Journal of Computer Science

Original Research Paper

Text Summarization Using Morphological Filtering of Intuitionistic Fuzzy Hypergraph

[1,2]Dhanya Prabhasadanam Mohanan, [1]Sreekumar Ananda Rao, [1]Jathavedan Madambi and [3]Ramkumar Padinjarepizharath Balakrishna

[1]Department of Computer Applications, CUSAT, Kochi, India
[2]Department of Computer Science, RSET, Kochi, India
[3]Department of Basic Sciences and Humanities, RSET, Kochi, India

Article history
Received: 11-04-2018
Revised: 18-05-2018
Accepted: 22-06-2018

Corresponding Author:
Dhanya Prabhasadanam Mohanan
Department of Computer Applications, CUSAT, Kochi, India
Email: dhanya.rajeshks@gmail.com

Abstract: Text Summarization has been an area of interest for many years. It refers to creating a concise text of a document without any lose of information. Researchers in the area of natural language processing have developed many abstractive and extractive methods for creating summary. Abstractive summaries modifies the sentences and creates a modified concise form, while extractive summaries pick relevant sentences. The extractive method used in this study is a novel one which models the document as an Intuitionistic Fuzzy Hypergraph (IFHG). This IFHG is subjected to morphological filtering in order to create a concise summary. This is the premier work which applies morphological operations on IFHG that is modeled on a text. The method has generated summary which is almost similar to a human generated summary and showed more accuracy when compared with other machine generated summaries.

Keywords: Dilation, Erosion, Filter, Hypergraph, Intuitionistic Fuzzy

Introduction

Overview

Intuitionistic Fuzzy Hypergraphs (IFHG) were introduced in (Parvathi *et al*., 2009), where the authors have also mentioned the (α,β) cut and the dual intuitionistic hypergraph. The same authors (Parvathi *et al*., 2012) have also developed many operations like complement, join, intersection etc on IFHG. Intuitionistic fuzzy sets and its applications in career determination has also been (Ejegwa *et al*., 2014) developed. They have used a normalized Euclidean distance method for finding out the suitable career for students depending upon their marks for various subjects. Rather than merely having a membership and non membership value, a hesitation margin was also introduced for every node in the IFHG. Generalized strong IFHGs were introduced (Samanta and Sumit, 2014), which can be used for partitioning and clustering. The modeling of a document as a hypergraph and its spectral partitioning (Dhanya *et al*., 2017) resulted in text clusters. Our paper shows how to model a document as IFHG and apply morphological filering on it to create a summary report. The organization of the paper is as follows. Section 2 is the related works in the field of text summarization, section 3 describes how a document can be modeled as intuitionistic fuzzy hypergraph, section 4 is an illustration of the application of various morphological operations like dilation, erosion on a document. Section 5 shows the filter operator on text, the design of summary filter with dilation and erosion, implementation of the system. The advantages of the system are given in section 6. The result analysis and comparison done with existing methods are given in section 7, section 8 is the conclusion and finally the references are included.

Related Works

Graph based Methods

The Google brain team has developed and open sourced the tensor flow model (TST, 2016) for text summarization for generating news headlines on an annotated English giga word, where tensor flow is an open source library for numerical computation using data flow graphs. Interesting parts of the document are extracted using some metric (tf-idf) to create summary. There are many other graph based summarization methods, out of

which five methods like HITS, positional power function, page rank methods, undirected graphs and weighted methods are compared (Mihalcea, 2004) and HITS and page rank seems to provide a better performance. Weighted directed graphs (Borhan *et al*., 2014) are also created by taking in to consideration the distortion measure. There, an edge is formed only if the distortion (semantic distance between node) is below a predefined threshold. In multigraph method (Fatima *et al*., 2015), there are more number of edges between two nodes (sentences). The number of edges equals the number of common words in both sentences. Results of this method are being compared against many online summarizers available and they have shown good performance. Lexical centrality (Erkan and Radev, 2004) is being used in LexRank method, where the sentences similar to many other sentences are found to be central to a topic. Given the similarity of each sentence to other sentences, the overall centrality of a sentence is calculated. The system has shown better results when compared to human summaries. On creation of a graph with multiple documents, sentence selection is done with segmented bushy path (Ribaldo *et al*., 2012) and depth first path method. Redundancy removal is being done at the end.

Neural Network based

Neural text summarization (Karthik, 2016) has defined the work as a task which generates an output sequence $y_1,y_2,.....y_m$ for an input sequence $x = x_1,x_2,.....x_n$. The best summary, the one under the scoring function argmax (x, y) is used. A subset of sentences of Document D is created by predicting the label $y_L \in 0,1$, where 0 stands for non inclusion in summary and 1 stands for inclusion in summary. All sentences are labelled by considering model parameters θ (Jianpeng and Mirella, 2016). Seven features of the document are extracted to create a feature vector, fed to the neural network, feature fusion is done (Kaikhah, 2004) and sentences are filtered. The system is tested for news articles and they got good accuracy. A set of eight features are extracted from a document and fed to a neural network with input layer, hidden layer and output neuron. After finding high ranked sentences, rhetorical structure theory (Kulkarni, 2015) is applied to find better summary. In the encoder decoder model (Urvashi, 2016) of text summary, an encoder reads the input sequence and computes the hidden state representation h_x, decoder uses the h_x to generate the target sequence y. Errors are back propagated from the decoder to encoder through h_x and a minimum entropy model is created. A feed forward auto encoder (Mahmood and Len, 2017) is trained to encode the input x in a concept space $c(x)$. In the encoding phase, the dimensionality is reduced to give a number of codes. Here features are learned by auto encoder rather than manually engineering them.

Genetic Algorithm Based

In a genetic algorithm based method (Carlos *et al*., 2004), each sentence of the document is represented by an attribute vector consisting of position, size, average *tf-sf*, similarity to title, similarity to keywords, cohesion w.r.to other sentences, w.r.to centroid, depth of sentence in tree, direction of sentence in a tree which is obtained after applying hierarchical clustering algorithm, indicators of main concepts, presence of anaphors, proper nouns, discourse markers. They have applied a multi objective GA and a single objective GA and they have shown better results for multi objective GA. Document is represented using a DAG (Vahed *et al*., 2008) where every sentence S_i is added to the graph in chronological order. Weights are then assigned using tf-isf, which are further used for calculating the similarity between two sentences. These similarities are used as edge weights. With these similarities yet other features like topic relation factor, cohesion factor, readability factor are formed. A fitness function designed based on the above three factors is used to calculate the fitness of a chromosome which consists of 1s and 0s, where 1 represents the inclusion of the sentence in the summary and 0 represents the non inclusion of the sentence in the summary. The system has demonstrated better results compared to others.

Fuzzy Logic Based

A lot of methods have been developed for text summarization using fuzzy based systems. A number of parameters like sentence position in paragraph, sentence length, similarity to title, similarity of keyword, similarity to text concept, proper noun, sentence cohesion are used in fuzzy systems. The authors (Farshad *et al*., 2008) have compared the score of vector based method and fuzzy method given by five judges and the fuzzy based summary gave a summary which reflects 77% of the concepts as opposed to 66% performance by the vector based method. In another method, the vector features (Rucha and Apte, 2012) created for each sentence in the document include title feature, sentence length, term weight, sentence position, sentence to sentence similarity, numerical data etc. The results compared with word summarizer, copernic summarizer has shown a better result. Almost the same set of features are used by a triangular membership function (Babar and Pallavi, 2015) which fuzzifies each score to three values low, medium and high. A parallel summary using latent semantic analysis is also taken and both are merged to get the final summary. Experimental results have shown an average precision of 89%. A comparison of fuzzy system is done with neural network with features like cue pharses, legal vocabulary, paragraph structure, citation, term weight, named entity, similarity to neighbouring sentences, absolute location etc and a better result is demonstrated by fuzzy based system (Megala *et al*., 2014; Rajesh *et al*., 2014). A fuzzy logic

based inference system computes the score of each sentence from highest to value above a threshold. The results compared with word summarizer shows a better output for fuzzy system (Farshad *et al*., 2010).

General Methods

Sentiment computation (Rucha and Apte, 2012) of sentences are done which is further used for text summarization. Here the total, absolute and average sentiment scores of sentences are calculated to generate a P% summary.

Different sentence selection methods (Babar and Pallavi, 2015) are implemented such as term weighting, similarity measure and coverage upon which, a human learning algorithm is being applied. In a DAG-structured topic hierarchy (Ramakrishna *et al*., 2015) method, submodular optimization is being done. They have tried it on 1 million topics and 3 million correlation links. Many features like transitive cover, truncated transitive cover and several quality notions like specificity, clarity, relevance, coherence etc were considered. Text summarization has been used for sentiment analysis (Rupal and Yashvardhan, 2017) of reviews of different products like iPhones, camera, hard disks. The reviews in four languages namely English, German, French, Spanish are extracted from amazon. in, conducted a language translation, aspect identification, text summarization and finally sentiment analysis. Here the summary is based on sentence to centroid score, cue phrase score, sentence position score, numerical data and tf-idf score. Textual summaries of long videos (Shagan *et al*., 2017) are generated using recurrent networks where key frames are taken from impactful segments and are converted to textual annotations. The sequence of events in the video are summarized to generate a paragraph description.

