Text Summarization Using Rough Sets

Nouman Azam and Afzaal Ahmad

nouman.azam@nu.edu.pk
National University Of Computer and Emerging Sciences, Peshawar

April 7, 2016

Overview

1. Problem Statement and Motivation

2. Contribution

3. Demostrative Example

4. Experimental Results

Text Summarization

Text Summarization is the process of reducing text documents while retaining important information.

Applications

- Web Content Mining.
- Financial Reports.
- JistWeb.
- News Stories.
- Online Blogs.

Existing Text Summarization Methods

Broad categories of text summarization methods:

- Abstractive Summarization (Pourvali and Abadeh, 2012)
- Extractive Summarization (Pourvali and Abadeh, 2012)
- Inductive Summarization (Pourvali and Abadeh, 2012)

Each of these categories has further different types and methods developed for them in Literature.

Pourvali and Abadeh, Automated Text Summarization Based on Lexicales Chain and graph Using of WordNet IJCSI 2012.

Problem Statement

Rough Sets has been utilized in several fields.

- Data Mining (Hung Son, 2010)
- Feature Selection (Megala and Kavitha, 2014)
- Conflict Analysis (Yao and Yan, 2007)
- Text Mining (Thanh,Yamada, and Unehara, 2011)

  We aim to explore rough sets for text summarization.

  WHY?

  There is no such investigation in existing literature.

Hung Son,Introduction to rough sets and data mining (2010).
Megala and Kavitha,Feature extraction based legal document summarization (2014)
Yao and Zhao, Conflict analysis based upon Indiscernibility and Discernibility,IEEE Computational Intelligence (2007)
Thanh, Yamada and Unehara,A Similarity Rough Set Model for Document Representation and Clustering (2011)

Contribution

We considered three different types of set relations to generate summaries of text documents based on rough sets.

- Discernibility Relation.
- Indiscernibility Relation.
- Equal To one Relation.

Rough Sets

Rough sets deal with information represented in the form of information table. Information table is formally represented as,

$S = (U, At, \{V_a | a \in At\}, \{I_a | a \in At\})$

- U is a finite non empty set of objects called universe.
- At is finite non empty se of attributes.
- V_a is non empty set of values for $a \in At$.
- $I_a : U \to V_a$ is an information function.

Demostrative Example

- Information table contain words in rows (Objects) and sentences in columns (attributes).
- Value **0** shows absence of a word in respective sentence.
- Value **1** represents presence of a word.

Table: Information Table

	S1	S2	S3	S4	S5
W1	1	0	1	1	0
W2	1	1	0	0	0
W3	1	0	1	0	1

Discernibility Matrix

Discernibility Relation

$$D = \{a \in At | I_a(x) \neq I_a(y), (x, y) \in U\}$$

Table: Discernibility Matrix

$U * U$	W1	W2	W3
W1	-	S2,S3,S4	S4,S5
W2	-	-	S2,S3,S5
W3	-	-	-

Indiscernibility Matrix

Indiscernibility Relation

$$I = \{a \in At | I_a(x) = I_a(y), (x, y) \in U\}$$

Table: Indiscernibility Matrix

$U * U$	W1	W2	W3
W1	S1,S2,S3,S4,S5	S1,S5	S1,S2,S3
W2	-	S1,S2,S3,S4,S5	S1,S4
W3	-	-	S1,S2,S3,S4,S5

Equal To One Matrix

Equal to one Relation

$$O = \{a \in At | I_a(x) = I_a(y) = 1, (x, y) \in U\}$$

Table: Equal to one Matrix

$U * U$	W1	W2	W3
W1	S1,S3,S4	S1	S1,S3
W2	-	S1,S2	S1
W3	-	-	S1,S3,S5

Matrix Simplification Operations

Matrix Absorption(Yao and Zhao, 2009)
The matrix absorption operation is a sequence of all possible element absorption operations on pair of elements whenever the following condition holds:

$$\phi \neq M(x`, y`) \subset M(x, y)$$

After matrix absorption, no element in the matrix is proper subset of another element.

