Evidence Based Approach for Sentence Extraction from Single Documents

Sukanya Manna, Tom Gedeon, B. Sumudu U. Mendis, Richard L. Jones

School of Computer Science, The Australian National University, ACT 0200, Canberra, Australia
{sukanya.manna, tom, sumudu, richard.jones}@cs.anu.edu.au

Abstract: We present an evidence based sentence extraction model which is an application of subjective logic in a document computing scenario, to rank sentences according to their importance in a document. Elements from the Dempster-Shafer belief theory are used by this model to measure the subjective belief or opinion about a sentence. The important sentences extracted by this model can be seen to summarize a document partially. For qualitative analysis, this method is compared with two different open source summarizers along with human extracted sentences which are used as benchmarks for this purpose. This model also improves the effect of signal to noise ratio on sentence rank by applying the whole evidence based model on a reduced data set to evaluate its stability and accuracy. Since evidence based models are computationally very expensive, here we show that one third of the words of a document are sufficient to rank sentences similarly to human judgements, but if reduced further, the accuracy drops. The results show that our evidence based model outperforms standard summarizers when evaluated with human ranked sentences.

Keywords: subjective logic, evidence theory, summarization, sentence extraction, uncertain probability, summarization

1. Introduction

Intelligence analysis is a complicated time critical task which requires attention and a high degree of analytical judgement under considerable uncertainty. It generally requires that analysts choose from several alternative hypotheses in order to present the most plausible of these as likely explanations or outcomes for the evidence being analyzed [18]. In a wider context, people make their decisions based on subjective information which is rarely completely certain and reliable. To handle such a scenario, such as analyzing a

single document, we require some form of subjective data analysis as there is no volume information available about the source data. In this paper, our main motivation is to show how subjective belief works on single documents for sentence classification and also to show the effect of reduced available information on the analysis. Our approach is to deduce relative sentence importance based on frequency plus inter-sentence interaction which is determined by subjective logic.

Standard logic deals with propositions which are either true or false. This is very unlikely to be useful in a human situation where a condition cannot be determined with absolute certainty whether that proposition is true or false. There are other alternative logics which handle uncertainty and ignorance and have been applied practically to solve problems where there is insufficient evidence [5], [12]. Probabilistic logic was defined by Nillson [13] with the aim of combining the capacity of deductive logic to exploit the structure and relationship of arguments and events, with the capacity of probability theory to express degrees of truth about those arguments and events. Belief theory represents the extension of classical probability by making explicit the expression of ignorance i.e., lack of information, by assigning belief mass to the whole state space [17]. Classical belief representation is quite general, and allows complex belief structures to be expressed on arbitrary large state spaces as seen in Dempster-Shafer theory, which addresses interaction based on the evidence [17]. The main idea behind belief theory is to abandon the additivity principle of probability theory. Instead, belief theory gives observers the ability to assign so-called belief mass to any subset of state spaces. A limitation of this model lies in the combination of evidence which may lead to counterintuitive conclusions after applying normalization [19]. To overcome this limitation, we use Jøsang's [6] model of subjective belief, as it has a simpler representation of belief functions called *'opinions'*, which can be easily mapped to probability density function.

Subjective logic [6] operates on subjective beliefs about the world and uses *opinion* to denote the representation of a subjective belief. An opinion can be interpreted as a probability measure containing secondary uncertainty, and thus subjective logic can be seen as an extension to both probability calculus and binary logic [7]. It can be seen that real world situations are more realistically interpreted and analyzed using this subjective logic when applied manually. An aim of our paper is to apply this subjective logic automatically to rank sentences from a document according to their importance. The concepts of belief and disbelief [6] have been incorporated to measure the uncertainty. The probability expectation of the sentences form the scores that rank the sentences by their importance with respect to their context.

In [6], subjective logic presented by Jøsang is used to model and analyze real world situations realistically; these real world situations are instances of open environments which have no specific limitations on evidence to be gathered for giving an opinion about

a hypothesis. On the other hand, in a document computing environment, the evidence collected is from the document which is a closed environment. A document consists of sentences, and each sentence consists of words. These form the basis for the evidence which we either get directly or by deriving from them. If there are n words in a document, then we have 2^n possible states (or combinations of words) which might exist in the whole document. The co-occurrence of words found in a sentence represents evidence which we derive from existing words. These represent non-atomic events belonging to 2^n. Now, we can consider belief of a sentence to refer to the total number of states present in that sentence i.e, words or combinations of words existing in the sentence represent evidence. This presents how much information we get from a sentence about the whole document. Disbelief of a sentence in this context can be stated as the words which are not present in that sentence as well as are not present in other sentences with which it has some words in common. This is 'ignorance' in Dempster-Shafer theory; so disbelief does not have any role in supporting a sentence or a hypothesis. Uncertainty of a sentence is the evidence which is plausible to support it. This means, if the sentence has some words in common with another, then the interaction of the words in the other sentence will have some contribution to the meaning of this sentence. In this context, interaction means all possible combinations of words which are present, either in the sentence being considered, or any other sentences in the document.

Arguments in subjective logic are called "subjective opinions" or "opinions" for short. An opinion can contain degrees of uncertainty in the sense of "uncertainty about probability estimates". The uncertainty of an opinion can be interpreted as ignorance about the truth of the relevant states, or as second order probability about the first order probabilities. Thus, opinion about a sentence presents the importance of that sentence in a given context containing degrees of uncertainty.