Modeling Document using Intuitionistic Fuzzy Hypergraph (IFHG)

Preliminaries

Let $[H_{IF}, (\mu_n, \gamma_n), (\mu_e, \gamma_e), H^n, H^e]$ be an intuitionistic fuzzy hypergraph with membership degree μ_n and non membership degree γ_n defined on the set of nodes H^n and membership degree μ_e and non membership degree γ_e defined on a set of hyperedges H^e of H_{IF}. While using the concept of hypergraphs in document modeling, the sentences in the document forms the hyperedges H^e and the words in the document forms the nodes H^n. The same method can be used in the case of an IFHG where it includes membership and non membership degrees for nodes and hyperedges. The membership value μ_n of a node H^n is the term priority p_n of a word. i.e., the membership value of a word depends on the priority of the word. The words which are having less priority will have a high non membership value, so also the node H^n which represents that word will have a less membership value μ_n and high non membership value γ_n. The words which are having high priority will have a high membership value, so also the nodes H^n which represent those words will have a high membership value μ_n and less non membership value γ_n. The membership and non membership values of the words are assigned according to the Table 1 to 3 respectively. All other words in the document other than those given in Table 1 to 3 will have $\mu_n = 0.5$ and $\gamma_n = 0.5$. Those words are medium words whose presence won't affect the result of morphological operations which are defined on sub IFHG X_{IF} of H_{IF}.

Assigning Membership Degrees

The membership degree $\mu(n_i)$ of some node H^n is the sum of (normalized term frequency, membership value (as given in Table 1,2)) of the word. For such words, non membership degree is <= 1-$\mu(n_i)$. The non membership degree $\gamma(n_i)$ of some of the node H^n is the sum of (normalized term frequency, non membership value of the node (as given in Table 3)). Here the normalized term frequency is the count of the word in the document/number of words in the document. For such words, the membership degree is <= 1-$\gamma(n_i)$. The membership degree of a hyperedge can be written as follows:

$$\mu\left(e_j\right) = \vee_{\forall i,j}\left\{\mu\left(n_i\right) / n_i \in e_j \cap n_i \in P_j\right\} \quad (1)$$

Table 1: Priority set - Words in various domains with high membership values

Domain Words	Sports membership	Domain Words	Health membership
Board	0.6	Disease/illness	0.8
Indian	0.6	Problem	0.7
Failure	0.8	Severe	0.7
Success	0.8	Result	0.7
Score	0.7	Medicine	0.6
Team	0.7	Medical	0.6
Amount	0.6	Medicine	0.8
Player	0.6	Treatment	0.7
Cricket	0.7	Harmful	0.7
Football	0.8	Reason	0.8
Reception	0.8	Severe	0.7
Domain Words	Travel Membership	Domain Words	Politics Membership
Bus	0.8	Failure	0.8
Metro	0.8	Success	0.8
Distance	0.7	Election	0.7
Kilometer	0.7	Chief minister	0.7
Hotel	0.6	Minister	0.7
Road	0.6	Prime minister	0.8
Rail	0.6	Panchayat	0.6
Plane/flight	0.6	Muncipality	0.6
Train	0.8	Corporation	0.6
History	0.7	Result	0.7
Nature	0.8	State/country	0.7

As per Equation 1, The membership degree $\mu(e_j)$ of the hyperedge H^e is the supremum of the membership degrees of all the nodes H^n in it, provided all H^n in it belong to the priority set P_j. The non membership degree $\gamma(e_j)$ of such a hyperedge H^e is <= 1-$\mu(e_j)$. The non membership degree $\gamma(e_j)$ of a hyperedge H^e can be written as follows:

$$\gamma(e_j) = \vee_{\forall i,j} \left\{ \left\{ \gamma(n_i) / n_i \in e_j \right\} \cap \left\{ \exists n_i / n_i \in nP_j \right\} \right\} \quad (2)$$

It is the supremum of the non membership degrees of all the nodes H^n in it, provided atleast one H^n belongs to the non priority set nP_j. The membership degree of such edges will be <= 1-$\gamma(e_j)$. Let us illustrate this IFHG modeling with a small sample text. The text under consideration as in Fig. 1 is a preprocessed one from which the stop words are removed and which is subjected to lemmatization.

This sample text consists of seven sentences. The membership value and the non membership value of these words are calculated from Table 1 to 3. This membership/non membership value along with the normalized term frequency give the membership and non membership degree. For all words other than those in Table 1 to 3, the membership and non membership values are 0.5. Here we consider that the sum of the membership degree and non membership degree of the node (word) is <= 1 (Parvathi *et al*., 2009). i.e., $\mu(n_i)$ + $\gamma(n_i)$ <= 1. So also the sum of the membership degree and non membership degree of the hyperedge (sentence) is <= 1. i.e., $\mu(e_i)$ + $\gamma(e_i)$ <= 1 (Parvathi *et al*., 2009). The IFHG for the above sample text can be drawn as in Fig. 2.

Table 2: Priority set - Words with high membership values

Domain Words	Automobile membership	Domain words	Gadgets membership
New	0.8	Model	0.8
Engine	0.8	Price	0.8
Company	0.8	Market	0.7
Market	0.7	Memory	0.7
Speed	0.7	Speed	0.7
Metro	0.6	Storage	0.7

Table 3: Non priority set -Words with high non membership values

Domain words	Sports non membership	Domain words	non Health membership
Medicine	0.8	Surgery	0.8
Drugs	0.8	Delivery	0.7
Police	0.7	Cancer	0.7
Custody	0.7	Death	0.7
Arrest	0.7	Failure	0.7
Domain words	Travel non membership	Domain words	non Politics membership
Disaster	0.8	Strike	0.8
Accident	0.8	Police	0.7
Death	0.8	Expel	0.7
Deep	0.7	Arrest	0.7
Expensive	0.6	Court	0.7
Luxurious	0.6	Strike	0.8
Expense	0.6	Harthal	0.8
Domain words	Automobile non membership	Domain words	non Gadgets membership
Bike	0.6	Expensive	0.8
Lorry	0.7	Expense	0.8
Bus	0.7	Old	0.8
Minibus	0.7	Tablet	0.7
Railer	0.8	Ipod	0.7
Expensive	0.8	Earphone	0.7
Luxurious	0.8	Outdated	0.8
Old	0.8	Cheap	0.8

"He is an Indian cricket board player. The board has seen his arrest for using drugs. Still the success was with Indian team. The team scored an amount of 20,00,000. The cricket player was arrested on 25/10/17.Police has stopped the reception . Well, the next match is in the city....."

- $indian - n_1,$ $cricket - n_2,$ $board - n_9,$ $player - n_8.$
- $board - n_9,$ $drugs - n_{11},$ $arrest - n_{15}.$
- $indian - n_1,$ $team - n_6,$ $success - n_4.$
- $team - n_6,$ $score - n_5,$ $amount - n_7.$
- $cricket - n_2,$ $player - n_8,$ $arrest - n_{15}.$
- $receipt - n_{16},$ $police - n_{13},$ $stop - n_{17}$.
- $next - n_{18},$ $match - n_{19},$ $city - n_{20}$.

