International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

Automatic Summarization Using an LDA-Based Similarity Measure

Mrs.B Sasikala , Mrs.Y Basanthi , Mrs.V Hemasree

Assistant Professor[1,2], Associate Professor[3]

Department of CSE,

Viswam Engineering College (VISM) Madanapalle-517325 Chittoor District, Andhra Pradesh, India

Abstract

To facilitate the automated summarizing of texts, this work introduces a new similarity measure. Latent Dirichlet Allocation is used to construct the topic space model. Each individual word, phrase, sentence, document, and corpus are all vectors in the same underlying subject space. The LMMR and LSD algorithms for creating summaries are shown. Experimental findings using DUC data demonstrate the efficacy and performance of the suggested measure and algorithm.

Keywords:*Latent Dirichlet Allocation, similarity measure, and automated text summarization.*

Introduction

An integral part of the Nature Language Processing community, automatic text summarization (ATS) (NLP). Abridged text summaries (ATSs) are described as "texts that are produced from one or more texts, which convey important information in the original text(s), and that are no longer than half of the original text(s) and usually significantly less than that"[1]. When studying ATS, researchers often use one of two approaches: either an abstraction or an extraction. The former prioritizes form above content, with the goal of creating a grammatical summary—an approach that often involves sophisticated language creation techniques [6]. However, thanks to advances in natural language processing (NLP), the abstraction can only be used in specific contexts. The focus of extraction is on the appropriateness of the summary's content, with sentences being the primary unit of extraction [6]. This method's broad applicability stems from the fact that it is independent of specific topic expertise. As a result of its superior performance, extraction has surpassed all other ATS methods in terms of popularity. The significance of the similarity measure in the extraction process cannot be overstated. There must be a method to calculate the degree of similarity between any two given sentences, documents, or corpora. Half of a successful extraction may be attributed to a decent similarity measure. In this study, we present the LDA-Sim measure, and the results of the trials show that it is successful. The paper will proceed as described below. Section 2 includes works that are relevant to this investigation. The Latent Dirichlet Allocation-based similarity metric is described in detail in Section 3. In Section 4, we cover the basics of the LMMR and LSD algorithms. In Section 5 we provide an example experiment to help illustrate the points made. In the last section, we provide our verdict.

Related work

Similarity between phrases may be determined by translating them into a common mathematical representation. In the popular Vector Space Model (VSM), documents are represented by vectors. Since the vocabulary is the space in which VSM operates, the dimension is often large. Latent semantic index (LSI) and probabilistic latent semantic index are only two examples of the types of dimension reduction studies that have received a lot of attention. (pulsi). Singular Value Decomposition (SVD) of the associated term/document matrix is used by LSI to implement a linear projection with reduced dimensions. Each document is converted to a probability distribution on a predetermined set of subjects by being expressed as a list of mixing proportions for various mixture components in pulsi. Next, we must settle on appropriate similarity functions. The cosine distance between the word vectors for each sentence or document, where the value for each word is the Term Frequency Inverse Document Frequency, is the typical similarity function used from Information Retrieval. (TFIDF). However, employing a basic word overlap metric yielded comparable results in certain experiments: shared non-stop word count adjusted for increased sentence length [7]. In 2003[4], a generative probabilistic model called the Latent Dirichlet Allocation (LDA) was introduced for collections of discrete data like text corpora. Recently,

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

the LDA has been successfully implemented in a wide variety of text information processing applications [3][5]. In this study, we implement the LDA as the basis for a similarity measure tool we call LDA-Sim. The subject space is a unified representation of the word, the phrase, the document, and the corpus. Then, we provide two LDA-Sim-involved methods (LMMR and LSD).

Similarity measure based on LDA.

Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a probabilistic model of a corpus that may be used to generate new data. Each latent subject is described by a distribution over words, and documents are then represented as random mixtures over these topics. Each document w in a corpus D is assumed to be generated using the following technique by LDA: Step 1: Decide on N Poisson (). Step 2: Select Dirichlet (). For each of the N words, won, in the list: (a) Select a Multinomial(z) Subject Area. (b) Pick a won from the list p (won | an,). It is assumed that the dimensionality k of the Dirichlet distribution (and thus z, the dependent variable) is constant. Parameterizing the word probabilities is a kV matrix where I j = p (we = 1|zi = 1), which we will assume is a constant for the time being. Finally, the remainder of the discussion does not need the Poisson assumption, thus other, more realistic distributions of document lengths may be employed if necessary. In addition, keep in mind that N is unrelated to the other variables (and z) that produce the data. As such, it is a non-essential factor, and its unpredictability will be disregardedin the next stage of development. In Figure 1, the LDA model is shown as a probabilistic graphical model. The image elucidates the LDA representation is three tiers. Both and are corpus-level parameters that are supposed to be sampled just once during the creation of a corpus. The variables are once-per-document samples at the document level. Word-level variables z and we are sampled once per document and once per word.

