TEXT SUMMARIZATION IN MONGOLIAN LANGUAGE

Chuluundorj Begz

University of the Humanities, Ulaanbaatar, Mongolia

ABSTRACT

Textual information in this new era, it is difficult to manually extract the summary of a large data different areas of social communication accumulates the enormous amounts of data. Therefore, it is important to develop methods for searching and absorbing relevant information, selecting important sentences, paragraphs from large texts, to summarize texts by finding topics of the text and frequency based clustering of sentences. In this paper, the author presents some ideas on using mathematical models in presenting the source text into a shorter version with semantics, graph-based approach for text summarization in Mongolian language.

KEYWORDS

graph representation, adjacency matrix, vector space, similarity measurement, quantum cognition, algorithm and encoding

1. INTRODUCTION

1. In human cognition mapping of perceptual signals is analyzed in the framework of gestalt concept of proximity (closeness) and similarity, enclosure and closure, continuity which presents psychological basis of graph based modeling of text perception. In relation to text perception, associative network of human mental lexicon serves as a basis to building graph of text as network of propositions.
   Semantics structure of a text or discourse, coherence and cohesion between propositions, sub-concepts or subtopics present an object of description in terms of Boolean algebra leading to matrix based representation of a text. Models as a BoW or BoC in close connection with set theory and operations of disjunction (union) and conjunction (intersection) serve as a basis for building graph based representation of a text.

2. LITERATURE REVIEW AND RESEARCH

In quantum interpretation computational scalability or semantic scaling is described from the bag of words to the bag of sentences and further (Ilya A.Surov, E.Semenenko, A.Bessmertny, F.Galofaro, Z.Toffano, Quantum semantics of text perception. Scientific reports, 11, 2021, p. 9).

Quantum approach to information retrieval is applied for meaning based processing of textual data. Units of cognition such as ideas, decisions, referred to logs are encoded by distributional neuronal ensembles. Perception of a text has potential to activate it (1) or not (0) in two dimensional vectors.

$$\varphi = C_o \mid O_x \rangle + C_1 \mid 1_x \rangle$$

According to distributional hypothesis, a text is composed of a hierarchy of topics. Topic is composed of a hierarchy of terms and a text can be treated as a bag of words (BoW). In terms of propositional semantics, a text can be described as a bag of concepts (BoC).

In graph based model of a text its semantic links between content components (propositions, subtopics) must be represented by keywords or terms.

Vector space representation of semantics and probabilistic nature of textual events present an effective way to modeling text semantic structure with encoding elementary units of cognition such as ideas,

thoughts and decisions as cogs. These cogs are associated with quiet states of functional group of neurons realizing the cog. (Ilya A.Surov, F.Galofaro, E.Semenenko, Z.Toffano, Quantum semantics of text perception. Scientific reports. 2021.11)

Co-occurrence of words in semantically connected sentences (or paragraphs), its frequency across a text as indicator of text coherence between subtopics present a basis for text reduction or summarization in the form of topic-term matrix using SVD. SVD is used to reduce number of rows (in case if rows represent words, columns-paragraphs of a text), preserving the similarity structure among columns or cohesion between paragraphs of a text (What is LSA, Vimarsh Karbhari, 2020. Acing AI).

Converting graph to matrix form serves as a basis to apply SVD to text analysis.

2. Methods and ideas presented above must be applied to analysis of text (or discourse) in Mongolian language related to the family of agglutinative languages.

Short text example on topic of negative effects of air pollution must be presented in list of following subtopics.

Утаат манан

Утаат манан нь аж үйлдвэржсэн бүс нутгаар нийтлэг бөгөөд энэ нь хотуудын хувьд танил харагдац байсаар байна. Утаат манан ихэвчлэн нүүрс түлснээс үүсдэг.

Орон гэрээ халаах болон хоол хийх зорилгоор нүүрс түлэх нь хорт хийн ялгарлыг бий болгож байна.

Утаат манан нь утаа болон манангийн нэгдэл юм. Утаат манан нь агаарын бохирдлын аюултай төрөл гэж үзэгддэг.