Fig. 1: The sample text to be modeled as intuitionistic fuzzy hypergraph

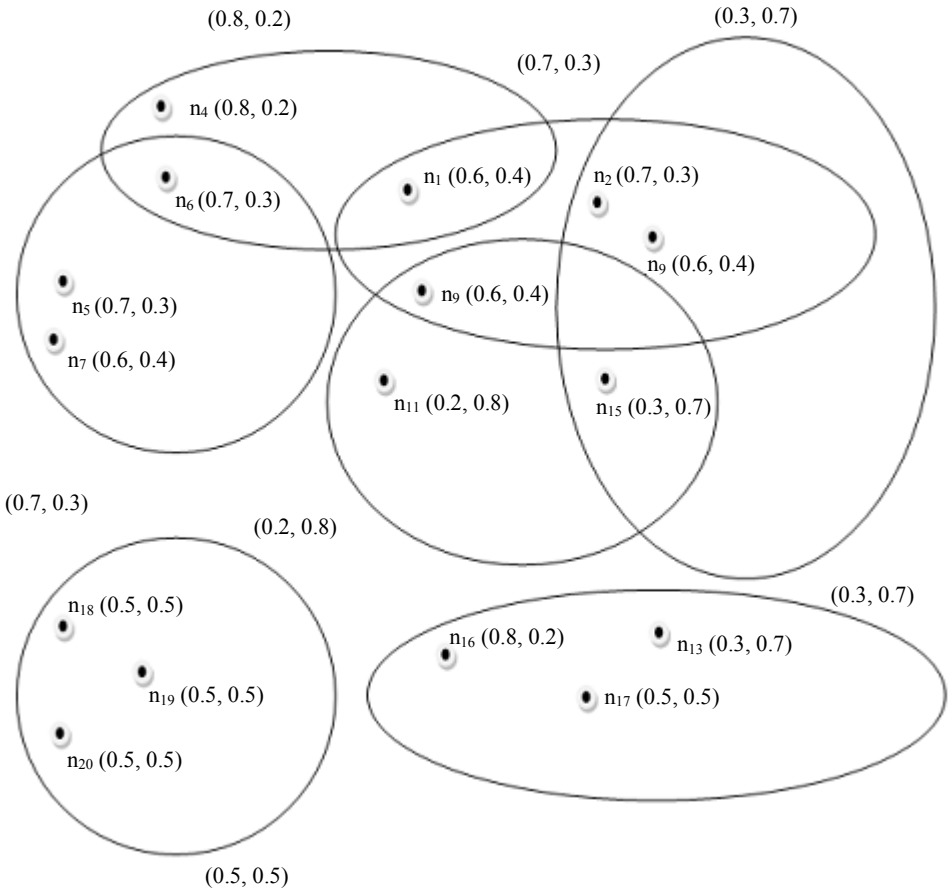

Fig. 2: Text modeled as hypergraph

In Fig. 2, we can see sentences modeled as hyperedges and words modeled as nodes. Nodes are having both membership degree $\mu(n_i)$ and non membership degree $\gamma(n_i)$. The hyperedges are also having both membership degree $\mu(e_i)$ and non membership degree $\gamma(e_i)$. Since there are seven sentences in the sample text in Fig. 1, there are seven hyperedges in Fig. 2. The hyperedge having the nodes n_1, n_2, n_8 and n_9 is an edge with only priority words so that it is having good membership degree. Due to the presence of nodes n_{11} and n_{15} which are having high non membership degree, the corresponding hyperedge is having less membership degree and high non membership degree. I.e., the presence of a single word with high non membership degree $\gamma(n_i)$ influences the non membership degree of the hyperedge.

Morphological Operations on Intuitionistic Fuzzy Hypergraph

Let $\left[X_{IF},\left(\mu'_n,\gamma'_n\right),\left(\mu'_e,\gamma'_e\right),X^n,X^e\right]$ be the sub IFHG obtained by applying the (α,β) cut on H_{IF}, where α corresponds to the membership degree and β corresponds to the non membership degree of nodes/edges. i.e., $H_{\alpha,\beta} = X_{IF}$. The (α,β) cut of H_{IF} can be written as the following:

$$H_{\alpha-\beta}=\left[X_{IF},\left(\mu'_n,\gamma'_n\right),\left(\mu'_e,\gamma'_e\right),X^n,X^e\right]$$
$$=\left\{\begin{array}{l}\left(\mu'^{\alpha}_n,\gamma'^{\beta}_n\right),\left(\mu'^{\alpha}_e,\gamma'^{\beta}_e\right)/\ \mu'^{\alpha}_n=\left\{\mu\left(n_i\right)/\ \mu\left(n_i\right)>\alpha\right\}\\ \cap\gamma'^{\beta}_n=\left\{\gamma\left(n_i\right)/\ \gamma\left(n_i\right)<\beta\right\}\cap\mu'^{\alpha}_e=\left\{\mu\left(e_i\right)/\ \mu\left(e_i\right)>\alpha\right\}\\ \cap\gamma'^{\beta}_e=\left\{\gamma\left(e_i\right)/\ \gamma\left(e_i\right)<\beta\right\}\end{array}\right\}$$

where $\mu(e_i)$ is defined by Equation 1 and $\gamma(e_i)$ is defined by Equation 2.

Here $X_{IF} \subset H_{IF}$, such that X_{IF} consists of nodes with membership degree > 0.5 . The hyperedges in X_{IF} has at least one node with membership degree > 0.5 and it should not contain any node with non membership degree > 0.5. i.e., the membership degree can be greater than 0.5, but the non membership degree should be less than 0.5. Now X_{IF} is a collection of priority sentences and priority words as given in Fig. 3.

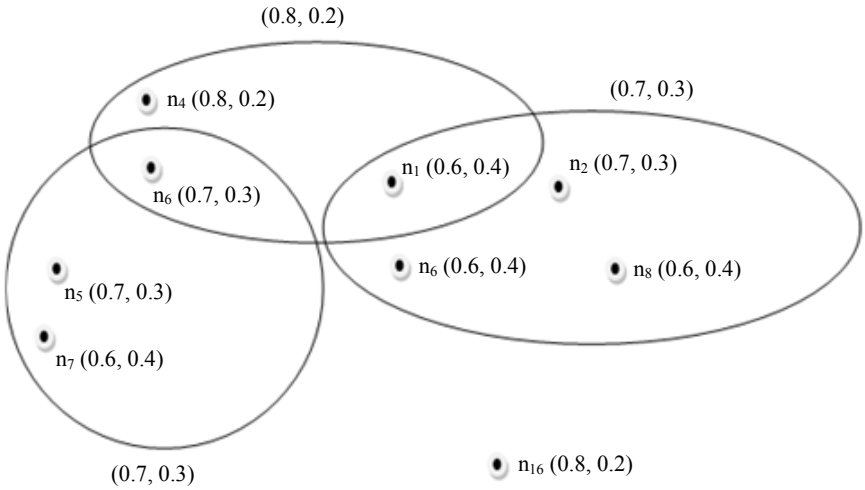

Fig. 3: IFHG X_{IF} obtained after (α,β) cut on H_{IF}

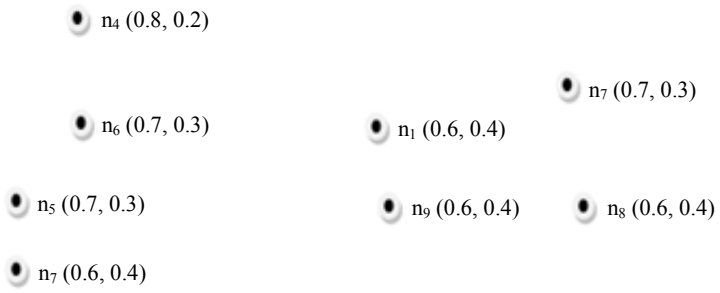

Fig. 4: IFHG obtained after dilation on X_{IF}

Now let us apply morphological operations (Bino *et al*., 2017; Dhanya *et al*., 2018a; 2018b) on this X_{IF}. Let X^n be the node set in X_{IF} and X^e be the edge set in X_{IF}.

$\delta^n(X^e)$-Dilation with Respect to nodes

This morphological operation is defined as:

$$\delta^n\left(X^e\right)=\left\{n_i \;/\; n_i \in X^e\right\} \tag{3}$$

Take all edges in X_{IF}. This will result in X^e. Take all nodes X^n in X^e. Here we are selecting all hyperedges from H_{IF}, which have atleast one node with membership degree >0.5 and which does not contain any node with non membership degree >0.5. Once we select such edges, we select the nodes in it with membership degree >0.5. This will ultimately give $\delta^n(X^e)$. This retreives a collection of priority words within priority sentences as shown in Fig. 4.

$\delta^e(X^n)$ -Dilation with Respect to Hyperedge

This dilation can be written as:

$$\delta^e\left(X^n\right)=\left\{e_i \;/\; e_i \in H^e \cap \left\{\exists n_i \in e_i \;/\; n_i \in X^n\right\}\right\} \tag{4}$$

Take all nodes X^n. Find from H_{IF} all the hyperedges which include X^n. Here we select from X_{IF} all nodes with membership degree >0.5. Find from H_{IF} all hyperedges which contain those nodes. This will give all hyperedges which contain atleast one node with membership degree >0.5. These hyperedges may or may not contain nodes with non membership degree >0.5. This dilation selects all text which has atleast one priority word as shown in Fig. 5.

$\delta(X_n)$ - Node Dilation

This dilation can be written as:

$$\delta\left(X_n\right)=\left\{e_i \;/\; e_i \in H^e \cap \left\{\exists n_i \in e_i \;/\; n_i \in X^e\right\}\right\} \tag{5}$$

Take all hyperedges X^e. Take all nodes in X^e. Find all the hyperedges with respect to H_{IF} which contain these nodes. This dilation gives all sentences in H_{IF}

which overlap with the priority sentences. This is shown in Fig. 6.