In this paper, besides finding the importance of sentences, we simultaneously reduce the complexity of the model by reducing information for analysis without significant loss of accuracy. It is known that most evidence based models are computationally expensive with the increase in size of the input space. We investigate the reduction (or purification) of the available information and its effect on sentence ranking using a reduced word set for the whole analysis. The quality of the top ranked sentence extracted is evaluated by comparing them with human ranked sentences of the same documents and also with open source summarizers.

The detailed implementation, modification and assumption of this application of the subjective logic based model are presented in sec. 2, followed by a detailed evaluation.

2. Modeling uncertain probabilities for sentence ranking

In this section we present the model for uncertain probabilities [6] for sentence ranking. The parameters of this model are defined using sentences in the form of hypotheses. Words (or terms) occurring in a sentence in a document are facts or evidence available to support or weaken the hypothesis. So the truth of evidence of the sentences is formulated using the given words or co-occurrence of words. Here *three basic assumptions* are made to proceed with this model:

1. All the words or terms (removing the stop words) in the document are atomic.
2. The sentences are unique, i.e., each of them occur only once in the given document.
3. It is a closed system where the evidence is confined within a single document.

A document consists of sentences. In this paper, a sentence is considered to be a set of words separated by a stop mark (".", "!", "?"). Non stop words are extracted and the frequencies (i.e. number of occurrences) of the words in each sentence are calculated.

Let us now define the notations which we will be using in the rest of the equations and explanations. Θ is the frame of discernment. We represent a document as a collection of words, which is

$$\Theta = D_w = \{w_1, w_2, ..., w_n\} \tag{1}$$

where, D_w is a document consisting of words w_1, $w_2 ... w_n$ and $|D_w| = n$. Now,

$$\rho(\Theta) = \{\{w_1\}, \{w_2\}, ..., \{w_1, w_2, w_3, ..., w_n\}\} \equiv 2^{\Theta} \tag{2}$$

$$|\rho(\Theta)| = 2^n, \tag{3}$$

where ρ represents the power set of the elements of Θ. We can also represent a document as a collection of sentences,

$$D_s = \{s_1, s_2, ..., s_m\} \tag{4}$$

where m is a finite integer and each s_i is an element of $\rho(\Theta)$. Each sentence is comprised of words, which belong to the whole word collection of the document D_w. We thus represent each sentence S_l by,

$$S_l = \{w_i, w_k, ..., w_r\} \in \Theta \tag{5}$$

where, $1 \leq i, k, r \leq n$ and $S_l \in \rho(\Theta)$.

The Belief Model The representation of uncertain probabilities [6] is based on a belief model similar to the one used in Dempster-Shafer theory of evidence. Initially a set of possible situations, frame of discernment are defined as in (1). It is assumed that the system cannot be in more than one elementary state at the same time. The elementary states in the frame of discernment Θ will be called atomic states because they do not contain substates.

Here, all the non stop words of the document are considered to be atomic and they are the elements of frame of discernment. The powerset of Θ, denoted by 2^{Θ}, contains the atomic states and all possible unions of the atomic states including Θ; this is the pattern of the words' occurrence or co-occurrence of words in the document. Sentences are events with non-atomic states. Similarly, co-occurrence of words represent sub-events (represented by non-atomic states as well). In this work, we actually support events using atomic and non-atomic states considering them to be sources of evidence within the document.

Suppose, we have a document D (fig.1) with 4 sentences, s_1, s_2, s_3, and s_4 and 5 words, w_1, w_2, w_3, w_4, and w_5 respectively. So the all possible states in the document will be 2^5 which are as follows:
$\{\emptyset, \{w_1\}, \{w_2\},..., \{w_1,w_2\}, \{w_2,w_3\},..., \{w_1,w_2,w_3\}, \{w_2,w_3,w_4\},..., \{w_1,w_2,w_3,w_4\},$
$... \{w_1,w_2,w_3,w_4,w_5\}\}$.
We consider only the ones which occur at least once in the document.

Now, the events with countable evidence are only considered for the calculations and a belief mass is assigned to each event.

Definition 1 (Belief Mass Assignment) *Let Θ be a frame of discernment. If with each substate $x \in 2^{\Theta}$ a number $m_{\Theta}(x)$ is associated such that:*
1. $m_{\Theta}(x) \geq 0$
2. $m_{\Theta}(\emptyset) = 0$
3. $\sum_{x \in 2^{\Theta}} m_{\Theta}(x) = 1$
then m_{Θ} is called a belief mass assignment in Θ, or BMA for short. For each substate $x \in 2^{\Theta}$, the number $m_{\Theta}(x)$ is called the belief mass of x.