Манан нь хүмүүст ижил төрлийн амьсгалын асуудлыг бий болгодог байхад утаа нь уушгийг хорт хавдар үүсгэгчээр дүүргэж байгаагаараа илүү аюултай юм.

Уушгины хорт хавдар нь нүүрсийг ахуйн хэрэглээний зорилгоор түлэхтэй бас холбоотой юм.

Нүүрс нь дэлхий нийтээр хэрэглэж буй нүүрстөрөгчийн хамгийн том эх сурвалж бөгөөд бидний эрүүл мэндэд зонхилох аюул заналыг учруулж буй гэж үздэг.

Smog

Smog is common in industrial areas, and remains a familiar sight in cities. The smog usually came from burning coal. Burning coal inside the home for the purposes of heating and cooking produces gas emissions.

Smog is combination of fog and smoke. Smog refers to a dangerous type of pollution. While fog can cause similar breathing problems in people, smog is more dangerous as it fills the lungs with cancer causing carcinogens. Lung cancer also is associated with household combustion of coal.

Coal remains the world’s single largest source of carbon pollution, presenting a major threat to our health.

Smog

Sentences

A. Smog is common in industrial areas, and remains a familiar sight in cities.
B. The smog usually came from burning coal.
C. Burning coal inside the home for the purposes of heating and cooking produces gas emissions.
D. Smog is combination of fog and smoke.
E. Smog prefers to a dangerous type of pollution.
F. While fog can cause similar breathing problems in people, smog is more dangerous as it fills the lungs with cancer causing carcinogens.
G. Lung cancer also is associated with household combustion of coal.
H. Coal remains the world’s single largest source of carbon pollution, presenting a major threat to our health.

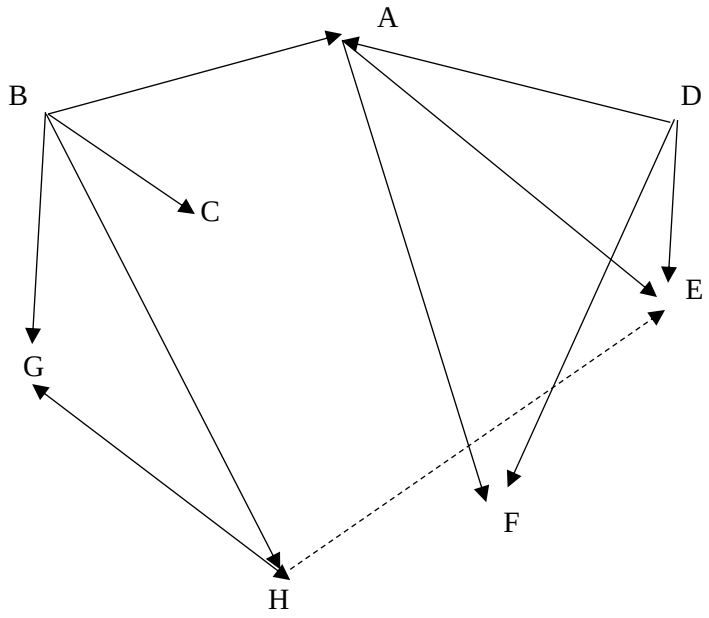
Keywords in Sentences

A – smog

B – coal

C – emission
D – fog, smoke
E – pollution
F – breathing, cancer
G – cancer
H – coal, pollution, health (as a concluding sentence)

Graph of the text:



In graph based representation of a text. Frequency of keywords emphasizes a weight of a vertex, and thus value of a vector. Number of vectors associated with the vertex indicates multiple cohesion between sentences and coherence between paragraphs (subtopics or propositions).

In addition, cohesion between terms (keywords), similarity and degree of coherence of a text must be analyzed not only by using TF-IDF, but with cosine similarity, least square method or correlational analysis. Cosine similarity is used to calculate the similarity between the centroid of the text and the sentence embedding.

