Automatic Evaluation of Extract Summaries Using Fuzzy F-Score Measure

G.Ravindra
Supercomputer Education
& Research Center
Indian Institute of Science
Bangalore-12
Email:*indravipico@hotmail.com*

N.Balakrishnan
Supercomputer Education
& Research Center
Indian Institute of Science
Bangalore-12
Email:*balki@serc.iisc.ernet.in*

K.R.Ramakrishnan
Department of
Electrical Engineering
Indian Institute of Science
Bangalore-12
Email:*krr@ee.iisc.ernet.in*

Abstract

This paper describes a fuzzy union based approach for automatically evaluating machine generated extract summaries. The proposed method represents every sentence within a machine generated summary as a fuzzy set. Sentences in the reference summary are assigned membership grades in each of these fuzzy sets using cosine distance measure. Finally Fuzzy union (s-norm operation) is used to compute an F-score measure. Max s-norm and Frank's s-norm operators are discussed and the advantages of using a modified Frank's s-norm operator for evaluating summaries is explained. The proposed evaluation method is compared to ROUGE evaluation system with some test cases.

Keywords:

Summary Evaluation, Fuzzy union, Frank's s-norm operator, Longest Common Sub-sequence

1 Introduction

Automatic text summarization involves algorithms and mathematical operations that can result in knowledge compaction. Summaries can be broadly classified as abstracts and extracts. An extract is a summary which is a collection of portions of the original document reproduced without any changes, while abstracts are re-phrased versions of an extract. One of the challenges faced in the field of text summarization is to find a method that quantitatively evaluates a summary. This has been a nagging problem as human evaluators tend to be inconsistent with their evaluations. More over, it is subject to their domain expertise and points of view with regard to "what is important". When there are a large number of summaries to be evaluated it takes a lot of time and logistic problems to engage a team of human evaluators. Every time algorithms are changed or fine tuned, we will have to rely on the same team of human evaluators. In this light, the Document Understanding Conferences (DUC) have been using human evaluators along with machine-based evaluation methods. Although there is no automatic evaluation method that matches human evaluators, some methods like ROUGE are popular.

In this paper we present a Fuzzy-union based framework for automatically evaluating machine generated sentence extracts given a standard human generated extract. We compare this framework with methods like ROUGE and discuss the strong and week points with example sentences.

1.1 Literature Review

Methods to evaluate summaries can be classified into human evaluation methods and machine-based evaluation methods. DUC 2002 summarizer contest employed manual evaluators who were aided by Summary Evaluation Environment (SEE 2.0)[Lin, 2001] as the front-end tool. NIST assessors who created the "model" summaries, compared the automatically generated summaries with the former and with the baseline summaries. The assessors step through each model-unit and mark all the units in the machine generated summary sharing content with the current model-unit. Depending on the extent to which content in the marked unit matches with the current model unit the latter is classified as *fully* matches, *almost* matches, *partially* matches and *hardly* matches. To measure quality of the summaries, assessors rated the machine generated summary on a scale quantized as: 100%, 80%, 60%, 40%, 20% and 0% based on whether the summary observes English grammatical rules independent of the content, sentences in the summary fit-in with their surrounding sentences and whether the content of the summary is organized in an effective way.

Recall score is a popular measure for quantitatively comparing two summaries[Mani et al., 1998]. This score is a measure of how well a machine-generated summary retains important content of original documents . But it has been shown [Donaway et al., 2000] that simple recall score cannot differentiate system performance effectively and ***coverage score*** defined as

$$C = \frac{(Number\ of\ units\ marked) \times E}{Total\ units\ in\ model\ summary} \tag{1}$$