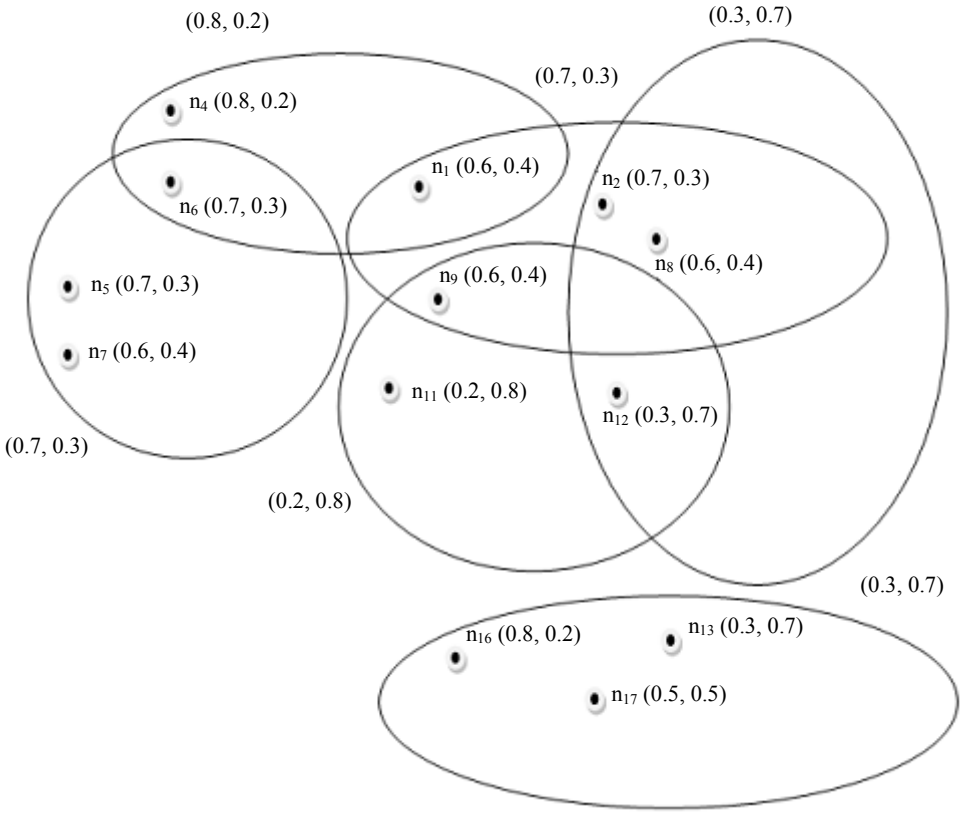

Fig. 5: Dilation w.r.to hyperedge

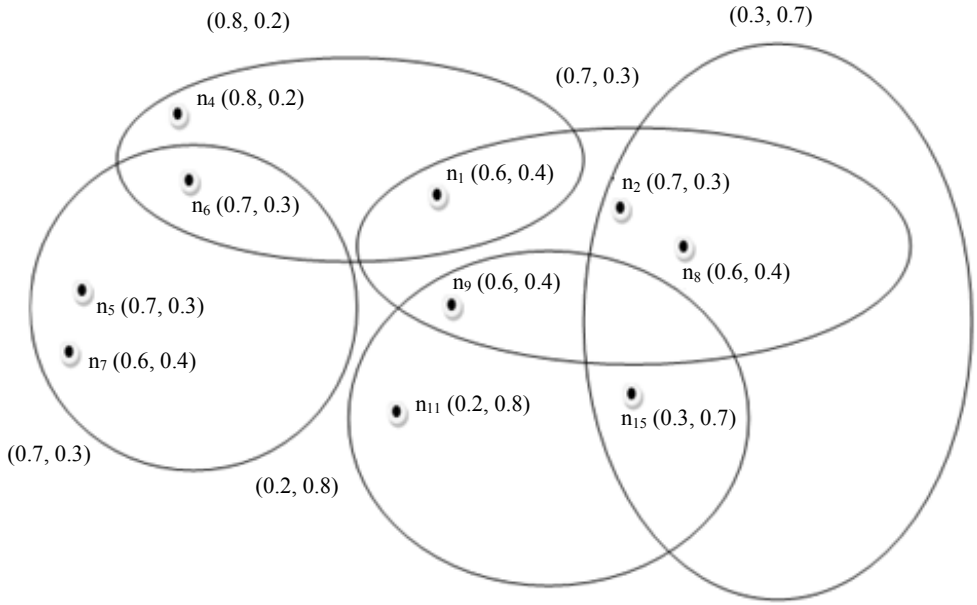

Fig. 6: Dilation

Δ(X^e)- Dilation

This dilation can be written as:

$$\Delta\left(X_n\right)=\left\{e_i \;/\; e_i \in H^e \cap \left\{\exists n_i \in n_i \;/\; n_i \in \left\{X^e \cap H^e\right\}\right\}\right\} \qquad (6)$$

Find all hyperedges X^e. Find all nodes in X^e. Let it be X^{n1}. Find all hyperedges H^e and the nodes in it. Let it be H^{n1}. For all $X^{n1} \cap H^{n1} \neq$ empty, find the hyperedges from H_{IF}. This will retreive all sentences which has atleast one priority word in priority sentences of X_{IF}. The same is represented in Fig. 7.

ε^e(X^n)- Erosion w.r.to Hyperedge

So far we have seen dilation operations of X_{IF}. Now let us see how different types of erosion can be defined on X_{IF}. The erosion $\varepsilon^e(X^n)$ can be defined as the following:

$$\varepsilon^e\left(X^n\right)=\left\{e_i \;/\; e_i \in H^e \cap \left\{\forall n_i \;/\; \left\{n_i \in e_i \cap n_i \in X^n\right\}\right\}\right\} \qquad (7)$$

Take all nodes X^n in X_{IF}. Take all hyperedges in H_{IF} which consists of these nodes only. This erosion as seen in Fig .8 strictly retreives priority sentences.

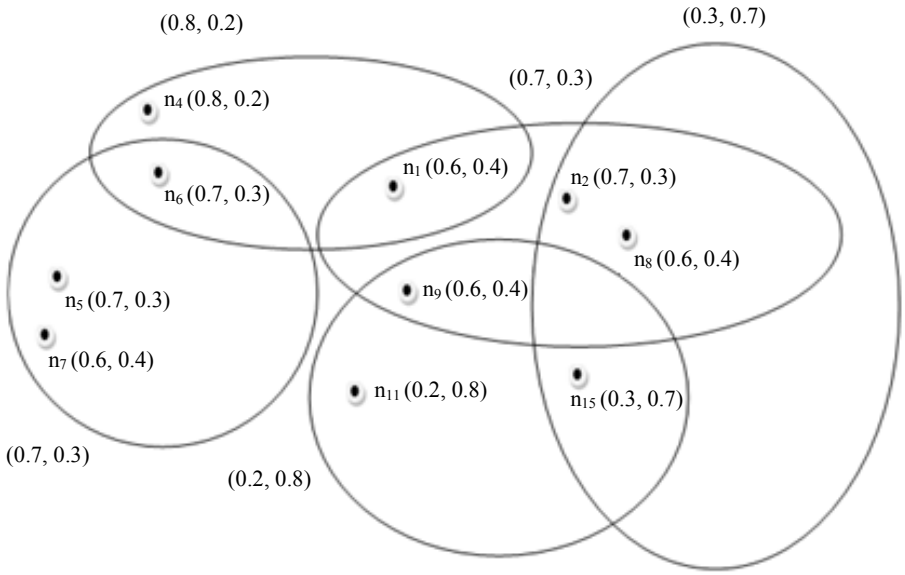

Fig. 7: Dilation

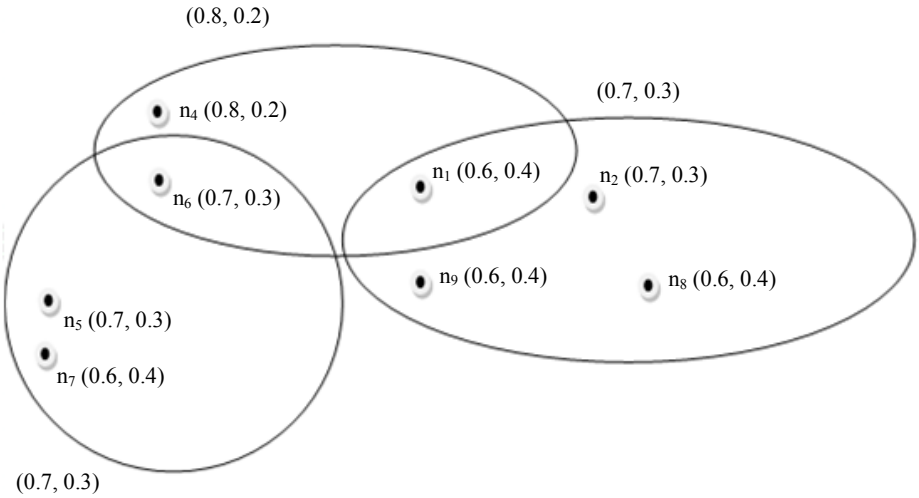

Fig. 8: Erosion with respect to nodes

$\varepsilon^n(X^e)$- Erosion w.r.to node

The erosion $\varepsilon^n(X^e)$ can be written as follows:

$$\varepsilon^n\left(X^e\right)=\left\{n_i \;/\left\{n_i \notin\left\{X^e \cap X^{e'}\right\} / X^{e'}=H_{IF}-X^e\right\}\right\} \tag{8}$$

Take all hyperedges X^e. Take its complement edges $X^{e'}$ in H_{IF}. Take all nodes X^n which are not in $X^e \cap X^{e'}$. This will retrieve all priority sentences which do not overlap with any non priority sentences. Now take the priority words in it as shown in Fig. 9.

$\varepsilon(X^e)$ - Hyperedge Erosion

The erosion $\varepsilon(X^e)$ is defined as the following:

$$\varepsilon\left(X^e\right)=\left\{e_i \;/\; e_i \in H^e \cap\left\{n_i \in e_i \cap n_i \in \varepsilon^n\left(X^e\right)\right\}\right\} \tag{9}$$

Take all nodes in $\varepsilon^n(X^e)$. Take all edges from X_{IF} which fully contains these nodes. This will retreive all priority sentences which do not overlap with the non priority sentences. This is illustrated in Fig. 10.