In case of multiple effect (direct and non-direct effect, B → and C⇢A, A→ D and B ⇢D) number of edges associated to given vertex must express value of the vertex.

Value of an edge (or vector) must be measured with frequency of terms (words) expressing same meaning or semantic value.

Description of a vertex within a graph and edges sharing common information of two or more vertices is important to develop techniques for modeling text content and matrix based representation of a text.

Keyword associated to a keyword in following sentence or paragraph must be described as a linear continuation of in semantical dimension if a direction of a vector is not changed. Change in vector direction leads to interpretation of a keywords association in both of semantical and pragmatical dimensions.

3. Creating a graph matrix using vector in combination with other techniques measuring similarity between content topics and keywords distribution presents a basis for developing algorithms for text analysis and human verbal cognition.

   Subtopic sentences are compared by taking cosine of the angle between two vectors where dot product between the normalization of two vectors is formed by any two rows.

The rows of the matrix U contains information about each keyword (term) and how it contributes to each factor (i.e. the "factors" are just linear combinations of our elementary term vectors). The columns of the matrix V^t contain information about how each paragraph (sentence) is related to each factor (i.e. the paragraph or text is linear combinations of these factors with weights corresponding to the elements of V^t).

Adjacency list:

B → A → E
B → A → F
B → C
B → G
D → A
D → E
D → F
B → H
G → H
H → E

B	→	A	→	E
B	→	A	→	F
B	→	C		
B	→	G		
D	→	A		
D	→	E		
D	→	F		
B	→	H		
G	→	H		
H	→	E		

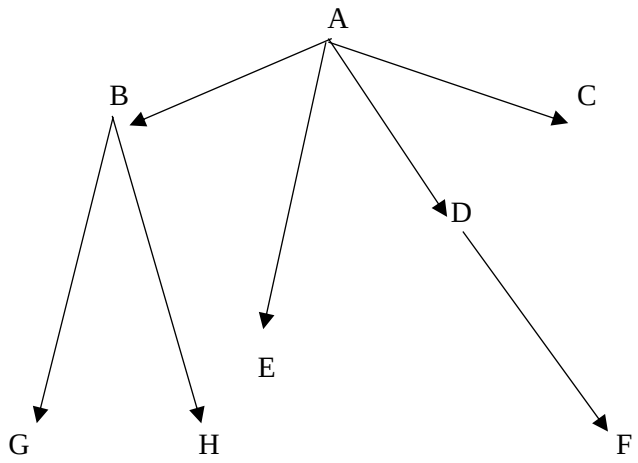
Adjacency matrix of the directed graph

	A	B	C	D	E	F	G	H
A					1	1		
B			1				1	
C		1						
D	1				1	1		
E								
F								
G								!
H					?		!	

Term (keyword) x paragraph (sentence) matrix

		Sentences							
		A	B	C	D	E	F	G	H
Keywords	smog		1		1	1	1		
	coal		1	1				1	1
	emission			1					
	smoke				1				
	fog				1		1		
	pollution					1			1
	breathing						1		
	cancer						1	1	
	health								1
	lung						1	1	

Following simplified version of text content presents a case for modeling a content of above presented text in terms of the graph (and matrix).



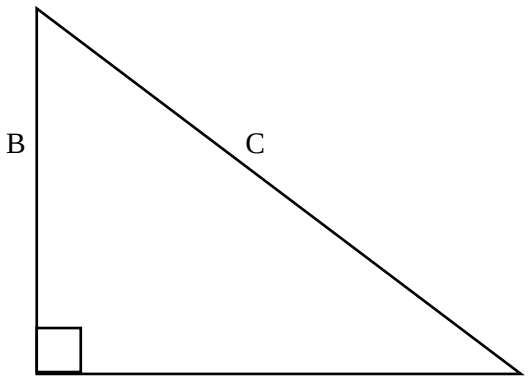
Term (keyword) x paragraph (sentence) matrix

	A	B	C	D	E	F	G	H
smog	o							
coal		o						
emission			o					
smoke				o				
pollution					o			
fog						o		
cancer							o	
health								o

LSA, PCA, cosine distance measure also must be applied to the estimation of semantic connection between words, sentences and paragraphs scaling of bag of words, bag of sentences with application of PCA, LCA and the techniques for dimensionally reduction tend to find correlations between keywords, sentences, paragraphs or documents computing eigenvectors of matrix.