Element Deletion(Yao and Zhao, 2009)
For an attribute $a \in At$, the attribute deletion operation deletes {a} from all the elements if the following condition holds:

$$\forall (M(x, y) \neq \phi)(M(x, y) - \{a\}) \neq \phi$$

Yao and Zhao,Discernibility matrix simplification for constructing attribute reducts,Information Sciences(2009)

Example

Let us suppose a Discernibility Matrix:

$$\left\{\begin{array}{ccccc} \phi & & & & \\ \phi & \phi & & & \\ \underline{\{a,b,f\}} & \underline{\{c,d,f\}} & \underline{\{b,e,f\}} & & \\ \underline{\{a,c\}} & \underline{\{b,d\}} & \underline{\{c,e\}} & \phi & \\ \underline{\{a,d\}} & \underline{\{b,c\}} & \underline{\{d,e\}} & \phi & \phi \end{array}\right\}$$

- We want to find out minimum number of attributes which will preserve a particular relation i.e Discernibility relation.
- For this purpose we will iteratively apply element deletion and absorption operations until we obtain a minimum matrix.

Matrix Absorption

The first iteration of row-wise simplification causes no change because no subset is available.

$$\left\{\begin{array}{ccccc}
\phi & & & & \\
\phi & \phi & & & \\
\{a,b,f\} & \underline{\{c,d,f\}} & \underline{\{b,e,f\}} & & \\
\underline{\{a,c\}} & \underline{\{b,d\}} & \underline{\{c,e\}} & \phi & \\
\underline{\{a,d\}} & \underline{\{b,c\}} & \underline{\{d,e\}} & \phi & \phi
\end{array}\right)$$

Element Deletion

Let $A = \{b, f\}$ and $M(4,1) = \{a\}$. We simplify part B into:

$$\left\{\begin{matrix} \phi & & & & \\ \phi & \phi & & & \\ \{a\} & \underline{\{c,d\}} & \underline{\{e\}} & & \\ \underline{\{a\}} & \underline{\{d\}} & \underline{\{c,e\}} & \phi & \\ \underline{\{a\}} & \underline{\{c\}} & \underline{\{d,e\}} & \phi & \phi \end{matrix}\right\}$$

Second Itteration

The second iteration of row-wise simplification absorbs $M(4,2)$ by B:

$$\left\{\begin{matrix} \phi & & & & \\ \phi & \phi & & & \\ \{a\} & \{d\} & \underline{\{e\}} & & \\ \underline{\{a\}} & \underline{\{d\}} & \underline{\{c,e\}} & \phi & \\ \underline{\{a\}} & \underline{\{c\}} & \underline{\{d,e\}} & \phi & \phi \end{matrix}\right\}$$

Let $A = \phi$ and $M(4,2) = \{d\}$. We simplify part B into:

$$\left\{\begin{matrix} \phi & & & & \\ \phi & \phi & & & \\ \{a\} & \{d\} & \underline{\{e\}} & & \\ \underline{\{a\}} & \underline{\{d\}} & \underline{\{c,e\}} & \phi & \\ \underline{\{a\}} & \underline{\{c\}} & \underline{\{d\}} & \phi & \phi \end{matrix}\right\}$$

After three iterations original matrix is simplified into minimum matrix which produces a reduct {a, c, d, e}.

Reducts

Reducts are the subset of attributes $R \subseteq Attributes$ if R meets following Conditions (Yao and Zhao, 2009)

For Discernibility Relation

- $DIS(R) = DIS(At)$
- For any $a \in R, DIS(R - \{a\}) \neq DIS(At)$

For Indiscernibility Relation

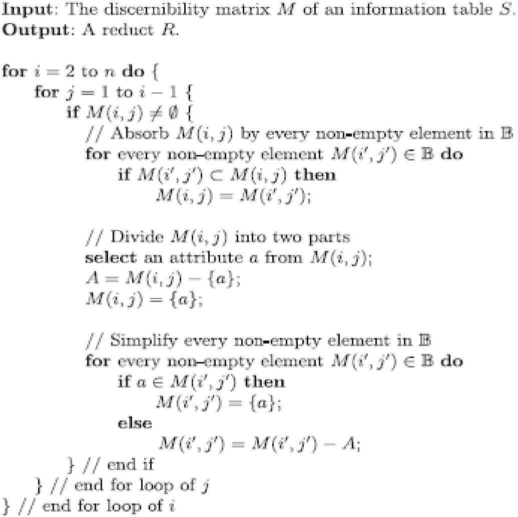
- $IND(R) = IND(At)$
- For any $a \in R, DIS(R - \{a\}) \neq DIS(At)$