Belief Mass Assignment (BMA) is defined here in the same way as modeled by Jøsang [6]. We also call this probability of evidence. Let Θ be a frame of discernment. If with each substate $x \in 2^{\Theta}$ a number m_{Θ} is associated such that:

1. $m_{\Theta}(x) \geq 0$
2. $m_{\Theta}(\Phi) = 0$
3. $\sum_{x \in 2^{\Theta}} m_{\Theta}(x) = 1$

In fig. 1, we depict the above example. The state space contains events that form evidence for the problem. Each of these words w_1, w_2, w_3, w_4, and w_5 are atomic states, and each of these sentences s_1, s_2, s_3 and s_4 are non atomic states (events). All these states are from power set of the frame of discernment. In this case, the possible states we get from the

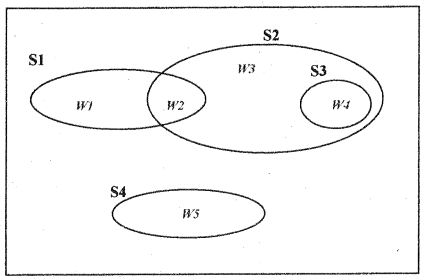

Figure 1: Example showing the occurrence of words in the sentences

example are:
$\{\{w_1\}, \{w_2\}, \{w_3\}, \{w_4\}, \{w_5\}, \{w_1,w_2\}, \{w_2,w_3\}, \{w_3,w_4\}, \{w_2,w_4\}, \{w_2,w_3,w_4\}\}$.

We calculate BMA for each event by,

$$m(x) = \frac{F(x)}{Z}, \tag{6}$$

where $F(x) = \sum_{k=1}^{N} f_{x_k}$, where N is the total number of sentences in the document, $x \in 2^{\Theta}$, and f_{x_k} is the frequency of occurrence of event x in sentence k. In words, it is the total frequency of that event in all the sentences (or the whole document).

$$Z = \sum_{\substack{\forall x \neq \emptyset \\ f_x \neq 0}} F(x), \quad x \in 2^{\Theta} \tag{7}$$

Z is the total frequency of the all the events which has valid evidence of truth (whose frequency is non zero). In this example, as shown in 1, let us assume that frequency of each of these words be 1 in each sentence. So, $Z = 12$. Now, we can see that s_1 has two words, w_1 and w_2. We have three different sub-states with non zero evidence; $m(w_1)$, $m(w_2)$, and $m(w_1,w_2) = m(s_1)$.

Definition 2 (Belief Function) *Let Θ be a frame of discernment, and let m_Θ be a BMA on Θ. Then the belief function corresponding with m_Θ is the function $b : 2^{\Theta} \rightarrow [0,1]$ defined by:*

$$b(x) = \sum_{y \subseteq x} m_\Theta(y), \quad x,y \in 2^{\Theta} \tag{8}$$

Now, in context to the example, we calculate the belief of a sentence, $b(s_1) = m(w_1) + m(w_2) + m(w_1,w_2)$. Similarly, an observer's disbelief must be interpreted as the total belief that a state is not true.

Definition 3 (Disbelief Function) *Let Θ be a frame of discernment, and let m_Θ be a BMA on Θ. Then the disbelief function corresponding with m_Θ is the function $d: 2^\Theta \rightarrow [0,1]$ defined by:*

$$d(x) = \sum_{y \cap x = \emptyset} m_\Theta(y), \quad x,y \in 2^\Theta. \tag{9}$$

If we now consider the example, we calculate disbelief of s_1 by $d(s_1) = m(w_3) + m(w_4) + m(w_3,w_4) + m(w_5)$.

Definition 4 (Uncertainty Function) *Let Θ be a frame of discernment, and let m_Θ be a BMA on Θ. Then the uncertainty function corresponding with m_Θ is the function $u: 2^\Theta \rightarrow [0,1]$ defined by:*

$$u(x) = \sum_{\substack{y \cap x \neq \emptyset \\ y \not\subseteq x}} m_\Theta(y), \quad x,y \in 2^\Theta. \tag{10}$$

From Josang's research concept, we can get **Belief Function Additivity** which is expressed as:

$$b(x) + d(x) + u(x) = 1, \quad x \in 2^\Theta, x \neq \emptyset. \tag{11}$$

One can simply calculate the uncertainty of a sentence by using (11), i.e., $u(s_1) = 1 - (b(s_1) + d(s_1))$.

Definition 5 (Relative Atomicity) *Let Θ be a frame of discernment and let $x,y \in 2^\Theta$. Then for any given $y \neq \emptyset$ the relative atomicity of x to y is the function $a: 2^\Theta \rightarrow [0,1]$ defined by:*

$$a(x/y) = \frac{|x \cap y|}{|y|}, \quad x,y \in 2^\Theta, y \neq \emptyset. \tag{12}$$

In this case, we get the following relative atomicity for sentence s_1 as:

$$a(s_1/w_1) = \frac{|s_1| \cap |w_1|}{|w_1|} = \frac{1}{1} = 1$$
$$a(s_1/w_2) = \frac{|s_1| \cap |w_2|}{|w_2|} = \frac{1}{1} = 1$$
$$a(s_1/\{w_1,w_2\}) = a(s_1,s_1) = \frac{|s_1| \cap |\{w_1,w_2\}|}{|\{w_1,w_2\}|} = \frac{2}{2} = 1$$
$$a(s_1/w_3) = \frac{|s_1| \cap |w_3|}{|w_3|} = \frac{0}{1} = 0$$
$$a(s_1/w_4) = a(s_1/s_3) = \frac{|s_1| \cap |w_4|}{|w_4|} = \frac{0}{1} = 0$$

$a(s_1/\{w_2,w_3\}) = \frac{|s_1|\cap|\{w_2,w_3\}|}{|\{w_2,w_3\}|} = \frac{1}{2}$
$a(s_1/\{w_2,w_3,w_4\}) = a(s_1/s_2) = \frac{|s_1|\cap|\{w_2,w_3,w_4\}|}{|\{w_2,w_3,w_4\}|} = \frac{1}{3}$
$a(s_1/w_5) = a(s_1/s_4) = \frac{|s_1|\cap|w_5|}{|w_5|} = \frac{0}{1} = 0$
Likewise, we calculate the atomicity for other sentences.