$[\delta,\Delta](X_{IF})$ - Dilation

This dilation can be written as the following:

$$[\delta,\Delta]\left(X_{IF}\right)=\left\{\begin{matrix}\left(e_i,n_i\right)/\left\{e_i \in\left\{\Delta\left(X^e\right)\cap\delta^e\left(X^n\right)\right\}\right\}\\ \cap\left\{n_i \in e_i \notin\left\{\Delta\left(X^e\right)\cap\delta^e\left(X^n\right)\right\}\right\}\end{matrix}\right\} \tag{10}$$

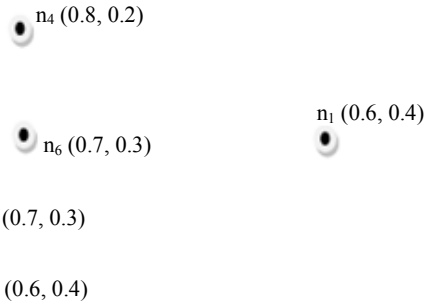

Fig. 9: Erosion with respect to nodes

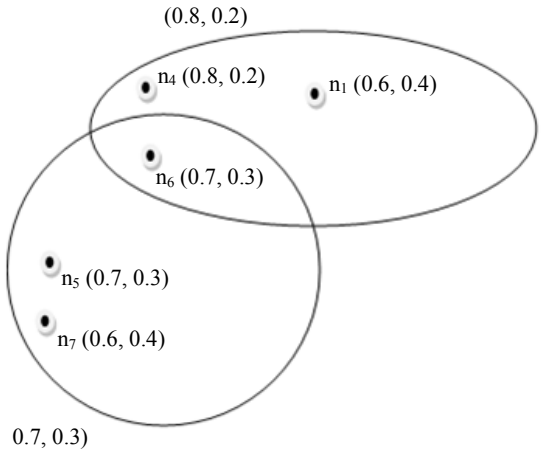

Fig. 10: Erosion

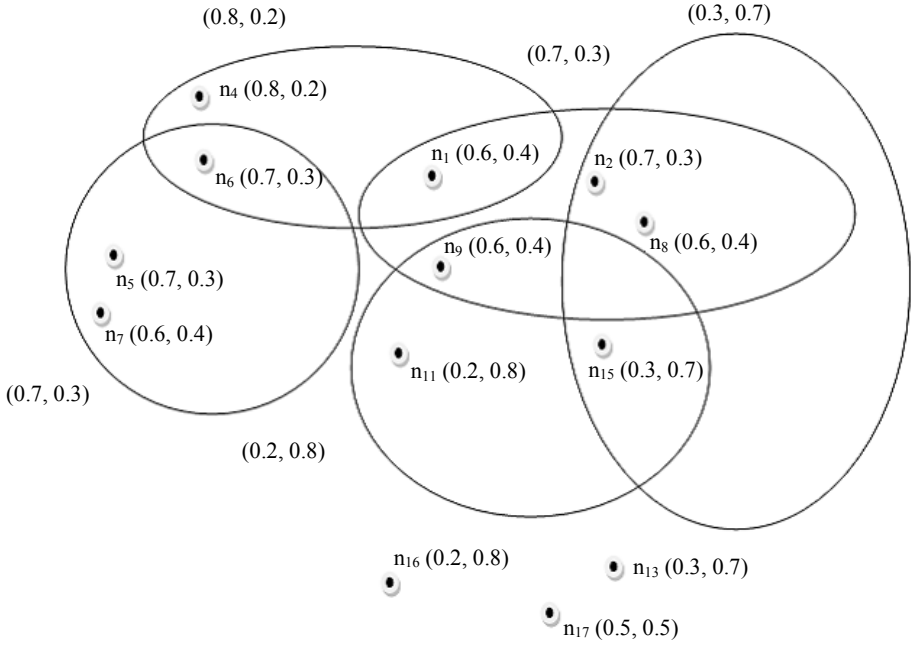

Fig. 11: Dilation

As seen in Fig. 11, this is obtained by joining $\Delta(X^e)$ and $\delta^e(X^n)$. Take all edges which are common in $\delta^e(X^n)$ and $\Delta(X^e)$. Include all such hyperedges and its nodes as output. For other edges in $\delta^e(X^n)$, include only nodes in it. This will retreive all sentences which overlaps with the priority sentences and the words in it. It also retreives all words in sentences which has both priority and non priority words and which do not overlap with others.

Implementation

The implementation of the summarization as shown in Fig. 12 and algorithm 1, is done with the help of a filter system developed in python for input English news taken from online news sites. The English news related to various topics are being subjected to stop word removal and stemming. The preprocessed text is then represented as a weighted hypergraph (Dhanya *et al*., 2017). The weighted hypergraph is subjected to spectral partitioning. Spectral partitions lead to text clusters. The summary filter is then applied to each cluster formed. The sentences which do not fall under any of the clusters are treated as outliers and are removed. A Malayalam summarization system is also developed using the same method, where a Malayalam stemmer (Dhanya *et al*., 2018c) is used to stem the words.

Filter Design

Filter is an operator which is idempotent and increasing defined on domain D. Let X_{IF} be the sub IFHG defined in section 4, then If $f(f(X_{IF})) = f(X_{IF})$, then f is idempotent. If X and Y are sub IFHGs then if $f(X) \subset f(Y)$, then f is increasing. F is a filter if both of these are satisfied. Let ε be the erosion operator and δ be the dilation operator. Then let $\varepsilon \circ \delta$ be an operator and if $\varepsilon \circ \delta\ (\varepsilon \circ \delta(X)) = \varepsilon \circ \delta\ (X)$ then $\varepsilon \circ \delta$ is a filter. That is, here filter consists of a erosion which is composed of dilation or we can say that we have dilation followed by erosion. Such a filter can be used for text summarization. Text summarization basically can be considered as a filter which removes all unwanted sentences from a text. We can also call summarization as a filter operator which selects only the needed sentences from the given text.

Table 4: Star words irrespective of the domain of the text

Words	Membership	Words	Membership
Famous	0.9	Excel	0.9
Fame	0.9	Excellent	0.9
Well known	0.9	Attract	0.9
Famed	0.9	Attractive	0.9
Popular	0.9	Pleasing	0.9
Important	0.9	Pretty	0.9
Prominent	0.9	Alluring	0.9
Main	0.9	Good	0.9
Chief	0.9	Handsome	0.9
Major	0.9	Significant	0.9
Key	0.9	Powerful	0.9
Formost	0.9	Urgent	0.9
Supreme	0.9	Influential	0.9
Overriding	0.9	Momentous	0.9
Essential	0.9	Indispensable	0.9

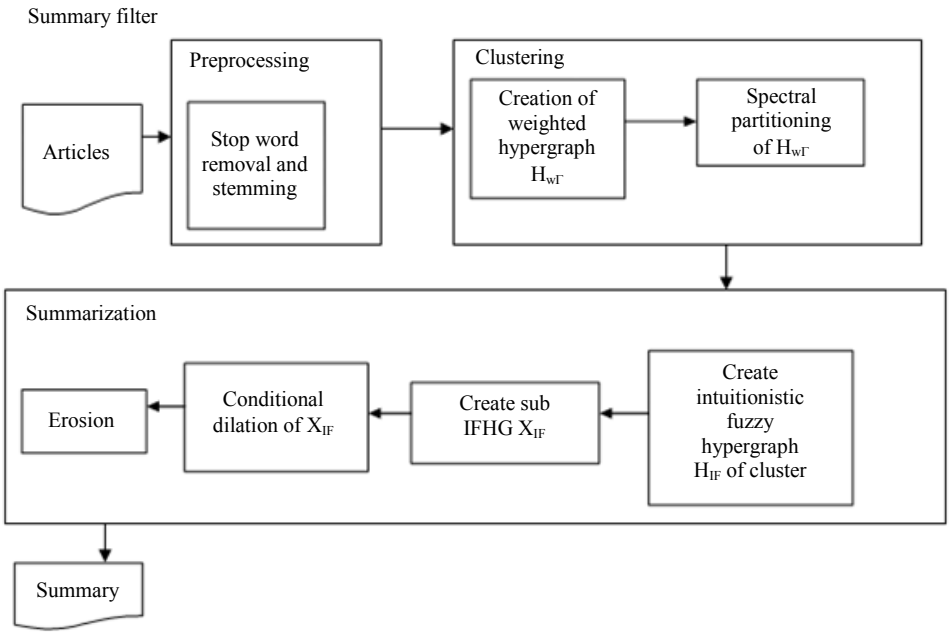

Fig. 12: Architecture of Summarization system

Summary Filter

Text summarization can be done with the help of this filter operator which is applied to the intuitionistic fuzzy hypergraph created from the text under consideration. This filter is designed as a combination of two morphological operators namely dilation and erosion. Here dilation is designed as a conditional one as explained in section 5.2.1 and erosion explained in section 5.2.2 is designed as the one which performs complement operation. For implementing this conditional dilation, let us assume that our text consists of certain star words, whose occurance in sentences are valid even if they co-occur with non priority words. So for this summary filter, let us assume that our text consists of words which are of high prority, words which are of low priority, words with neutral priority and star words. Let us redefine the intuitionistic fuzzy hypergraph as [H_{IF}, (μ_n, γ_n), (μ_e, γ_e), H^n, H^{*n}, H^e, H^{*e}], where H^{*n} is the star node and H^{*e} is the edge which has the star node H^{*n}. These star words are domain independent. Some of the star words are given in Table 4. Sentences which contain star words are definitely included in the summary text. To illustrate this, let us add one more sentence to our sample text as the following. ”The arrest of the famous player.....”. Now this will result in new hyperedge with the following nodes. famous- n_{21} player- n_8 arrest-n_{15}.