For Equal To one Relation

- $O(R) = O(At)$
- For any $a \in R, O(R - \{a\}) \neq O(At)$

Yao and Zhao,Discernibility matrix simplification for constructing attribute reducts, Information Sciences(2009)

Reduct Construction Algorithm



Input: The discernibility matrix M of an information table S.
Output: A reduct R.

```
for $i = 2$ to $n$ do {
    for $j = 1$ to $i - 1$ {
        if $M(i,j) \neq \emptyset$ {
            // Absorb $M(i,j)$ by every non-empty element in $\mathbb{B}$
            for every non-empty element $M(i',j') \in \mathbb{B}$ do
                if $M(i',j') \subset M(i,j)$ then
                    $M(i,j) = M(i',j')$;

            // Divide $M(i,j)$ into two parts
            select an attribute $a$ from $M(i,j)$;
            $A = M(i,j) - \{a\}$;
            $M(i,j) = \{a\}$;

            // Simplify every non-empty element in $\mathbb{B}$
            for every non-empty element $M(i',j') \in \mathbb{B}$ do
                if $a \in M(i',j')$ then
                    $M(i',j') = \{a\}$;
                else
                    $M(i',j') = M(i',j') - A$;
        } // end if
    } // end for loop of $j$
} // end for loop of $i$
```

Figure: A row-wise simplification reduct construction algorithm

Evaluation Measures

- $Summ_{ref}$ refers to the summary of text constructed by other summarization systems i.e Microsoft word Summarizer, Auto summarizer (Pourvali and Abadeh, 2012)
- $Summ_{cand}$ is the summary of our proposed method (Pourvali and Abadeh, 2012)
- F_1 Measure combines both precision,recall and takes their harmonic mean (Pourvali and Abadeh, 2012)

$$Precision = (Summ_{ref} \cap Summ_{cand})/Summ_{cand}$$
$$Recall = (Summ_{ref} \cap Summ_{cand})/Summ_{ref}$$
$$F_1 = 2(Precision)(Recall)/Precision + Recall$$

Pourvali and Abadeh, Automated Text Summarization Base on Lexicales Chain and graph Using of WordNet, IJCSI (2012)

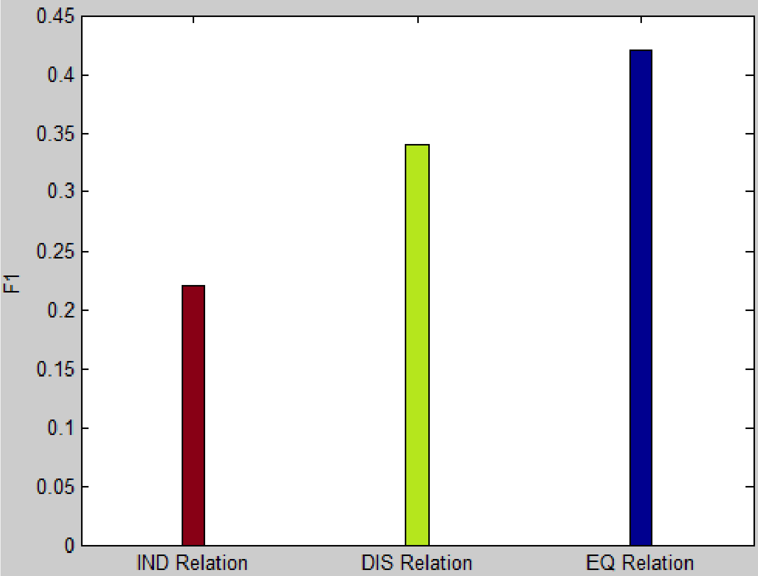
Comparison With Lexical Chains (elhadad)