Definition 6 (Probability Expectation) *Let Θ be a frame of discernment with BMA m_Θ then the probability expectation function corresponding with m_Θ is the function $E: 2^\Theta \rightarrow [0,1]$ defined by:*

$$E(x) = \sum_y m_\Theta(y)a(x/y), \quad y \in 2^\Theta. \tag{13}$$

So, for the given example, we calculate *ProbExp* for sentence s_1 as follows:
$E(s_1) = m(w_1)a(s_1/w_1) + m(w_2)a(s_1/w_2) + m(\{w_1,w_2\})a(s_1/\{w_1,w_2\}) + ... + m(w_5)a(s_1/w_5)$

We calculate PE of each sentence using (13). We consider sentences to be important if they have higher probability expectation and lower uncertainty. By doing this, we have seen that sentences with higher PE and lower uncertainty, have more words interacting with other sentences.

Definition 7 (Opinion) *Let Θ be a binary frame of discernment with 2 atomic states x and $\neg x$, and let m_Θ be a BMA on Θ where $b(x)$, $d(x)$, $u(x)$, and $a(x)$ represent the belief, disbelief, uncertainty and relative atomicity functions on x in 2^Θ respectively. Then the opinion about x, denoted by w_x is the tuple defined by:*

$$w(x) \equiv (b(x), d(x), u(x), a(x)). \tag{14}$$

For compactness and simplicity of notation we will in the following denote belief, disbelief, uncertainty and relative atomicity functions as b_x, d_x, u_x and a_x respectively. Thus opinion about a sentence s_1 can be expressed using these four parameters as, $w(s_1) = (b(s_1), d(s_1), u(s_1), a(x))$.

3. Experiment

In this paper, Jøsang's subjective logic [6] is implemented in a document computing scenario to classify sentences in a document according to their importance. The motivation and context of the paper is different from that of [6], where the idea was to find opinion about an unknown event from the known ones at hand. But here, we are building evidence from a document to support an event based on its importance. The concept of evidence

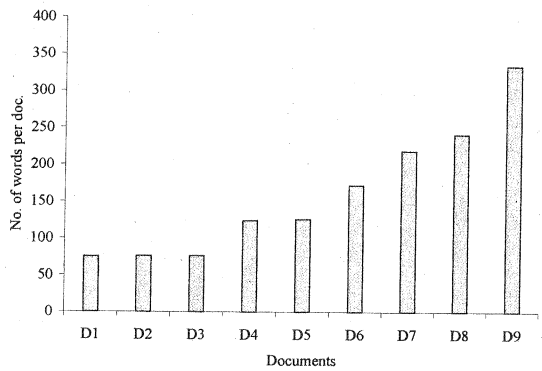

Figure 2: Number of words per document

is used in the form of word occurrence and its interaction with other words in a sentence and probability expectation of a sentence is calculated using (13) and (10) to calculate uncertainty. Sentences are then arranged according to descending probability expectation and ascending uncertainty to rank then according to their importance.

3.1. Data processing

The experiment is carried out using different Cross Document Structure Theory (CST) data sets [15]. Each data set consists of documents related to a specific topic such as plane crash, space shuttle mishap, and so on; consisting of fewer documents ranging from nine to ten or eleven. Our main aim is to see how this model works on single documents for content analysis purposes, so we focussed on this kind of data set unlike other information retrieval areas. Among the results, we present here are documents related to a Milan plane crash which consists of nine single documents. These documents were parsed, tokenized, cleaned, and stemmed. The cleaning is done by removing the stop words. The term list is generated from each of the documents. The documents are comparatively of moderate length as shown in fig.2.

3.2. Method

The subjective logic implementation of sentence classification has exponential time complexity. So, we prepared a reduced set of data to see the effects of the algorithms to extract sentences with them accurately. Two data sets were created for each documents; one with top 25 words and the other with every third word, together 25 in number. First, we found the frequency of occurrence of the words (excluding the stop-words) from a document. Then we arranged them in descending order of their occurrence and chose the top 25 words; as well as choosing every third element from the list keeping 25 here also. According to Gedeon *et al.*, [3] choosing every third word is appropriate as we suspect there is significant noise in the document so eliminating 2 out of 3 words can improve the model by increasing the signal to noise ratio so long as enough points remain to detect the underlying trend.

For shorter length documents, we kept approximately one third of the total number of words (approximately 25), then we gradually decreased to one sixth and then one ninth for the longer documents in order to maintain the count approximately to 25. It is seen that word count after the top 20 words have frequency of 1 in the word list for each documents. Statistically these words with count one will have similar contribution in the documents. So, this is another reason to analyze the effect of reduced word list for determining the sentence ranking.

We computed the probability expectation of each sentence and uncertainty using (13) and (10) respectively. We ranked each sentence with increasing probability expectation (PE) and decreasing uncertainty; the higher the PE and the lower the uncertainty, the greater is the importance of the sentence in the documents thus assigning higher rank to it.

Our main motivation is to extract important sentences from a document and use them for content analysis. To analyze the quality of the sentences extracted, we need some methods to proceed with the evaluation. So in the next phase, we perform the evaluation of this model where human accessors were involved to mark the important sentences which are used as a benchmark. This is then compared with two different open source summarizers MEAD [14] and OTS [1] respectively as also used by [1] for their evaluation.