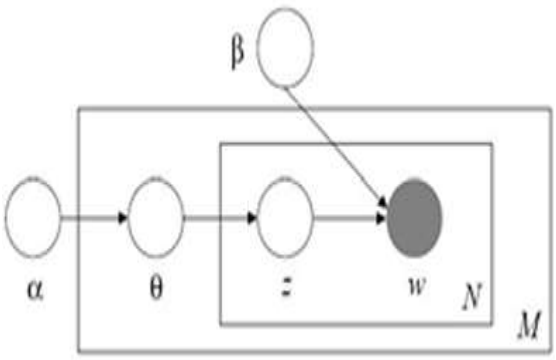


Fig 1: Graphical model representation of LDA

The EM method may be used to estimate the parameters in LDA. Determine the optimal values of the variational parameters for each document in the E-step. The log likelihood lower limit should be =maximizedabout the model parameters and in the M-step. This is analogous to identifying maximum likelihood estimates for each document that have statistically adequate expectations under the approximate posterior derived in the E-step. These two procedures are performed until the log likelihood convergence bound approaches zero.

LDA-Sim

Similarity of phrases may be quantified by expressing them in a common mathematical form. The word-space vector for traditional form. There is an incongruity between the extensive vocabulary and the sparseness of the sentences. That will make each line seem like a sparse vector represented in high dimensions. It is difficult to do calculations or discover connections between texts. A latent topic layer is created using the LDA model

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

described above. There are many fewer themes than there are words. The discussions here are intrinsically linked to the text itself. Therefore, it is enough for use as the starting point of the utterance. Word, phrase, document, and corpus may all be expressed in a unified way in the subject space. A word WI may be represented as a vector in the topic space, with each element's value being the subject's likelihood if it were associated with WI. This is denoted by L(WI)=(P(z1|wi), P (z2|wi) ..., P(kiwi)). Bayles's formula states:

$$P(z_i|w_i)=P(z_i)P(w_i|z_i)/P(w_i)$$

The latent topic distribution for a learned LDA model is a Dirichlet distribution with parameter k, where k= (1, 2..., k) Dirichlet (). In equation (1), P(zi) = I. The parameter allows us to calculate P(win). Simple statistical procedures allow us to compute P(WI), with P(WI)=count (WI)/N, where N is the total number of words in the vocabulary. Therefore, given WI, we can determine the likelihood of the subject. A word may be represented as a vector of concepts.

For a sentence $S=\{w_1,w_2,...,w_n\}$, calculating the average of the topic vectors of all words in S, we can get the topic vector of S, that is $L(S)=L(w_1,w_2,...,w_n)=(\sum P(z_1|w_i)/n, \sum P(z_2|w_i)/n, ..., \sum P(z_k|w_i)/n)$.

As the same way, we can get the topic vectors of a document $D=\{S_1,S_2,...S_n\}$ and a corpus $C=\{D_1,D_2,...D_n\}$.

So the word, sentence, document, corpus can be expressed as vectors in the same dimensionality. Then we can use the cosine distance to measure the similarity of any two of them. The similarity of two sentences is defined as:

$$LDA\text{-}Sim(S1,S2)=COS(L(S1),L(S2))$$

And the similarity between a sentence and a document can be defined as:

$$LDA\text{-}Sim(S,D)=COS(L(S),L(D))$$

The document's subject vector represents its most important ideas. The subject vector represents the overarching theme of a collection of texts. The LDA-Sim provides a solid basis for automated summarization due to its ability to assess any degree of similarity between a sentence and a document or corpus. The automated summarization may be achieved via the development of a number of algorithms.

The algorithms

LMMR

Automatic summarization is described by traditional MMR as a method of selecting the best candidate sentence to include in the summary. Select the statement that most accurately portrays the content of the text while minimizing repetition with the current summary. Iterate the steps above until the summary has grown to its maximum length. The formula they use to determine the following phrase is as follows:

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

$$MMR = argmax(t \cdot Sim1(Si,Q)-(1-t) \cdot max(Sim2(Si,Sj)))$$

The letter Q stands for a query, which comes from the field of data retrieval. Q is the essence of the material when using automated summarization. The two evaluative functions are called Sim1 and Sim2. The LDA-Sim may be used in this case. If we replace both with LDA-Sim, we get:

$$LMMR1=argmax(t(LDA\text{-}Sim(S,D))-(1-t)max(LDA\text{-}Sim(Si,Sj)))$$

If we substitute LDA-Sim for the Sim1 and maintain Sim2, we get:

$$LMMR2= argmax(t(LDA\text{-}Sim(S,D))-(1-t)max(Sim2(Si,Sj)))$$

LSD:

The LDA Sentence Descending Algorithm is provided here as an alternative. (LSD). One by one, eliminate the less crucial sentences until the summary is short enough. To determine how significant each phrase is, we use the LDA-Sim.