Table: Results For Comparison With Lexical Chains Method

Our Method	Precision	Recall	F_1
Equal to one Relation	**0.4**	**0.45**	**0.42**
Discernibility Relation	0.31	0.38	0.34
Indiscernibility Relation	0.21	0.27	0.22

http://www.cs.bgu.ac.il/ elhadad/summary-test.html

Bar Graph of F1 Measure



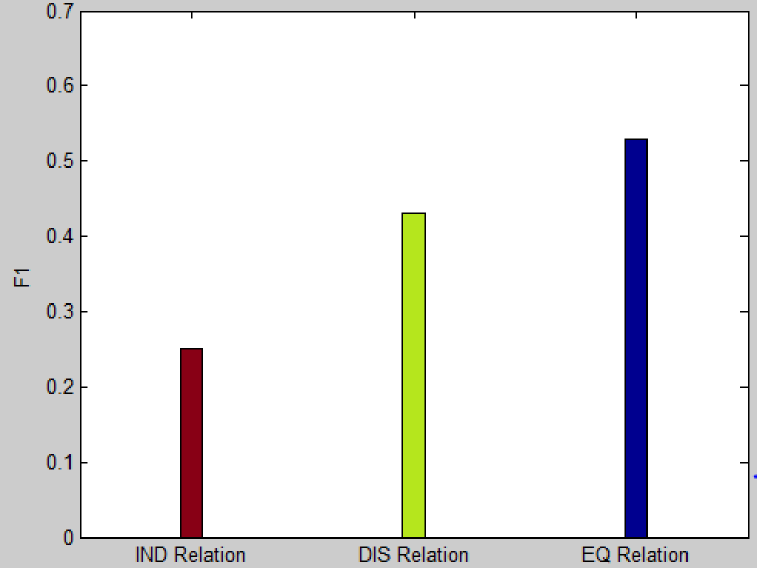
Comparison With Auto Summarization Toolkit (tools4noobs)

Table: Results For Comparison With Auto Text Summarizer Toolkit

Our Method	Precision	Recall	F_1
Equal to one Relation	**0.58**	**0.52**	**0.53**
Discernibility Relation	0.47	0.42	0.43
Indiscernibility Relation	0.29	0.23	0.25

http://www.tools4noobs.com/summarize/

Bar Graph of F1 Measure

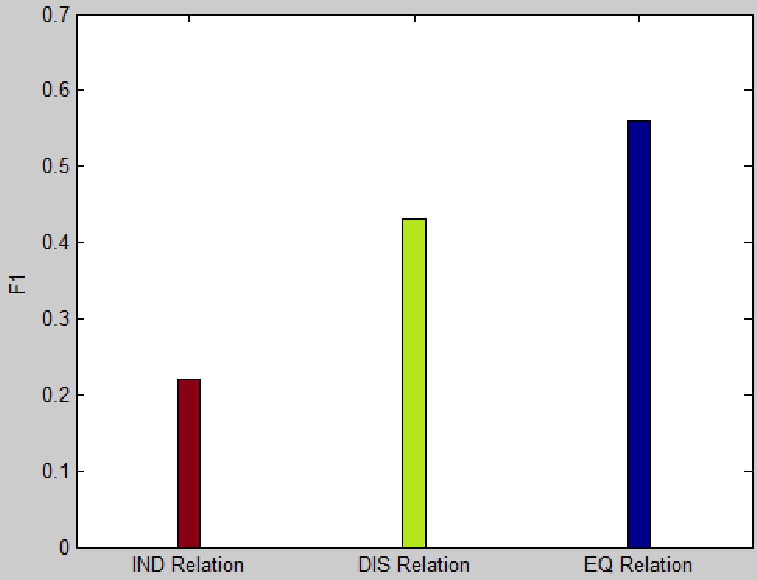


Comparison With Microsoft Summarizer

Table: Results For Comparison With Microsoft Word Summarizer Tool

Our Method	Precision	Recall	F_1
Equal to one Relation	**0.6**	**0.53**	**0.56**
Discernibility Relation	0.46	0.43	0.44
Indiscernibility Relation	0.27	0.22	0.23