can be used as a better measure. E is the ratio of completeness on a scale 1-0: 1 for all match, 0.75 for most match, 0.5 for some match, 0.25 for hardly any and 0 for none. Meanwhile, the machine-translation community, having faced similar problems of quantitative evaluation of machine generated translations adopted an N-gram Co-Occurrence measure called BLEU. The foundation principle of BLEU[Papineni et al., 2002],[Papineni et al., 2001] is the modified n-gram precision that captures both "Fluency" and "Adequacy". Recently a modified version of BLEU [Lin and Hovy, 2003] was proposed as a metric for automatic evaluation of summaries. Unlike BLEU, this modified version is a recall centric approach that uses a brevity bonus which gives higher weights to sentences that are shorter. BLEU used a brevity penalty that penalized short sentences. Most recent evaluation methods like ROUGE "Weighted LCS" [Chin-Yew and E.Hovy, 2003] and normalized pairwise LCS used by the MEAD summarization group [Radev et al., 2002] use "Longest Common Subsequence" (LCS) match [Saggion et al., 2002]. In the light of these developments we propose a new F-score based evaluation method that harnesses the power of Fuzzy set union theory.

Remaining sections of the paper are organized as follows: section-2 discusses the fuzzy set based frame-work for automatically evaluating summaries. Section-3 discusses the fuzzy union operators followed by a comparison to ROUGE in section-4. Section-5 concludes the paper.

2 Fuzzy Precision, Recall and Fscore

In this section Precision, Recall and Fscore measures are presented in a fuzzy set theory framework. The advantages of such a framework and the ease with which external information can be used to produce reliable evaluation scores is explained. The frame work assumes that the summary produced is an extract, the most basic unit in the extract being a sentence.

Formally precision, recall and f-score measures are can be defined as follows:

There exists reference extract R and a candidate extract T that have N_R and N_T units respectively. The quality of T needs to be quantitatively assessed given the reference R. Precision is defined as $p = \frac{(N^T_{match}|R)}{N_T}$ where $(N^T_{match}|R)$ is number of units in T that occur in R. Similarly recall is defined as $r = \frac{(N^T_{match}|R)}{N_R}$. The f-score is a weighted combination of p and r and is given by $f = \frac{1}{\lambda\frac{1}{p}+(1-\lambda)\frac{1}{r}}$, where λ is the weighting factor. If $\lambda = \frac{1}{2}$ then precision and recall are both given equal weight. Figure-1 shows a plot of precision, recall and f-score values for the performance of 10 sentence extraction systems.

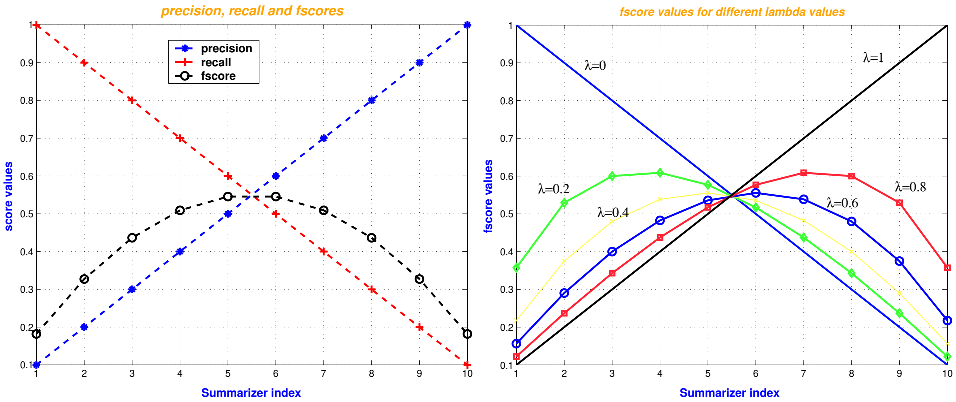

Figure 1: Behavior of precision, recall and fscores for different values of λ

This figure shows that of these systems, those with higher index values have a higher recall value but a lower precision value. The reference extract is a set of sentences picked from the original document. Each of the automatic extract generation systems were programed to pick the most important sentences from the original document and the chosen sentences were compared against sentences in the reference extract. The first system has the highest recall and the last system has the lowest recall. Based on the recall score alone, the first system can be declared as the winner. But a closer look at the extract produced by the first system showed that it was nothing but the original document itself. The last system on the other hand, produced an extract that had only one sentence and this sentence was also found in the reference extract. Hence its precision is 1 but its recall is very low. The curve for f-score is a combination of precision and recall scores and it takes into account the relevance of the automatically generated extract (recall) and the influence of its length (precision). The second plot in Figure-1 shows the behavior of the curve for f-scores, when computed with different λ values. An f-score computed with $\lambda = 0.5$ weights the precision and recall values equally. It is necessary to weigh precision and recall equally as we would like a candidate summary to be small and at the same time cover all the matter present in the reference summary.