3.3. Generation of Summaries

Summaries are broadly classified into text extraction and text abstraction [10], [8]. For text extraction, sentences from the documents are used as summaries and for text abstraction important pieces of information are extracted and then stitched together to form summaries

[1] `http://libots.sourceforge.net/`

following some linguistic rules. This evidence based model can be used as a text extraction as we use the original sentences like MEAD [14], an open source summarizer and OTS. Both these methods are used as a benchmark [1] for evaluation of summaries before.

Extraction of top ranked sentences play partially the role of summarization, so for qualitative analysis of our work, we compared the sentences extracted with open source tools such as MEAD and OTS.

Available summarizers *MEAD* [14] is a publicly available toolkit for multi-lingual summarization and evaluation. The toolkit implements multiple summarization algorithms (at arbitrary compression rates) such as position-based, Centroid, TF*IDF, and query-based methods. Methods for evaluating the quality of the summaries include co-selection (precision/recall, kappa, and relative utility) and content-based measures (cosine, word overlap, bigram overlap). We used single documents to summarize using MEAD.

The Open Text Summarizer (OTS) is an open source tool for summarizing texts. The program takes a text and decides which sentences are important and which are not. It ships with Ubuntu, Fedora and other linux distributions. OTS supports many (25+) languages which are configured in XML files. There is published researche on summarization, where OTS is used as a benchmark to evaluate the performance of summaries generated by their method [1], [16]. So we also used it for our comparative study for performance evaluation.

Human ranked sentences For the evaluation of our method, we involved two human assessors RJ and BP. We gave each of them the sets of documents. We asked them to rank 30% [2] of the important sentences as they read through the text. We then collected those sentences, and then formed extractive summaries maintaining the ranks assigned by them to the sentences.

Evidence based model (ProbExp): In sec.2, we described the methods of sentence ranking; subjective logic based where we ranked the sentences based on the probability expectation (ProbExp) and uncertainty of that sentence in that document. We ranked the sentences according to their importance by descending PE value and ascending uncertainty value. We took 30% [2] of the top ranked sentences. We consider these to represent the summary of the whole document. We then compared this top 30% with the summaries generated by the above summarizers (human as well as automated) to perform qualitative evaluation.

3.4. Evaluation

ROUGE evaluation ROUGE [9] stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. ROUGE is a recall based metric for fixed length summaries. The measures count the number of over lapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans.

For this experiment, we used both machine generated summaries as well as human generated summaries to compare to our evidence based approach.

In this experiment, we present the result with ROUGE-1 (n-gram, where n=1) at 95% confidence level. ROUGE is sensitive to the length of the summaries [11]. The results vary when the length of the summaries of the peer and the model differs. We have shown here both kinds of results:
1. the usual convention of ROUGE by fixing the peer and model summary length to 100;
2. varying the length of the summaries, simply taking 30% of the top ranked sentences of a document as summary.

The figures 4 to 8 present the recall curve whereas the tables 1 to 4 present the average recall, precision and F-measure [9] of the comparisons of all documents.

Results In this part, summarization evaluation results are discussed. In the figures 3 to 8 and tables 1 to 4 some abbreviations are used, which are as follows:
$ProbExp_{top}$: Probability Expectation with top words (reduced word set)
$ProbExp_{\frac{1}{3}}$: Probability Expectation with every third word (reduced word set)
$Human_{BP}$: Refers to the human assessor BP
$Human_{RJ}$: Refers to the human assessor RJ

We mentioned that ROUGE is sensitive to the length of the summaries being compared. So, two different forms of ROUGE evaluation are presented: by limiting the word length of peer and model summaries to 100 (subsec.3.4.1); and without any word limitation (subsec.3.4.2). In both cases it is seen that the performance of our evidence based model is as good as human ranked sentences than automated summarizers: OTS and MEAD in particular. The results for $ProbExp_{\frac{1}{3}}$ show more consistent performance than $ProbExp_{top}$. When these two are compared with other automated summarizers like MEAD and OTS, $ProbExp_{top}$ results are more similar to these than $ProbExp_{\frac{1}{3}}$. This suggests that MEAD and OTS are more focussed on important words in a document. The tables 1 to 4 present the overall performance of each summarization method with human assessors' one. $ProbExp_{\frac{1}{3}}$ shows consistent performance in all the cases.

3.4.1 Evaluation by limiting the length of summaries to hundred words

Two different sets of results are shown in this section. The first is the comparison of the machine generated summaries with two different human assessors $Human_{BP}$ (fig.3) and $Human_{RJ}$ (fig.4). The second is the comparison of our evidence based method (with top words, $ProbExp_{top}$ and with every third, $ProbExp_{\frac{1}{3}}$) with MEAD and OTS (fig.5).

In this case, our evidence based model $ProbExp$ performs very similar to human generated summaries, $Human_{BP}$ and $Human_{RJ}$ than other automated summarizers, MEAD and OTS. The average performance $ProbExp_{\frac{1}{3}}$ is better than $ProbExp_{top}$, showing improvement of signal to noise ratio with reduction of words. Degradation of performance in our model is also noticed with excess removal of words. When $ProbExp$ is compared with MEAD and OTS, $ProbExp_{top}$ has higher similarity with them than $ProbExp_{\frac{1}{3}}$.