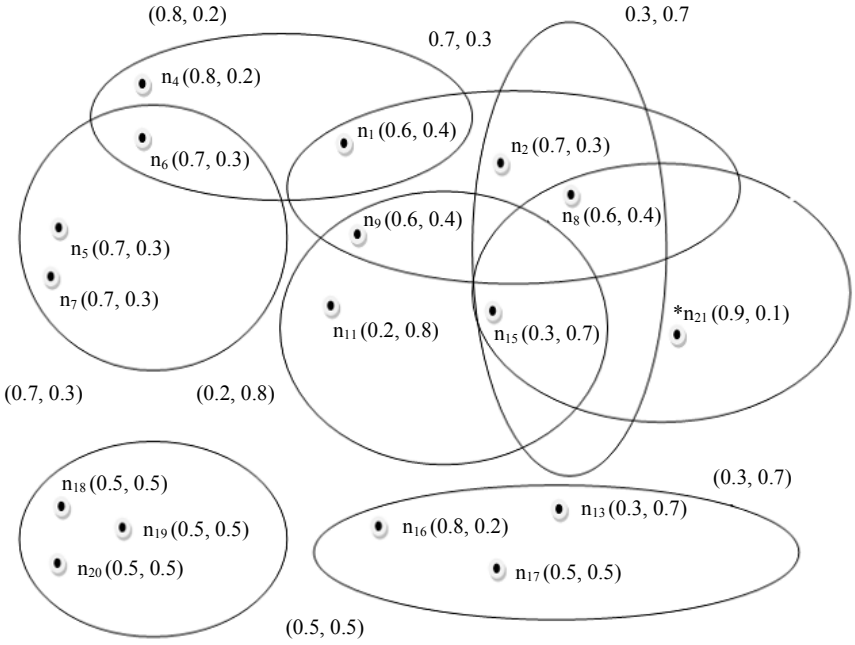

Fig. 13: Modified Intuitionistic fuzzy hypergraph H_{IF}

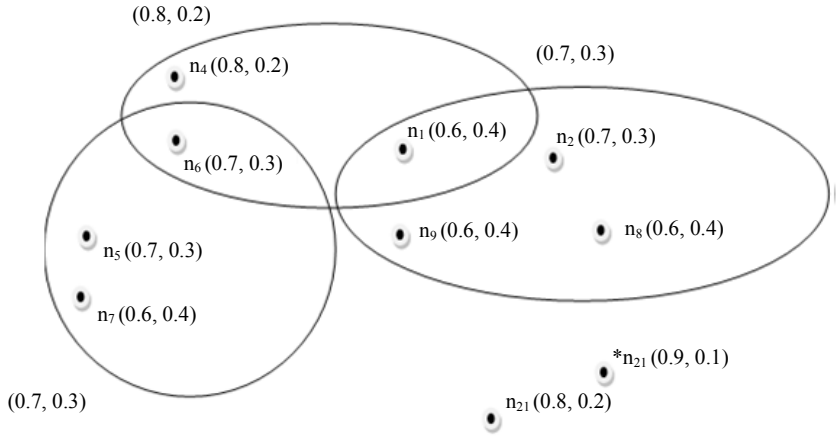

Fig. 14: Modified X_{IF}

The modified intuitionistic fuzzy hypergraph after the addition of the above sentence is given in Fig. 13. The sub IFHG X_{IF} is also getting modified since it will have the star nodes also in it. The modified X_{IF} can be shown as in Fig. 14.

Conditional dilation - $\delta^c(X_{IF})$

This conditional dilation is applied such that while dilating the sub IFHG X_{IF} we consider the condition specified by c, where c is designed such that it selects all hyperedges in H which consists of star nodes given in Table 4:

$$\delta^c\left(X_{IF}\right)=\left\{e_i \;/\; e_i \in H^{*e}\right\} \tag{11}$$

This conditional dilation will retrieve all edges from the intuitionistic fuzzy hypergraph, such that it consists of all edges H^{*e}, which consists of star nodes H^{*n} as given in Fig .15. Even though the non membership degree of the edge H^{*e} is 0.7, it is retreived in the dilation operation which is applied, since it contains the star node H^{*n}.

Erosion - $\varepsilon(H^{*e}, X^e)$

This erosion will retrieve all edges ε' from H_{IF} which are not in H^{*e}. Also take all edges ε'' from H_{IF} which are not in X^e. The intersection of the two will result in the retrieval of non priority edges. Now the complement of this will yield the priority edges from the hypergraph H_{IF}. This erosion will eliminate all duplicate edges from H^{*e} and X^e and retrieve us the most important sentences which itself is the required feature of a summary. This erosion can be written as the following:

$$\varepsilon\left(H^{*e}, X^e\right)=\left\{e_i \;/\; e_i \in\left[H_{IF}-\left[H^{*e'} \cap X^{e'}\right]\right]\right\} \tag{12}$$

where, $H^{*e'}$ is the complement of H^{*e} and $X^{e'}$ is the complement of X^e. The intutionistic fuzzy sub hypergraph retrieved after filter can be shown as in the Fig. 16.

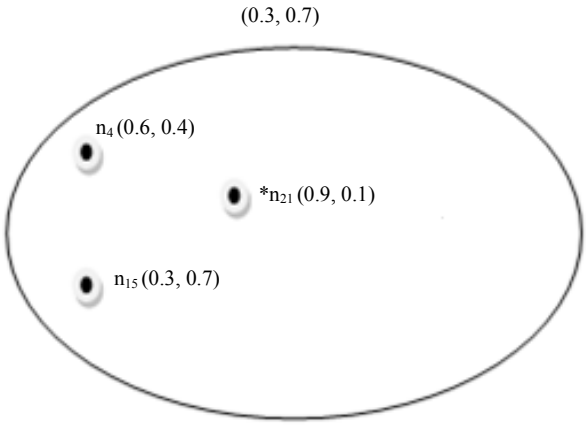

Fig. 15: Conditional dilation on XIF

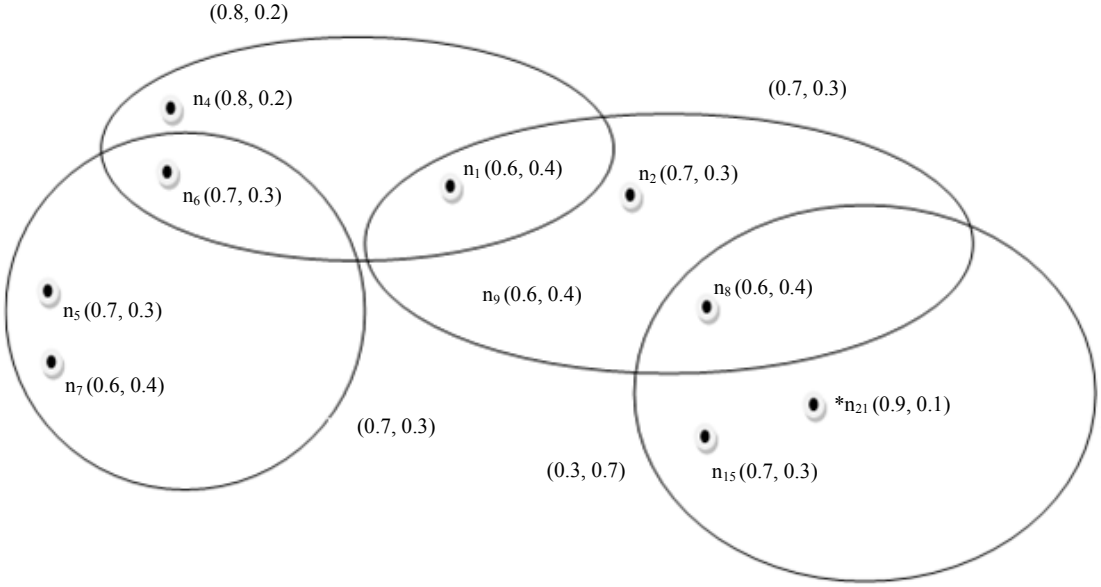

Fig. 16: Summary filter

Algorithm 1 Summarization of text

1: Collect news related to various topics from online sites
2: Preprocess the sentences by subjecting to stop word removal and stemming
3: Create weighted hypergraph $H_{w\tau}$ of the text τ
4: Cluster the text τ using spectral partitioning (Dhanya *et al*., 2017) of hypergraph $H_{w\tau}$
5: for each cluster C_i do
6: Assign $\mu(n_j)$ and $\gamma(n_j)$ for all words C_i
7: Assign $\mu(e_j)$ and $\gamma(e_j)$ for all sentences in C_i
8: Create intuitionistic fuzzy hypergraph H_{IF} with nodes H^n having $(\mu(n_j), \gamma(n_j))$ and hyperedges H^e having $(\mu(e_j), \gamma(e_j))$
9: Create subgraph X_{IF} of H_{IF} with hyperedges X^e having $\mu(e_j) > 0.5$ and nodes X^n having $\mu(n_j) > 0.5$
10: Apply conditional dilation $H^{*e} = \delta^e(X_{IF})$
11: Apply erosion $\varepsilon(H^{*e}, X^e)$ to form the summary
12: end for

Advantages Over Existing Systems

The summarization system which is designed here as a filter applied on IFHG has many advantages over existing summarization methods developed so far. They can be listed as the following.