The need for fuzzy versions of these standard scores can be better understood with an example. Consider the following three sentences collected from the DUC 2001 dataset:

R1: JAMMU-KASHMIR IS THE ONLY STATE WITH A MOSLEM MAJORITY IN PREDOMINANTLY HINDU INDIA.

R2: INDIA ACCUSES PAKISTAN OF ARMING MILITANTS.

C1: HINDU DOMINATED INDIA ACCUSES PAKISTAN OF ARMING AND TRAINING KASHMIRI MILITANTS FIGHTING FOR SECESSION OF JAMMU KASHMIR INDIA S ONLY STATE WITH A MOSLEM MAJORITY

Let the first and second sentence be present in the reference summary and the third sentence in the candidate summary. After the use of a stemmer, we have the following observations from this example:

- if the basic units for evaluation are sentences, then $N_R = 2$ and $N_T = 1$. Number of exact sentence matches is zero, and hence $p = r = 0$.
- if we compute $(C_1 \cap R_1) \cup (C_1 \cap R_2)$ with basic units for matching being words, we find that there are 11 words common between the reference summary and the candidate summary. Hence $p = 11/18$ and $r = 11/14$
- if the basic units are LCS then we have the following LC sub-sequences $(C_1 \cap R_1)$ ={JAMMU KASHMIR STATE MOSLEM MAJORITY} and $(C_1 \cap R_2) =$ {INDIA ACCUSES PAKISTAN ARMING MILITANTS }. Hence $p = 10/18$ and $r = 10/14$.

The above mentioned observations show that resorting to a "longest common sub-sequence" match results in lower scores, while word-based match results in a higher score. But word-based match has its own problems as any arbitrary collection of words can also produce a match. LCS has the drawback of matching short sequences and in many cases not finding a long enough sequence as, in realistic circumstances tense and other word modifiers like adjectives and adverbs can cause word re-ordering.

2.1 Fuzzy-set model for evaluation

This model for evaluation assumes that the reference extract is a fuzzy set and each sentence in the candidate extract has a membership grade in this set. Further it is assumed that a sentence in the candidate extract has a membership grade associated with every sentence in the reference extract, and the membership grade of this sentence in the reference extract is the union of its individual membership grades. Further, the quality of the candidate summary is expressed in terms of precision, recall and f-scores.

2.1.1 Membership grades

Let $R=\{r_1, r_2, r_3, ..., r_{|R|}\}$ be a reference extract consisting of $|R|$ sentences and $T=\{t_1, t_2, ..., t_{|T|}\}$ be a candidate extract consisting of $|T|$ sentences. A sentence $t_i \in T$ is said to have some similarity with every sentence in R on the scale $[0,1]$. This similarity is called the membership grade. Computation of membership grade should be ideal for summary evaluation. As we intend to evaluate an extract consisting of sentences, the membership grade becomes a measure for sentence similarity. Sentence similarity is computed on "basic units" defined for a sentence pair. These units can be

1. words (after removal of stop-words)
2. collocations
3. sub-sequences

A vector space model is used for computing sentence similarity $\mu_{r_j}(t_i)$ between $r_j \in R$ and $t_i \in T$ as follows:

1. let number of basic units in r_j be $|r_j|$ and in t_i be $|t_i|$.
2. let V be the cardinality of the set $r_j \cup t_i$.
3. then $\hat{r_j}$ and $\hat{t_i}$ are V dimensional vectors corresponding to r_j and t_i. Each dimension is the number of times the basic unit corresponding to that dimension has occurred in the corresponding sentence.