Comparison between evidence based model with human assessors We can see in fig.3 that $ProbExp_{\frac{1}{3}}$ and $ProbExp_{top}$ are more similar to human judgement given by $Human_{BP}$. Of the other two automated ones, MEAD and OTS, OTS performs better. It is closer to $ProbExp_{\frac{1}{3}}$ and $ProbExp_{top}$. Here the result for $ProbExp_{\frac{1}{3}}$ is better than $ProbExp_{top}$, since taking every third word increases the signal to noise ratio and purifies the ranking. Now, if we look at the length of the documents in fig.2, we find the number of words increasing with the documents. The performance of $ProbExp_{top}$ and $ProbExp_{\frac{1}{3}}$ initially outperformed the other two methods for the first three documents. The performance degraded from document 4 onwards. This is because the number of words per document started increasing from document 4 onwards (see fig.2), so in these documents approximately one sixth and one ninth of the words are considered for sentence ranking using our evidence based model. In this situation there is higher loss of useful information with greater reduction of words. But still it outperforms MEAD except for the last document, where we lost the maximum amount of information while reducing words to one ninth.

Fig.4 shows a similar effect to that seen in fig.3. $ProbExp_{top}$ as well $ProbExp_{\frac{1}{3}}$ is very close to human $Human_{RJ}$. Here again the drop in performance by our method is noticed for the documents with higher number of words. From these two figures we can say that if we consider one third of the total number of words, this evidence based model can give us close results to human judgement.

Comparison between evidence based model with automated summarizers In fig.5, $ProbExp_{top}$ and OTS have higher overlap than $ProbExp_{\frac{1}{3}}$ and OTS. $ProbExp_{top}$ and MEAD comes next to $ProbExp_{top}$ and OTS in terms of performance. Here too the performance degradation is observed as seen in figures 3 and 4.

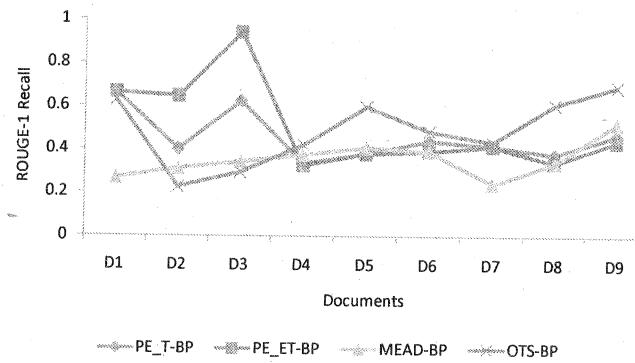

Figure 3: ROUGE-1 recall for Assessor BP with different automated methods

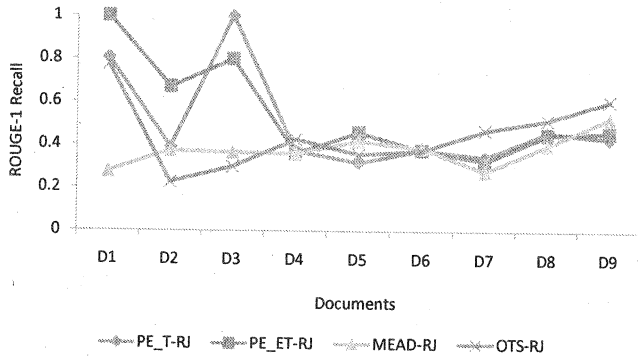

Figure 4: ROUGE-1 recall for Assessor RJ with different automated methods

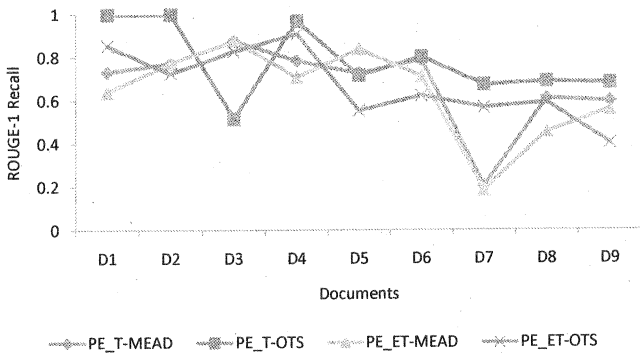

Figure 5: ROUGE-1 recall for evidence based model (PE) with automated summarizers

Table 1: Average ROUGE-1 (word limit=100) Recall(R), Precision(P) and F-measure(F) of all documents by comparing $Human_{BP}$ with different automatic summarization methods

	Avg R	Avg P	Avg F
$ProbExp_{top}$	0.46	0.50	0.47
$ProbExp_{\frac{1}{3}}$	0.51	0.51	0.51
MEAD	0.36	0.40	0.37
OTS	0.49	0.60	0.53

It should be noticed that using only one third of the words, our evidence based model works better than standard summarizers when compared with human assessors. The performance is boosted when we increase the signal to noise ratio by taking every third word from the document word list. But, summarizers like OTS and MEAD are more significant keyword focussed, so fig.5 shows our $ProbExp_{top}$ has higher similarity with them than $ProbExp_{\frac{1}{3}}$.

