia05/slides/jgomez.pdf`
16. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 517–519. Springer, Heidelberg (2005)