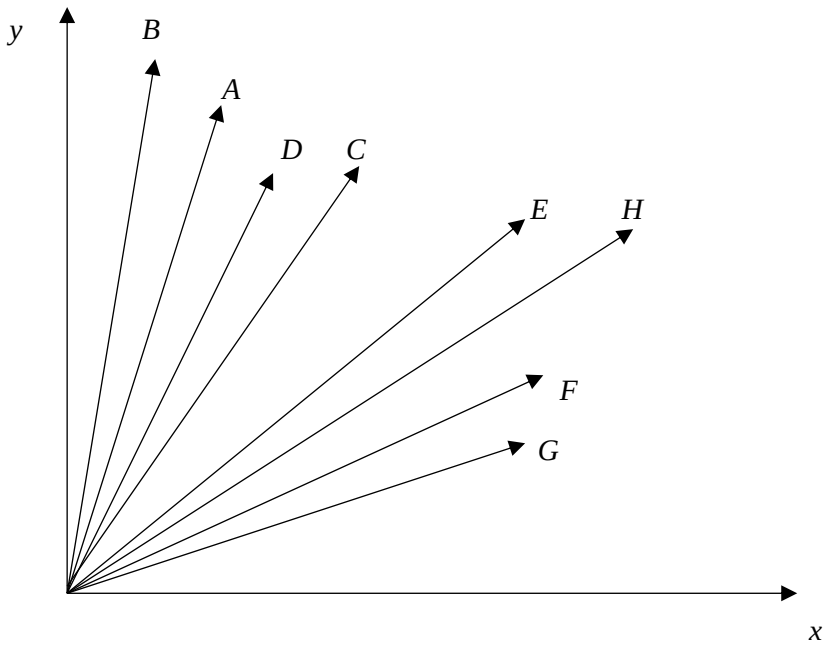
Idea of Pythagorean theorem is applied to calculate approximation by LCA or SVD algorithms to find singular vector or singular value. With combination of Pythagorean theorem SVD goes to stronger approximation of text content or date identifying a relevant subspace of all dimensions.

Pythagorean theorem is applied to minimized the distance between text components.



$$C=\sqrt{a^2}+b^2$$

(Guide to Singular value decomposition. Munkundh Murthy, 2012, p. 7)



Matrix (sentence-keyword, phrase-term etc) care be represented in Euclidean space as vectors. This matrix is step to model relationships among keywords/sentences (phrases) and this to perform SVD.

In SVD the matrix must be represented as a product of three matrices in some cases the matrix "mxn" has "m" rows representing document, paragraph or sentence, "n" columns representing terms or keywords.

$$A_{[mxn]} = U_{[RxR]} \sum_{[RxR]} (V_{[nxR]})^T$$

A – input data matrix mxn (m – document, paragraph, sentence, n – term, keyword)

U – left singular vectors mxR matrix (R - concept)

∑ - singular value – RxR diagonal matrix (strength of each concept)

V – right singular vectors (n – term, keyword, R - concept)

$$\sum = \begin{matrix} \delta_{11} & O_{000} & O \\ O & \delta_{22} & O \\ \vdots & \vdots & \\ O & O_{000} & \delta_{pp} \end{matrix}$$

Analysis of human mental associative mechanism must be used in developing graph or matrix based models for text extraction.

Semantic similarity of keywords or terms related to same node and logico-semantic coherence between keywords, terms are based on different associations in human cognition.

Paradigmatic and syntagmatic relations between words (keywords, terms) evoke different associations in human mental space (smog-smoke, fog, emission; smog-health, cancer).

These differences in association present interest for analysis in terms of covariance and contra variance of vectors on tensor analysis.