Tables 1 and 2 present the average recall, precision, and F-measure of all the documents for both human assessors' summaries with automated ones when ROUGE parameter is

Table 2: Average ROUGE-1 (word limit=100) Recall(R), Precision(P) and F-measure(F) of all documents by comparing $Human_{RJ}$ with different automatic summarization methods

	Avg R	Avg P	Avg F
$ProbExp_{top}$	0.50	0.49	0.49
$ProbExp_{\frac{1}{3}}$	0.55	0.48	0.51
MEAD	0.38	0.39	0.38
OTS	0.45	0.50	0.47

fixed to 100 words for evaluation. In both the tables similar results are noticed. As shown in the figures 3 and 4, here too in the tables, summaries generated by $ProbExp_{\frac{1}{3}}$ are most similar to humans than $ProbExp_{top}$ and then followed by OTS and MEAD.

3.4.2 Evaluation without any specific word limit

In this part, we present ROUGE evaluation results without limiting the lengths of model and peer summaries. The summaries are are 30% of the total length of a document. We found that ROUGE score tends to increase with increases in the length of the summaries. Like the previous evaluations for fixed length summaries, here the same comparisons are presented without fixing the length. The comparison results show that our evidence based models behave more similarly with human assessors' than MEAD and OTS. $ProbExp_{\frac{1}{3}}$ is even better than $ProbExp_{top}$ (like subsec.3.4.1). Similar degradation of performance is noticed here like the fixed summary length evaluation results (see subsec.3.4.1). In the tables 3 and 4, similar results are observed when averaged over all documents.

Comparison between evidence based model with human assessors In fig,6, $ProbExp_{\frac{1}{3}}$ and $ProbExp_{top}$ are more similar to human assessor BP in terms of overlap than the other two automated standard summarizers. Though there is performance degradation due to reduced data size to one sixth and then to one ninth for the documents with ascending word length, still $ProbExp_{top}$ and $ProbExp_{\frac{1}{3}}$ outperforms others. $ProbExp_{\frac{1}{3}}$ is best among all. The reason for this case is the same as in the other figures 3 to 5, due to purification of words increasing the signal to noise ratio.

In fig.7 similar behaviour is observed like fig.6. For majority of the documents, $ProbExp_{\frac{1}{3}}$ is higher than the other models and has maximum overlap with $Human_{RJ}$.

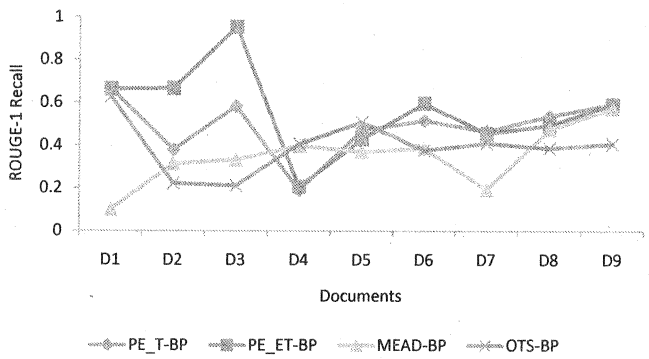

Figure 6: ROUGE-1 recall for Assessor BP with different automated methods

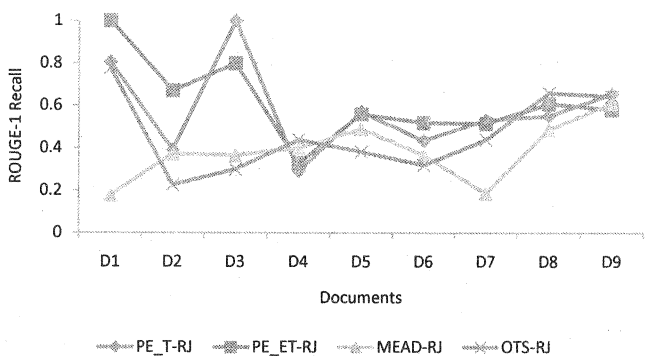

Figure 7: ROUGE-1 recall for Assessor RJ with different automated methods

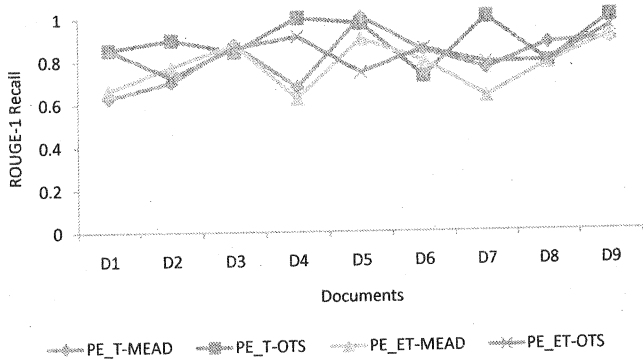

Figure 8: ROUGE-1 recall for evidence based model (PE) with automated summarizers

Table 3: Average ROUGE-1 (without specific word limit) Recall(R), Precision(P) and F-measure(F) of all documents by comparing $Human_{BP}$ with different automatic summarization methods

	Avg R	Avg P	Avg F
$ProbExp_{top}$	0.49	0.50	0.47
$ProbExp_{\frac{1}{3}}$	0.56	0.53	0.53
$MEAD$	0.35	0.46	0.38
OTS	0.40	0.64	0.48

Comparison between evidence based model with automated summarizers In fig.8, the picture is a bit different. $ProbExp_{top}$ has maximum overlap with OTS than MEAD. $ProbExp_{\frac{1}{3}}$ comes next in terms of overlap with OTS than MEAD. Here the performance degradation of $ProbExp_{top}$ and $ProbExp_{\frac{1}{3}}$ is not obvious.