There are various functions that can be used to compute the membership grade as a sentence similarity measure. The cosine distance defined as $\mu_{r_j}(t_i) = \frac{\sum_{k=1}^{k=V} r_{jk} t_{ik}}{\sqrt{\left(\sum_{k=1}^{k=V} r_{jk}^2\right)\left(\sum_{k=1}^{k=V} t_{ik}^2\right)}}$ produces a score in the range $[0,1]$ and hence is a simple and elegant measure for similarity. More over it has been extensively used to quantitatively find similarity between documents in information retrieval applications. Hence in the proposed method for summary evaluation we use cosine-distance to assign membership grades.

2.1.2 Computing precision

Let every sentence $r_j \in R$ be considered as a fuzzy set. As a result, R now becomes a collection of fuzzy sets and sentence $t_i \in T$ has a membership grade in each of these fuzzy sets. Let $\mu_{r_j}(t_i)$ be the membership grade of the sentence t_i in the fuzzy set r_j. The reference summary can be written as $R = \cup_{j=1}^{j=|R|} r_j$, a union of fuzzy sets. In classical set theory, the membership grade of an element in a set is 0 or 1. Hence, the membership grade of an element in the union can be written as

$$\mu_R(t_i) = \begin{cases} \max_{j=1..|R|}\left(\mu_{r_j}(t_i)\right) = 0; \quad if\ \mu_{r_j}(t_i) = 0\ \forall\, t_i \\ \max_{j=1..|R|}\left(\mu_{r_j}(t_i)\right) = 1; \ if\ \mu_{r_j}(t_i) = 1\ for\ some\ t_i \end{cases} \quad (2)$$

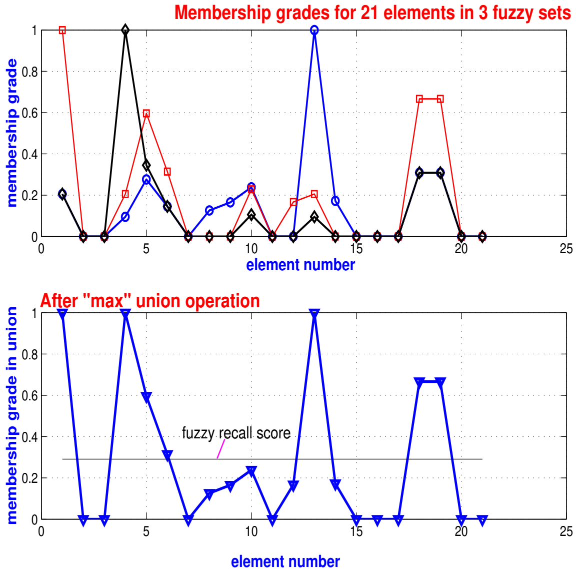

Figure 2: Result of “max” as union operator

The fuzzy set union can also be written as mentioned in eq:-2, but the membership grade in the union need not be only 0 or 1. Further an S-Norm operator(union operator) $\bigvee$ can be defined to replace the max operator used in eq:-2. Then fuzzy precision can be defined as $p_F = \frac{\sum_{i=1}^{i=|T|}\left(\bigvee_{j=1..|R|} \mu_{r_j}(t_i)\right)}{\sum_{i=1}^{i=|T|}\left(\bigvee_{j=1..|T|} \mu_{t_j}(t_i)\right)}$

2.1.3 Computing recall

Let every sentence $t_i \in T$ be considered as a fuzzy set. Similar to the case of precision computation, T now becomes a collection of fuzzy sets and sentence $r_j \in R$ has a membership grade in each of these fuzzy sets. Let $\mu_{t_i}(r_j)$ be the membership grade of the sentence r_j in the fuzzy set t_i. The candidate summary can be written as $T = \cup_{i=1}^{i=|T|} t_i$, a union of fuzzy sets and, the fuzzy recall score can be computed as $r_F = \frac{\sum_{j=1}^{j=|R|}\left(\bigvee_{i=1..|T|} \mu_{t_i}(r_j)\right)}{\sum_{j=1}^{j=|R|}\left(\bigvee_{i=1..|R|} \mu_{t_i}(r_j)\right)}$