Some researchers express an idea about relation of representation of human brain also in the framework of Euclidean distance. This idea leads to space-time (four dimensional) representation of human cognition as a quantum model of cognition. The central tenet of this idea is that brain is a tensorial system and central nervous system acts as a metric tensor which determines the relationships between the contravariant and covariant vectors. Quantum processes and functional geometry (Sisir Roy, 2004).

4. Matrix based representation is used to detect the main topics and the relations between them. Description of coherence between propositions or subtopics of a text using conversion of graph into a matrix creates basis for text encoding. Algorithms based on co-occurrence of terms (words) present effective technique to select important sentences, paragraphs from the text creating high level of shared subtopics (propositions) and concepts based on text coherence.

   The algorithm using TF-IDF to find most relevant words of the text as the centroid of the text sums up the vectors of the words as a part of the centroid. The sentences of the text are scored based on how similar they are to the centroid embedding.

   There are two types of encoding-binary and one hot encoding for text analysis. In one hot encoding every word represented as a vector, list of vectors, an array of vectors. One hot encoding has finite state where each state takes a separate flip flop.

   In binary-encoding all possible states are defined and there is no possibility for hang state.

   In matrix based encoding of a text, one hot encoding converts terms with original categorical value into numeric variables on values 0 or 1.

   Complex-valued text presents a specific object of analysis there cognitive superposition and cognitive coherence are referred to quantum superposition and quantum coherence, entanglement measure of semantic connection (Scientific report, p. 5).

   Quantum entanglement between cognitive subspaces leads to semantic connection between sub-concepts in text recognition.

   According to researchers, at variance with classical Kolmogorovian probability, quantum probability enables coping with the superposition in the subject's thought (D.Aerts, J.Broekaert, L.Gabora, S.Sozzo, Quantum structure and human thought. Behav.Brain.Sci. 36, p. 274).

   Two concepts A and B are combined in human thought to form the conjunction (A and B) or the disjunction, a new quantum effect comes to a superposition in the subject's thought. This phenomenon has some association with metaphorical representation in human thoughts related to guppy effect of context sensitivity (Patrizio E.Tressoldi, Lance Storm, Dean Radin, Extrasensory perception and quantum models of cognition, NeuroQuantology, 2010, Vol 8, Issue 4, p. 585).

   In this case, text perception under uncertainty in the terms of expected utility theory (EUT – John von Neumann, Oskar Morgenstern. Theory of games and economic behavior. Princeton, 1853) has some association with metaphorical representation of human thought. It means that the idea of quantum probability in addition to Bayesian model must be applied to an analysis of text perception, particularly with respect to possible superposed state and complex interferences.

3. CONCLUSION

Developing techniques for graph representation of text content using vertex as feature term (or keyword) and edge as relation between keywords presents effective way to develop text summarization methods. Vector model in combination with probability analysis creates basis for text summarization converting a graph to a matrix form. Matrix representation of graph serves as effective model of text analysis in agglutinative languages with specific characteristics of syntax. Coherence between sub-concepts or subtopics of the text is closely connected with ideas of neural network and associative semantics and in that way presents interest for developing text summarization methods in terms of quantum semantics.

REFERENCES

[1] Surov. I.A, Semenenko. E, Bessmertny. A, Galofaro. F, Toffano. Z. Quantum semantics of text perception. Scientific reports. 11. 2021, p9

[2] Vimarch Rarbhar. What is LSA. 2020. Acing AI

[3] Mukundh Murthy. Guide to singular value decomposition. 2012, p7

[4] Patrizio. E, Tressoldi, Lance Storm, Dean Radin. Extrasensory perception and quantum model of cognition. NeuroQuantoloty. 2010. Vol 8. Issue 4, p585

[5] Tai Linzen T.Florian Jaeger. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions p1386. Cognitive science. 2016. Vol40. No6

[6] Arthur C. Graesser, Danielle S.McNamara. Computational analysis of multilevel discourse comprehension. P386. Topics in cognitive science. 2011. Vol3. No2