Given a document $D=\{S_1, S_2, \ldots, S_n\}$, the LSD is as follows:

- 1, Let $E=D$, i.e. $E=\{S_1, S_2, \ldots, S_n\}$.
- 2, find $j=argmax_{1\leq i\leq n}(LDASim(E, E-\{Si\}))$.
- 3, Let $E=E-\{S_j\}$.
- 4, go to step 2, until the length of E meets the threshold.

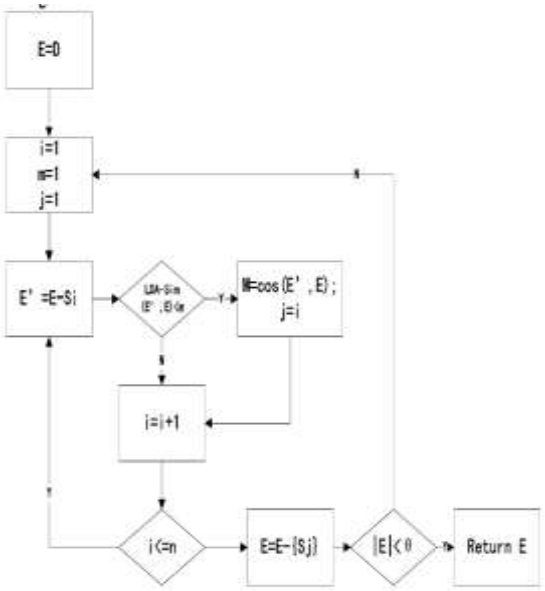
Figure 2 shows the flow chart.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86



Experiments:

Data

To promote study in automated multiloquent summarizing of news stories, the Text Analysis Conference (TAC) and its forerunner the Document Understanding Conference (DUC) host yearly conferences and accompanying contests. These conferences are the gold standard in the ATS literature since they compile a wealth of information and reference summaries into standardized datasets. In this paper, we provide findings from our analysis of the DUC datasets from 2006, 2007, 2008, and 2009. In 2006 and 2007, participants wereenquired to summarize sets of twenty-five items in 250 words. The 2008 and 2009 sets requested 100-word summaries of ten papers to move research away from focusing only on phrase extraction. About fifty similar issues are included in the annual problem set.

Results and analysis

As a starting point, we choose to take the first L words from the most recent input document, where L is the maximum length. We used LMMR and LSD on data sets from 2006, 2007, 2008, and 2009. The first thing we did was remove any potential stop words and then stem the words. The evaluation and final score are determined by using ROUGE2 and ROUGESU4. We evaluated the effectiveness of MMR and the Sum Basic algorithm against our own methods. The outcomes of ROUGE2 are shown in Table 1, whereas those of ROUGE-SU4 are shown in Table 2.

Table 1, ROUGE2 results

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

Dataset	Baseline	MMR	SumBasic	LMMR	LSD
2006	0.0060	0.0085	0.0092	0.0089	0.0094
2007	0.0059	0.0100	0.0119	0.0102	0.0121
2008	0.0060	0.0076	0.0097	0.0085	0.0105
2009	0.0063	0.0084	0.0103	0.0098	0.0108

Table 2, ROUGE-SU4 results

Dataset	Baseline	MMR	SumBasic	LMMR	LSD
2006	0.0108	0.0138	0.0140	0.0138	0.0141
2007	0.0106	0.0150	0.0165	0.0165	0.0169
2008	0.0091	0.0113	0.0130	0.0128	0.0135
2009	0.0099	0.0119	0.0134	0.0127	0.0135

Both LMMR and LSD perform better than the control group, as shown by the data. LMMR outperforms MMR but is inferior to Sum Basic. The LSD outperforms the Sum Basic by a little margin.

Conclusion

In this study, we suggest a new similarity metric for use in computer-generated summaries of texts. Using Latent Dirichlet Allocation, we construct a shared topic space in which individual words, sentences, documents, and whole corpora may beepresented as vectors. The conventional approach of creating summaries has been replaced with the LMMR and LSD algorithm. The results of the experiment support the usefulness of our similarity metric. The LMMR outperforms the MMR, while the LSD outperforms most current approaches.

Reference

[1] Inderjeet Mani. 2001. Automatic Summarization. John Benjamins Publisher.

[2] Mark Stivers, Tom Griffiths. 2007. Probabilistic Topic Models

[3] William M. Darling and Fei Song. 2011. Path Sum: A Summarization Framework Based on Hierarchical Topics

[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan .2003. Latent Dirichlet Allocation

[5] Rachit Arora, Balaraman Ravindran. 2008. Latent Dirichlet Allocation Based Multi-Document Summarization

[6] Dipanjan Das, Andre F.T. Martins. 2007. A Survey on Automatic Text Summarization

[7] Daniel Jacob Gillick. 2011. The Elements of Automatic Summarization