3 The S-Norm operator $\bigvee$

3.1 MAX union operator

The operator $\bigvee$ is the union operator(S-Norm) in fuzzy set theory. One of the most popular union operators is the max operator. The use of max operator is valid both in the classical set theory union and fuzzy unions. In a matrix Φ, of membership grades, an element Φ_{ij} is the membership grade of the i^{th} sentence of the reference summary, in the j^{th} sentence of the candidate summary. This can also be written as $\Phi_{ij} = \mu_{t_j}(r_i)$. Hence elements along the i^{th} row correspond to the membership grades of this sentence in $|T|$ fuzzy sets, each of these sets representing a sentence from the candidate summary. Similarly the elements along the j^{th} column correspond to the membership grades of the corresponding sentence of the candidate summary, in $|R|$ fuzzy sets, each of these sets representing a sentence from the reference summary. The membership grade of the i^{th} sentence of the candidate summary can be computed using the max-union operator as $\mu_R(t_i) = \max_{j=1..|R|}\left(\mu_{r_j}(t_i)\right)$. Similarly to find the membership grade of the i^{th} sentence of the reference summary in the fuzzy union the relation $\mu_T(r_i) = \max_{j=1..|T|}\left(\mu_{t_j}(r_i)\right)$ can be used.

The top part of fig:-2 shows the membership grades of 21 sentences in 3 fuzzy sets. The bottom part of the figure shows the resultant membership grade of these 21 sentences after using the *max* s-norm operator. Because of the

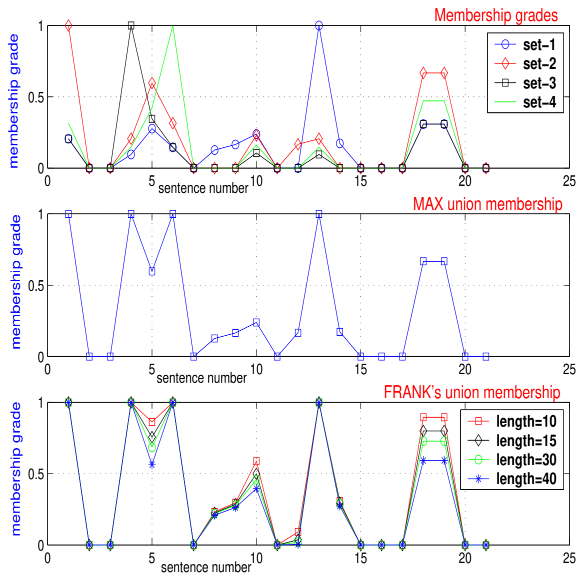

Figure 3: Behavior of FRANK's union operator with changing sentence length

nature of *max* s-norm operator, the score for a sentence in the reference summary can never exceed its maximum membership grade. This becomes a problem when a candidate sentence is a union of words in more than two sentences of the reference summary.

3.2 FRANK's S-Norm operator

The max union operator does not take into consideration the distribution of membership grades in various sets. It also is independent of the length of the sentence. Membership grade distribution and sentence length play an important role. Hence an union operator that non-linearly combines individual membership grades and at the same time confirms to all the laws of fuzzy union is an ideal choice. The $FRANK's$ union operator defined as

$$\left(\mu_A(x) \bigvee \mu_B(x)\right) = 1 - \log_F \left[1 + \frac{\left(F^{1-\mu_A(x)} - 1\right)\left(F^{1-\mu_B(x)} - 1\right)}{F-1}\right] \tag{3}$$

non-linearly combines the membership grade $\mu_A(x)$ of the element x in set A, and $\mu_B(x)$ of the element x in set B. The base of the logarithm F plays an important role in the combination process and is defined as $F \neq 1, F > 0$. To adopt $FRANK's$ union operation to summary evaluation, the value of F is defined as

$$F = \exp\left(-10 \times m \times \frac{S_L}{\max_L}\right)$$

where m is the mean of the non-zero membership grades of a sentence S, S_L is the length of this sentence in terms of the basic units and $\max_L$ is the length of the longest sentence in the set of sentences being evaluated. The result of using the exponential function for F can be observed in fig-3.