Variety of Summary Filters

As we all know, a filter is basically a composition of dilation and erosion or erosion and dilation. The proposed new method helps in the creation of series of different types of filters by combining the morphological operators like dilation and erosion discussed in section 4. Using these different types of filters, different types of summaries can be generated. Some of the filter designs other than the one discussed in section 5 are shown below:

- Filter 1 - $\delta(\varepsilon^n(X^e))$ This filter is a composition of erosion $\varepsilon^n(X^e)$ and dilation δ. The erosion will retrieve all nodes in $X^e \cap X^{e'}$. Now the dilation operation will retrieve all hyperedges H^e which contains the nodes retrieved by the erosion operator. This summary filter will retrieve all sentences from the text with atleast one priority word. But this summary will consider star words only if they are part of priority edges in X. Well, this summary is not that short.
- Filter 2 - $\varepsilon(\delta^n(X^e))$ This is a composition of dilation $\delta^n(X^e)$ and erosion ε. The dilation operator retrieves the collection of priority nodes within priority edges. The erosion operator will retrieve all hyperedges H^e in H which consists of only the nodes returned by the dilation operator. This summary retrieves only pure priority sentences that have no non priority words in it. This is a very short summary.
- Filter 3 - $\varepsilon(\delta^e(X^n))$ This is a composition of dilation $\delta^e(X^n)$ and erosion e. The dilation defined by $\delta^e(X^n)$ takes all nodes in X and retrieves all edges from H which consists of these nodes. The erosion will take the double complement of $\delta^e(X^n)$. This is also a very short summary and it will be almost similar to the summary generated in section 5.2. More number of filters can be designed by combining the morphological operators defined in sections 4.1 to 4.8 resulting in the generation of different types of summaries.

Customized Summary

The summary generated by the filter is a customized one as it requires the priority of the user to be submitted before the summary being generated. Thus the summary generated is not a blind one as it takes in to consideration the preferences of the reader. The reader can give as input the priority and non priority words and the summary will be generated accordingly. So the summary report will definitely be a one which satisfies the reader.

Result Analysis

The system is tested on google cloud platform with 8 cores, 30 GB memory. A comparison of the proposed system with the existing online text summarization systems like tools4noobs, summarization. net, splitbrain.org/services is done for various data set. The data set consists of English news taken from online news sites. The news belongs to various domains like travel, politics, health, sports, gadgets etc.The same is uploaded in Mendeley repository. First of all, the news is subjected to clustering and then to summary generation using IFHG method. The summaries generated by each of the above system is compared with human summaries created. About 50 human summarizers are asked to create summaries for each of the data set. The maximum repeating sentences amoung all the 50 summaries are output to create the final human summary with which the existing systems and the IFHG method are compared. The Rouge-L, Rouge-2 and Rouge-1 scores are calculated and summarized in Table 5 to 7. In the following tables 'P' stands for the Precision, 'R' stands for recall and 'F' stands for F-measure.The proposed work has shown an average precision of 0.88, average recall of 0.84 and average F-measure of 0.86. The similarity of the output of the proposed system and the three online systems are compared with the human summaries as shown in Fig. 17. For all the datasets, the system has generated summaries which has more than 90% similarity with human summaries.

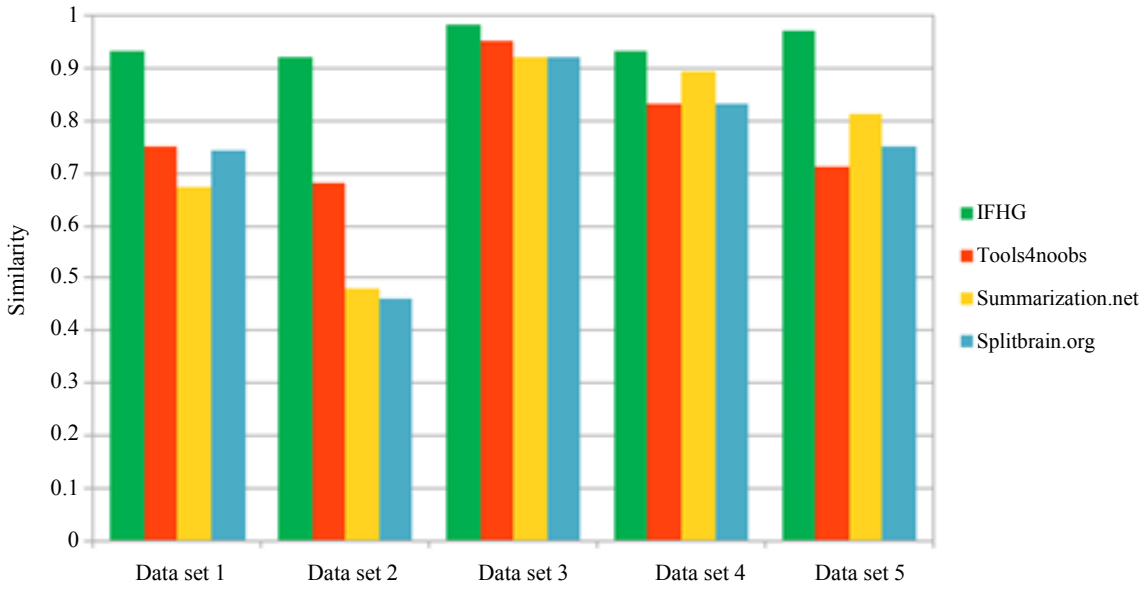

Fig. 17: Similarity with human summary

Table 5: Rouge-L score

Data set size (words)	IFHG P	IFHG R	IFHG F	Tools4noobs P	Tools4noobs R	Tools4noobs F
600	0.89	0.88	0.88	0.46	0.39	0.42
1071	0.81	0.78	0.79	0.33	0.23	0.25
2774	0.95	0.95	0.95	0.25	0.23	0.24
5044	0.67	0.69	0.68	0.19	0.16	0.17
6436	0.97	0.72	0.79	0.39	0.19	0.21
Data set size (words)	Summarization.net P	Summarization.net R	Summarization.net F	Splitbrain.org P	Splitbrain.org R	Splitbrain.org F
600	0.27	0.31	0.29	0.49	0.51	0.50
1071	0.17	0.22	0.19	0.21	0.26	0.23
2774	0.22	0.35	0.24	0.36	0.48	0.39
5044	0.33	0.27	0.29	0.19	0.19	0.19
6436	0.29	0.27	0.28	0.29	0.32	0.31

Table 6: Rouge-2 score

Data set size (words)	IFHG P	IFHG R	IFHG F	Tools4noobs P	Tools4noobs R	Tools4noobs F
600	0.87	0.88	0.88	0.51	0.45	0.48
1071	0.79	0.76	0.77	0.41	0.29	0.34
2774	0.97	0.97	0.97	0.60	0.55	0.58
5044	0.71	0.72	0.72	0.23	0.19	0.21
6436	0.95	0.69	0.79	0.31	0.17	0.23
Data set size (words)	Summarization.net P	Summarization.net R	Summarization.net F	Splitbrain.org P	Splitbrain.org R	Splitbrain.org F
600	0.31	0.36	0.33	0.51	0.56	0.53
1071	0.07	0.09	0.08	0.17	0.20	0.18
2774	0.39	0.60	0.47	0.46	0.62	0.53
5044	0.46	0.41	0.43	0.20	0.21	0.21
6436	0.28	0.29	0.29	0.24	0.27	0.25

Table 7: Rouge-1 score

Data set size (words)	IFHG			Tools4noobs		
	P	R	F	P	R	F
600	0.88	0.92	0.91	0.58	0.55	0.57
1071	0.81	0.79	0.79	0.49	0.37	0.42
2774	0.97	0.97	0.97	0.69	0.66	0.67
5044	0.79	0.78	0.78	0.37	0.34	0.36
6436	0.97	0.74	0.84	0.49	0.34	0.39
Data set size(words)	Summarization.net			Splitbrain.org		
	P	R	F	P	R	F
600	0.40	0.46	0.43	0.55	0.65	0.59
1071	0.19	0.25	0.22	0.29	0.35	0.32
2774	0.52	0.69	0.59	0.54	0.73	0.61
5044	0.55	0.5	0.52	0.34	0.38	0.36
6436	0.42	0.48	0.45	0.40	0.49	0.44

Conclusion

The system developed here has successfully modeled text using IFHG, where words become nodes and sentences become hyperedges. Membership degrees and non membership degrees are assigned for nodes. Based on that, membership degrees and non membership degrees of hyperedges are calculated. Various morphological operations are defined on IFHG. Summary of the text is created by applying a filter operator on IFHG. The system has given a better performance when compared to other existing systems. The summary filter has shown more similarity with human summaries generated. The system combines multiple text and treat it as a single one. The system can also be extended with multiple documents, where important words can be modeled as nodes and documents as hyperedges. In our system, there is only a single sub IFHG with which morphological operations are defined. Other enhancements like creating more than one sub IFHG and morphogical operations with intersection/union of those are also possible. All these are left as future enhancements of the proposed work.