Like tables 1 and 2, tables 3 and 4 present the average recall, precision and F-measure of all the documents for both human assessors' ($Human_{BP}$ and $Human_{RJ}$) generated summaries with automated ones without limiting word limits on ROUGE parameter. In both the tables, $ProbExp_{\frac{1}{3}}$ results are consistent for and similar to human assessors. $ProbExp_{top}$ is next to this. But overall performance of our evidence based model is higher than OTS and MEAD.

Table 4: Average ROUGE-1 (without specific word limit) Recall(R), Precision(P) and F-measure(F) of all documents by comparing $Human_{RJ}$ with different automatic summarization methods

	Avg R	Avg P	Avg F
$ProbExp_{top}$	0.64	0.45	0.51
$ProbExp_{\frac{1}{3}}$	0.63	0.44	0.50
MEAD	0.39	0.38	0.36
OTS	0.47	0.54	0.49

4. Conclusions

In this work we presented a subjective belief model for ranking sentences according to their importance from a single document. This evidence based model uses interaction and word occurrence among sentences. We presented the effect of a reduced word set for the evidence based model on sentence extraction. One of our hypotheses is supported by the results which show that a reduced filtered (or purified) data set can increases the signal to noise ratio, and can be used for extraction of significant sentences for summarization which are almost as good as human analysis. Another observation from this experiment is that the summaries generated by the top word set closely resemble the standard summarizers rather than the word set having every third word; but the word sets with every third word resembles more closely the human annotated results. This suggests that machine summarizers are too focussed on the *important* words, while human summarizers may be focusing elsewhere, which is likely to be the *content* or the *meaning*. The illustrations in the experimental result section also showed that ROUGE score for $ProbExp_{top}$ and $ProbExp_{\frac{1}{3}}$ is consistently higher for all the documents having fewer words, where we have considered one third of total words. But performance started degrading with the increase in the number of words in the documents where we increased the reduction ratio to one sixth and one ninth respectively. But it is interesting to notice that with only few words, it is still possible to rank the sentences meaningfully according to their importance closely matching with human judgements. Overall, it is clear from the experimental results that PE, the evidence based model, though having higher complexity, is effective and consistent in sentence extraction and summarization and outperforms the other standard summarizers with only one third of the words; thus reducing complexity to a greater extent.

The high complexity of belief based model is also the issue that we encountered. Though we have obtained good performance with our algorithm with a reduced word set, we are

working on further improvements. There are some methods to reduce the computational complexity of the state space which is similarly used in fuzzy measures, called the K-additivity method [4]. We also aim to improve the effectiveness (or accuracy) of the belief based model by data filtration (or reduction), some initial results have already been shown here in the form of reduced data sets in our experiments on significant sentence extraction and summarization.

References

[1] O. Boydell and B. Smyth. From social bookmarking to social summarization: an experiment in community-based summary generation. In *Proceedings of the 12th international conference on Intelligent user interfaces*, page 51. ACM, 2007.

[2] H. Dalianis. SweSum-A Text Summarizer for Swedish http://www. dsv. su. se/% 7Ehercules/papers. *Textsumsummary. html*, 2000.

[3] T. D. Gedeon and T. G. Bowden. Heuristic pattern reduction. *International Joint Conference on Neural Networks*, pages 449–453, 1992.

[4] M. Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2):167–189, 1997.

[5] A. Hunter. *Uncertainty in information systems*. Mc-Graw Hill, London, 1996.

[6] A. Jøsang. A Logic for Uncertain Probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–311, 2001.

[7] A. Jøsang. Probabilistic logic under uncertainty. *Proceedings of the thirteenth Australasian symposium on Theory of computing-Volume 65*, pages 101–110, 2007.

[8] Elizabeth DuRoss Liddy. The discourse-level structure of empirical abstracts: an exploratory study. *Inf. Process. Manage.*, 27(1):55–81, 1991.

[9] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[10] C.Y. Lin and E. Hovy. Identifying topics by position. In *Proceedings of the fifth conference on Applied natural language processing*, pages 283–290. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1997.

[11] C.Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, page 78. Association for Computational Linguistics, 2003.

[12] A. Motro and P. Smets. *Uncertainty Management in Information Systems: From Needs to Solutions*. Kluwer Academic Pub, 1997.

[13] N. J. Nilsson. Probabilistic Logic. *Artificial Intelligence*, 28(1):71–87, 1986.

[14] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, et al. MEAD-a platform for multidocument multilingual text summarization. *Proceedings of LREC*, 2004, 2004.

[15] D. Radev, J. Otterbacher, and Z. Zhang. CST Bank: A Corpus for the Study of Cross-document Structural Relationships. In *Proceedings of LREC 2004*, 2004.

[16] Lawrence H. Reeve, Hyoil Han, and Ari D. Brooks. The use of domain-specific concepts in biomedical text summarization. *Inf. Process. Manage.*, 43(6):1765–1776, 2007.

[17] G. Shafer. *A mathematical theory of evidence*. Princeton University Press Princeton, NJ, 1976.

[18] M. Warner. Wanted: A Definition of Intelligence. *Studies in Intelligence*, 46(3):15–22, 2002.

[19] L. A. Zadeh. Reviews of Books. *AI Magazine*, 5(3):81–83.