The first graph in this figure shows the membership grades of 21 sentences in 4 fuzzy sets (4 sentences of the reference summary). It can be observed that membership grade for some sentences is 1, in at least one of the sets. This is likely to be the case when a sentence matches exactly with another sentence, but also has partial matches with other sentences. There are also sentences whose membership grades are always less then 1 and have matches in all the 4 sets. This is likely to be the case when a single sentence has parts matching with multiple sentences. The second graph in fig-3 is the result of the union operation using the max union operator. The third graph in fig-3 shows the result of union operation using the $FRANK's$ union operator with exponential base F for various length of sentences. For example, sentence number 5 has membership grades $\{0.6, 0.4, 0.34, 0.28\}$ in the 4 fuzzy sets. The max union results in a membership grade of 0.6 for this sentence in the reference summary set and it is independent of the length of this sentence. But the $FRANK's$ union operation, results in membership values of $\{0.86, 0.75, 0.68, 0.56\}$ for sentence lengths of $\{10, 15, 30, 40\}$ words respectively. This means sentences which are shorter and have higher mean value computed using non-zero memberships, get a higher score after the union operation. This is very useful when the sentence being evaluated is a super set of a number of sentences in the reference set.

4 Comparison of Fuzzy evaluation with ROUGE

The fuzzy framework based evaluation method was compared with ROUGE by taking test cases. ROUGE has 3 classes of recall based scores, namely ROUGE-n, ROUGE-LCS and ROUGE-WLCS. ROUGE-n is an n-gram score where any two sentences having the same set of n-gram sequences are assumed to have contributed to a match. ROUGE-LCS is the LCS-based score while ROUGE-WLCS is a non-linearly weighted LCS score.

As the first case for comparison, a machine generated summary which is an exact replica of a human summary was used and recall-based ROUGE evaluation system was considered. As every sentence in the machine summary shall have an exact match with a sentence in the model(human generated) summary, we expect ROUGE scores to be 1 and the Fuzzy precision,recall and F-scores to be 1 as well. But ROUGE-WLCS does not have a normalized non-linear weighting as a result of which it gave a recall score of 0.35. Hence the absolute value of the ROUGE-WLCS score is misleading, as we would expect a score of 1 if there is an exact match. Exact values of these scores are listed in table-1.

Evaluation Type	Score Value
ROUGE-1	1
ROUGE-2	1
ROUGE-3	1
ROUGE-4	1
ROUGE-LCS	1
ROUGE-WLCS	**0.35764**
Fuzzy Precision	1
Fuzzy Recall	1
Fuzzy Fscore	1

Table 1: Reference and candidate summaries are the same

As a second case for comparison, sentences which are almost similar in word composition but not exactly the same were considered. For example consider the following model summary:

MODEL: JAMMU KASHMIR IS THE ONLY STATE WITH A MOSLEM MAJORITY IN PREDOMINANT (PREDOMINANTLY) HINDU INDIA. INDIA ACCUSE (ACCUSES) PAKISTAN OF ARM (ARMING) MILITANT (MILITANTS).

and the following candidate machine generated summary:

CAND-1: INDIA DOMINATE (DOMINATED) BY HINDU (HINDUS) ACCUSE (ACCUSES) PAKISTAN OF TRAIN (TRAINING) AND ARM (ARMING) KASHMIR MILITANT (MILITANTS) FIGHT (FIGHTING) FOR SECESSION OF JAMMU KASHMIR A MOSLEM MAJORITY STATE