Acknowledgement

We are extremely thankful to Dr. Kannan Balakrishnan, Dept of Computer Applications, Cochin University of Science and Technology, Kochi, India for his timely advice , help and support for this work.

Authors Contributions

Dhanya Prabhasadanam Mohanan: Project proposal, data modeling, design of the proposed work, python programming, acquisition of data and manuscript editing and final approval.

Sreekumar Ananda Rao: Contribution as research guide, technical corrections and article review and final approval.

Jathavedan Madambi: Contribution as research guide, technical corrections, manuscript editing and article review and final approval.

Ramkumar Padinjarepizharath Balakrishna: Mathematical modeling of work, suggesting survey paper, project design and critical review of the work and final approval.

Conflicts of Interest and Ethics

The authors declare that they have no conflicts of interest in publishing this manuscript. This manuscript is not published elsewhere and there are no ethical issues.

References

Babar, S.A. and D.P. Pallavi, 2015. Improving performance of text summarization. Proc. Comput. Sci., 46: 354-363. DOI: 10.1016/j.procs.2015.02.031

Bino, V.S., A. Unnikrishnan, B. Kannan and P.B. Ramkumar, 2017. Morphological filtering on hypergraphs. Discrete Applied Math., 216: 307-320. DOI: 10.1016/j.dam.2015.02.008

Borhan, S., E. Marzieh, K. Fazel and H. Sattar, 2014. Multi-document summarization using graph-based iterative ranking algorithms and information theoretical distortion measures. Proceedings of the 27th International Conference of the Florida Artificial Intelligence Research Society, (IRS' 14), pp: 214-218.

Carlos, N.S., L.P. Gisele, F. Alex and C.A.A. Kaestner, 2004. Automatic text summarization with genetic algorithm-based attribute selection. Proceedings of the 9th Ibero-American Conference on Advances in Artificial Intelligence, Nov. 22-26, Springer, Puebla, México, pp: 305-314.

Dhanya, P.M., A. Sreekumar, M. Jathavedan and P.B. Ramkumar, 2017. Document modeling and clustering using hypergraph. Int. J. Applied Eng. Res., 12: 2127-2135.

Dhanya, P.M., A. Sreekumar, M. Jathavedan and P.B. Ramkumar, 2018a. Algebra of morphological dilation on intuitionistic fuzzy hypergraphs. Int. J. Scientific Res. Sci. Eng. Technol., 4: 300-308.

Dhanya, P.M., A. Sreekumar, M. Jathavedan and P.B. Ramkumar, 2018b. On constructing morphological erosion of intuitionistic fuzzy hypergraphs. J. Anal. DOI: 10.1007/s41478-018-0096-3

Dhanya, P.M., A. Sreekumar and M. Jathavedan, 2018c. Vriksh: A tree based malayalam lemmatizer using suffix replacement dictionary. Int. J. Emerg. Technol. Eng. Res., 6: 31-42.

Ejegwa, P.A., A.J. Akubo and O.M. Joshua, 2014. Intuitionistic fuzzy set and its application in career determination via normalized Euclidean distance method. Eur. Scientific J., 10: 529-536.

Erkan, G. and D.R. Radev, 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artificial Intell. Res., 22: 457-479.

Farshad, K., K. Hamid, E. Esfandiar and D. Mohsen, 2010. Extraction-based text summarization using fuzzy analysis. Iran. J. Fuzzy Syst., 7: 15-32. DOI: 10.22111/ijfs.2010.185

Farshad, K., K. Hamid, E. Esfandiar, K.D. Pooya and T. Asghar, 2008. Optimizing text summarization based on fuzzy logic. Proceedings of the 7th IEEE/ACIS International Conference on Computer and Information Science, May 14-16, IEEE Xplore Press, Portland, OR, USA, pp: 347-352. DOI: 10.1109/ICIS.2008.46

Fatima, Q., A. Saif and C. Martin, 2015. New graph-based text summarization method. Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Aug. 24-26, IEEE Xplore Press, Victoria, BC, Canada, pp: 396-401. DOI: 10.1109/PACRIM.2015.7334869

Jianpeng, C. and L. Mirella, 2016. Neural summarization by extracting sentences and words. coRR, arXiv preprint arXiv:1603.07252.

Kaikhah, K., 2004. Automatic Text summarization with neural networks. Proceedings of the 2nd International IEEE Conference on Intelligent Systems, Jun. 22-24, IEEE Xplore Press, Varna, Bulgaria, pp: 40-44. DOI: 10.1109/IS.2004.1344634

Karthik, B.M., 2016. Text summarization using deep learning and ridge regression. coRR, arXiv preprint arXiv:1612.08333.

Kulkarni, A.R., 2015. Text summarization using neural networks and rhetorical structure theory. Int. J. Adv. Res. Comput. Commun. Eng., 4: 49-52. DOI: 10.17148/IJARCCE.2015.4612

Mahmood, Y.A. and H. Len, 2017. Text summarization using unsupervised deep learning. Expert Syst. Applic., 68: 93-105. DOI: 10.1016/j.eswa.2016.10.017

Megala, S.S., A. Kavitha and A. Marimuthu, 2014. Enriching text summarization using fuzzy logic. Int. J. Comput. Sci. Inf. Technol., 5: 863-867.

Mihalcea, R., 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Proceedings of the on Interactive Poster and Demonstration Sessions, Jul. 21-26, Association for Computational Linguistics, Barcelona, Spain, pp: 1-4. DOI: 10.3115/1219044.1219064

Parvathi, R., S. Thilagavathi and M.G. Karunambigai, 2009. Intuitionistic fuzzy hypergraphs. Cybernet. Inform. Technol., 9: 46-53.

Parvathi, R., S. Thilagavathi and M.G. Karunambigai, 2012. Operations on intuitionistic fuzzy hypergraphs. Int. J. Comput. Applic., 51; 46-54. DOI: 10.5120/8041-1357

Rajesh, D.S., H.R. Suraj, S.J. Savita and R.S. Smita, 2014. Enforcing text summarization using fuzzy logic. Int. J. Comput. Sci. Inf. Technol., 5: 8276-8279.

Ramakrishna, B., I. Ramakrishna, K. Rishabh, G. Ramakrishnan and A.B. Jeff, 2015. Summarization of multi-document topic hierarchies using submodular mixtures. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Jul. 26-31, Beijing, China, pp: 553-563.

Ribaldo, R., A.T. Ademar, H.M.R. Lucia and A.S.P. Thiago, 2012. Graph-based methods for multi-document summarization: Exploring relationship maps, complex networks and discourse information. Proceedings of the 10th international conference on Computational Processing of the Portuguese Language, Apr. 17-20, Springer, Coimbra, Portugal, pp: 260-271. DOI: 10.1007/978-3-642-28885-2_30

Rucha, S.D. and S.S. Apte, 2012. Improvement of text summarization using fuzzy logic based method. IOSR J. Comput. Eng., 5: 5-10. DOI: 10.9790/0661-0560510

Rucha, S.D. and S.S. Apte, 2012. Improvement of text summarization using fuzzy logic based method. IOSR J. Comput. Eng., 5: 5-10.

Rupal, B. and S. Yashvardhan, 2017. MSATS: Multilingual sentiment analysis via text summarization. Proceedings of the 7th International Conference on Cloud Computing, Data Science and Engineering-Confluence, Jan. 12-13, IEEE Xplore Press, Noida, India, pp: 71-76. DOI: 10.1109/CONFLUENCE.2017.7943126

Samanta, T.K. and M. Sumit, 2014. Generalized strong intuitionstic fuzzy hypergraph. Math. Moravica, 18: 55-65. DOI: 10.5937/MatMor1401055S

Shagan, S., K. Sourabh, G. Allison, V. Subhashini and P. Emily *et al*., 2017. Semantic text summarization of long videos. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Mar. 24-31, IEEE Xplore Press, Santa Rosa, CA, USA, pp: 989-997. DOI: 10.1109/WACV.2017.115

TST, 2016. Text summarization with tensor flow.

Urvashi, K., 2016. Neural text summarization. CS224d-Stanford University.

Vahed, Q., S.H. Leila and H. Ramin, 2008. Summarizing text with a genetic algorithm-based sentence extraction. Int. J. Knowl. Manage. Stud., 2: 426-444. DOI: 10.1504/IJKMS.2008.019750