




Leveraging Non-negative Matrix Factorization for Document Summarization

Alka Khurana^(✉) 

Department of Computer Science, University of Delhi, Delhi 110007, India
akhurana@cs.du.ac.in

Abstract. This paper outlines the doctoral research work carried out to develop unsupervised approaches for extractive single document summarization using Non-negative Matrix Factorization (NMF). NMF is a popular topic modeling technique, which divulges inter-relation between three prime semantic units of the document. We exploit the inter-relationship among the three semantic units, viz. *terms*, *sentences*, and *latent topics*, in a novel manner to extract informative sentences from the salient topics in the document. The three methods developed during this doctoral study are language-, domain- and collection-independent.

Keywords: Extractive summarization · Non-negative matrix factorization · Entropy · Semantic similarity · Language independence

1 Problem and Motivation

Automatic Document Summarization (ADS) aims to shorten the text without compromising its essence. Additionally, it reduces readers' time and cognitive effort to comprehend the information in the document. Recent advances in ADS confirm its popularity and appositeness among the research community. The current revolution in neural methods and resulting masked language models in NLP have remarkably enhanced the performance of document summarization methods. However, the advantage comes with the ancillary cost of preparation of annotated data and the time required to train the models. The trained models inherit the training data characteristics and unintentionally acquire domain- and collection-dependence. Further, these models lack transparency and interpretability. The overheads associated with neural models galvanized the research work presented here. The approaches presented in this work are unsupervised, language-, domain- and collection-independent. Moreover, the methods are explainable and fast enough to summarize documents in real-time.

Supervised by Vasudha Bhatnagar.

2 Related Work

The earlier work by Lee et al. [4], who applied NMF for extractive single document summarization, provided the original impetus to our study. Peyrard postulated the idea of employing Shannon’s Entropy to capture *informativeness* required for document summarization, which is another stimulant for our research work [5]. We briefly describe NMF and Entropy in the present context. **(i) Non-negative Matrix Factorization:** NMF is a matrix decomposition technique to obtain reduced rank approximation in lower dimensional space. NMF holds the promise to uncover the latent semantic space of the document. Consider a document D represented as a sequence of n sentences (S_1, S_2, \dots, S_n) and consisting of m terms (t_1, t_2, \dots, t_m) . Let A be $m \times n$ term-sentence matrix for D , where an element a_{ij} in A denotes the occurrence of term t_i in sentence S_j . NMF decomposes A into W and H (i.e. $A \approx WH$), where W is $m \times r$ term-topic (feature) matrix, H is $r \times n$ topic-sentence (co-efficient) matrix and r is the number of latent topics $(\tau_1, \tau_2, \dots, \tau_r)$ in the document. Starting with initial seed values for factors, W and H , NMF iteratively improves these matrices such that Frobenius norm, $\|A - WH\|_F^2$ is minimized. Element w_{ij} in W quantifies the strength of term t_i in latent topic τ_j and element h_{ij} in H denotes contribution of sentence S_j in latent topic τ_i .

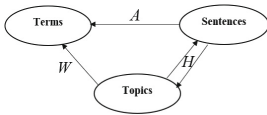


Fig. 1. Relationship among semantic units revealed by NMF decomposition. A : term-sentence, W : term-topic, H : topic-sentence matrix

Thus, Non-negative Matrix Factorization of *term-sentence* matrix A into *term-topic* feature matrix W and *topic-sentence* co-efficient matrix H , reveals the inter-relationship between the three semantics units, viz. terms, sentences, and latent topics (Fig. 1). Non-negativity constraints on NMF factor matrices enhance the interpretability of semantic units in the latent space. Random initialization of NMF factor matrices is the major caveat, which results in different W, H pairs (base models). We experimented with ensemble methods by combining the base models to create a consensus summary and found that ensembles often perform worse than the best base model [2].

(ii) Entropy of Semantic Units: *Informativeness* is the key attribute for capturing the essence of a document while generating the summary [5]. Inspired by the idea, we delve into the latent semantic space of the document revealed by NMF and mutate the semantic units into their corresponding probabilistic form. We provide different interpretations of topic and sentence entropy, and investigate their interplay.

3 Summarization Methods

This ongoing doctoral research proposes three novel approaches that excavate inter-relationship among the semantic units - *terms*, *sentences*, and *topics* in the document. The first approach *NMF-TR* is term-oriented, which leverages the

information carried by the terms in the latent semantic space exposed in the NMF feature matrix. Next is *NMF-TP*, a topic-oriented approach that exploits the information inferred by the latent topics uncovered by NMF [1]. The third approach is *E-Summ*, which follows an information-theoretic approach for selecting the informative sentences [3]. The three approaches score the sentences differently by using distinct inter-relationship between semantic units. Top scoring sentences are included in the summary.