CAND-2: HINDU DOMINATE (DOMINATED) INDIA ACCUSE (ACCUSES) PAKISTAN OF ARM (ARMING) AND TRAIN (TRAINING) KASHMIR MILITANT (MILITANTS) FIGHT (FIGHTING) FOR SECESSION OF JAMMU KASHMIR INDIA (INDIA'S) ONLY STATE WITH MOSLEM MAJORITY

CAND-3: ARM (ARMED) AND TRAIN (TRAINED) KASHMIR MILITANT (MILITANTS) FIGHT (FIGHTING)

FOR SECESSION ACCUSE MAJORITY OF MOSLEM (MOSLEMS) IN JAMMU KASHMIR STATE IN INDIA AND HINDU (HINDUS) IN PAKISTAN FOR THE PRESENT CONDITION

In this example we observe that each candidate summary has only one sentence. The sentence in first two candidate summaries is a union of information present in the two sentences of the model summary. But the third candidate summary presents a completely different information. Further, the candidate summaries are similar in word composition but different in word order. Table-2 shows the scores produced by ROUGE and Fuzzy evaluators. One of the draw backs of ROUGE can be observed from the results for candidates 2 and 3. In case of candidates 2 and 3 the longest common sub-sequence defaults to a sequence of non-contiguous words and hence becomes a bag of words. Tight word ordering is lost and candidates 2 and 3 are given almost the same score. On the other hand, fuzzy F-score uses bi -word and tri-word collocations and hence can easily capture not only longer sub-sequences but can also detect completely different meaning (owing to drastic change in word order). This is the reason for candidate-3 getting a very low fuzzy score.

Scoring Type (F-score)	Cand-1	Cand-2	Cand-3
ROUGE-1	0.65	0.69	0.63
ROUGE-2	0.18	0.3	0.18
ROUGE-3	0	0.12	0.09
ROUGE-4	0	0	0
ROUGE-LCS	0.51	0.56	0.4
ROUGE-WLCS	0.25	0.29	0.21
Fuzzy bi-word window	0.23	0.36	0.07
Fuzzy tri-word window	0.22	0.324	0.09

Table 2: Comparison of Evaluators for case-2

More over if the language is known and if a WordNet dictionary is available, it is possible to normalize words in any two sentences to include the same target synonyms and then apply the Fuzzy evaluation. Synonyms can be selected based on the sense number assigned by the WordNet. The assigned sense number could be weighted and made part of membership function computation. The two cases of evaluation discussed so far shows that ROUGE-WLCS penalizes extracts more than required and ROUGE-LCS score is poorer in cases where sentences in the candidate extract have words in a different order.

As the third case for comparison we see if there is any relationship between ROUGE scores and Fuzzy scores. If there exists a relationship then we can conclude that ROUGE and Fuzzy scores are consistent although they may vary in the exact value of the score. This is important as it has been demonstrated by the proposers of ROUGE that it closely resembles human evaluation. Fig-4 shows F-scores produced by ROUGE-LCS, ROUGE-WLCS, Fuzzy match and Exact sentence match. It can be observed that F-scores for ROUGE-WLCS and Fuzzy match vary similarly and this indicates consistency in scores produced by the two techniques. But there seems to be a deviation from this observation for some cases of ROUGE-LCS. This probably can be attributed to case-2 described previously.