Bar Graph of F1 Measure



Conclusion

- Results advocate that Rough Sets can play essential role in text summarization.
- There exist similarities between our summaries and three other systems which make our method reliable.
- Equal to one relation has high similarity ratio when compared with other summarization systems.
- After Equal to one relation, discernibility has high similarity ratio.
- Indiscernibility has the minimum similarity ratio.

Future Work

- This method can be extended for Multi Document Summarization system.
- We have several other types of set relationships on the basis of which we can generate summaries.

Several other Reduct construction algorithms have been developed

- Discernibility Function (Yao and Zhao, 2009)
- Indiscernibility Fuction (Yao and Zhao, 2009)
- Deletion algorithm for Reduct Construction(Yao and Zhao, 2009)
- addition algorithm for reduct calculation (Yao and Zhao, 2009)

Yao and Zhao,Discernibility matrix simplification for constructing attribute reducts, Information Sciences(2009)

Text Summarization Using Rough Sets

Nouman Azam and Afzaal Ahmad

nouman.azam@nu.edu.pk
National University Of Computer and Emerging Sciences, Peshawar

April 7, 2016

additional notes

Sets from imperfect, imprecise, incomplete data may not be precisely defined they have to be approximated these approximations are of three types:

1. Lower approximation and Positive Region

$$LowerApproximation = \{X_i \in U_i[X_i] \subset X\}$$

2. Upper approximation and Negative Region

$$UpperApproximation = \{X_i \in U_i[X_i] \cap X \neq 0\}$$

3. Boundary Region

$$BoundaryRegion = UpperApproximation - LowerApproximation$$

additional notes

To understand these approximations let us explain an example suppose we have an information Table below:

Table: Information Table

[U/A]	a_1	a_2	a_3
$\{X_1, X_3, X_9\}$	2	1	3
$\{X_2, X_7, X_10\}$	3	2	1
$\{X_4\}$	2	2	3
$\{X_5, X_8\}$	1	1	4
$\{X_6\}$	1	1	2

additional notes

Suppose a target set X

$$X = \{x_1, x_3, x_4, x_5, x_9\}$$

Lower approximation for X:

$$LowerApproximation = \{x_1, x_3, x_9\} \cup \{x_4\}$$
$$LowerApproximation = \{x_1, x_3, x_4, x_9\}$$

Upper approximation for X is:

$$upperApproximation = \{x_1, x_3, x_9\} \cup \{x_4\} \cup \{x_5, x_8\}$$
$$UpperApproximation = \{x_1, x_3, x_4, x_5, x_8, x_9\}$$

Boundary Region for X is:

$$Boundary = \{x_1, x_3, x_4, x_5, x_8, x_9\} - \{x_1, x_3, x_4, x_9\}$$
$$Boundary = \{x_5, x_8\}$$

additional notes

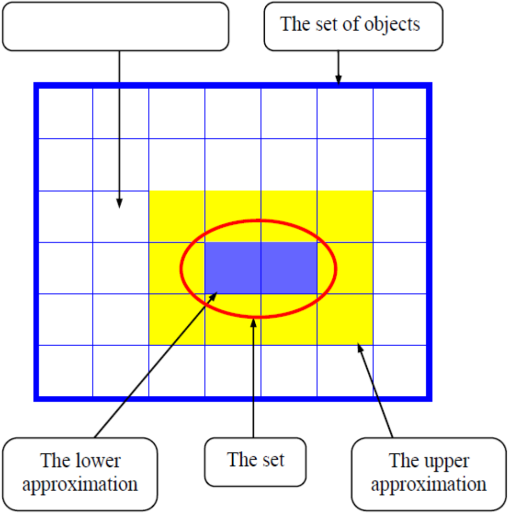


Figure: Rough Sets approximation Granules