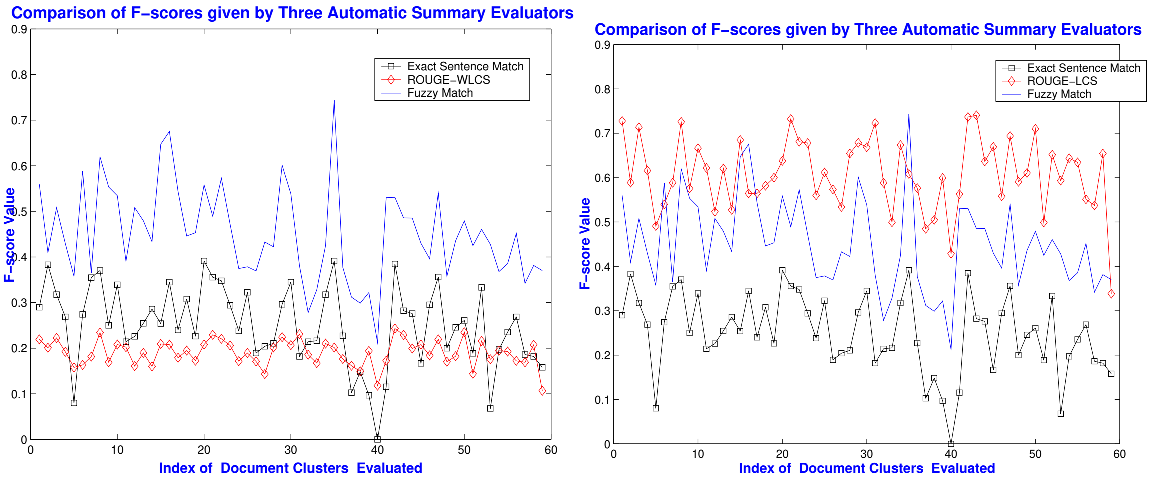

Figure 4: Comparison of F-scores of ROUGE, Fuzzy and Exact-match evaluators

5 Conclusion

In this paper we discussed issues regarding automatic evaluation of summaries. It was shown through argument that F-score based measures are better than just recall based scores. Further a fuzzy variant of F-score using modified Frank's union operator was described and its performance was compared to ROUGE scores. It was suggested that proposed fuzzy F-score based evaluation method can be modified further to include WordNet dictionaries and word sub-sequences as basic units so that the accuracy of evaluation improves. It would also be interesting to use linguistic methods for assigning membership grades although language independent evaluation cannot be guaranteed. Use of a probabilistic membership score augmented with linguistic queues can be experimented with, but it might wrongly bias the automatic evaluator. Hence use of a membership function based on frequency is a safer option. Further, use of cosine distance is intuitively appealing as a degenerate case would immediately lead to exact sentence match.

References

[Chin-Yew and E.Hovy, 2003] Chin-Yew and E.H, H. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*. Edmonton, Canada.

[Donaway et al., 2000] Donaway, R. L., Drummey, K. W., and Mather, L. A. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, U. Hahn, C.-Y. Lin, I. Mani, and D. R. Radev, Eds. Association for Computational Linguistics, 69–78.

[Lin, 2001] Lin, C.-Y. 2001. Summary Evaluation Environment. http://www.isi.edu/~cyl/SEE.

[Lin and Hovy, 2003] Lin, C.-Y. and Hovy, E. 2003. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In *Proceedings of the Human Language Technology Conference*.

[Mani et al., 1998] Mani, I., House, D., Klein, G., Hirshman, L., Orbst, L., Firmin, T., Chrzanowski, M., and Sundheim, B. 1998. The TIPSTER SUMMAC Text Summarization Evaluation. Tech. Rep. MTR 98W0000138, The Mitre Corporation, McLean, Virginia.

[Papineni et al., 2001] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2001. BLEU: A Method for Automatic Evaluation of Machine Translation. Research Report RC22176, IBM.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 4Oth Meeting of the Association for Computational Linguistics*. Philadelphia, USA, 311–318.

[Radev et al., 2002] RADEV, D. R., TEUFEL, S., SAGGION, H., LAM, W., BLITZER, J., ÇELEBI, A., QI, H., DRABEK, E., AND LIU, D. 2002. Evaluation of Text Summarization in a Cross-lingual Information Retrieval Framework. Tech. rep., Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD. June.

[Saggion et al., 2002] SAGGION, H., RADEV, D., TEUFEL, S., AND LAM, W. 2002. Meta-Evaluation of Summaries in a Cross-Lingual Environment Using Content-Based Metrics. In *Proceedings of COLING-2002*. Taipei, Taiwan.