(i) NMF-TR (Term-Oriented Sentence Scoring): Short documents are characterized by the small amount of information, which is insufficient to generate demarcated latent topics. Accordingly, considering the terms as the prime carriers of information in the document, we propose an algorithm *NMF-TR*, which is term-oriented sentence scoring method. *NMF-TR* considers that a term with high relative contribution is more important and that the importance of a sentence is an additive combination of importance of terms in the sentence.

$$Score(S_q) = \sum_{i=1}^m a_{iq} \phi_i, \text{ where } \phi_i = \frac{\sum_{q=1}^r w_{iq}}{\sum_{p=1}^m \sum_{q=1}^r w_{pq}} \quad (1)$$

That is, the sentence consisting of terms with higher contribution in latent space is preferred for inclusion in the summary. *NMF-TR* scores a sentence in the document according to Eq. 1.

(ii) NMF-TP (Topic-oriented Sentence Scoring): Long documents contain adequate information and concepts to yield clearly separated latent topics, and that the summary should reflect these latent topics proportionate to their importance in the document. Grounded on this idea, we propose *NMF-TP*, which is a topic-oriented method that explicitly considers the importance of latent topics in the document to extract salient sentences for the summary.

$$Score(S_q) = \sum_{i=1}^r \omega_i h_{iq}, \text{ where } \omega_i = \frac{\sum_{q=1}^m w_{qi}}{\sum_{p=1}^m \sum_{q=1}^r w_{pq}} \quad (2)$$

Since sentences are carriers of the information reflected by the latent topics, therefore, *NMF-TP* scores a sentence in the document using the formula in Eq. 2.

(iii) E-Summ: *E-Summ* algorithm is grounded on the principle of selecting informative sentences from the document, which convey important topics. *E-Summ* follows an information-theoretic approach, exploits the information contained in topic and sentence entropies, and identifies candidates by scoring a sentence as follows.

$$Score(S_j) = \zeta^\tau(S_j) + \Psi^S(\tau_i), \text{ where}$$

$\zeta^\tau(S_j)$ is sentence entropy in topic space and $\Psi^S(\tau_i)$ is topic entropy in sentence space (Refer [3] for details). Subsequently, it uses Knapsack optimization algorithm to maximize the information conveyed by sentences selected for inclusion in the summary. Since *E-Summ* works on the premise of maximizing information contained in summary and does not explicitly consider the terms and topics as in the case of *NMF-TR* and *NMF-TP*, it is reasonable to employ *E-Summ* to summarize short and long documents.

4 Data-Sets and Evaluation

We use well known public data-sets - DUC2001, DUC2002, CNN and DailyMail for performance evaluation of the proposed algorithms. The proposed methods are language-, domain- and collection-independent. The methods are faster than state-of-the-art extractive summarization algorithms (Refer [1, 3] for details) and are suitable for real time summarization of web documents.

We evaluate the performance of all the three algorithms viz. *NMF-TR*, *NMF-TP*, and *E-Summ* on DUC2001, DUC2002, and CNN/DailyMail data-sets. Additionally, we employ WikiHow data-set consisting of general Wikipedia articles and CL-SciSumm data-sets containing scientific articles, for justifying the domain independence of *E-Summ* algorithm. We demonstrate the language independence of *E-Summ* using two Indian and three European languages.

We use standard ROUGE measure for qualitative assessment of the algorithmic summaries. ROUGE is a measure of computing lexical overlap between algorithmic summaries and gold standard ground truth summaries. We additionally employ semantic similarity measure to evaluate the quality of summaries.

5 Conclusion and Future Directions

Growing demand of automatic summarization methods in diverse domains, genres and for resource poor languages necessitate development of unsupervised, language independent and fast methods. We use Non-negative Matrix Factorization to tease out the inter-relationship between terms, sentences and latent topics in the latent semantic space of the document and proposed three single document summarization algorithms. As expected, the performance of the proposed algorithms does not match with that of deep neural methods.

Existing research majorly focuses on generating generic summaries of documents. However, generic summaries ignore the subjectivity aspect as these summaries do not consider the background knowledge of the user. Our current focus is on automatically generating personalized summary of the document.

References

1. Khurana, A., Bhatnagar, V.: Extractive document summarization using non-negative matrix factorization. In: Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) DEXA 2019. LNCS, vol. 11707, pp. 76–90. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27618-8_6
2. Khurana, A., Bhatnagar, V.: NMF ensembles? not for text summarization! In: Proceedings of the First Workshop on Insights from Negative Results in NLP, pp. 88–93 (2020)
3. Khurana, A., Bhatnagar, V.: Investigating entropy for extractive document summarization. *Expert Syst. Appl.* **187**, 115820 (2021)
4. Lee, J.H., Park, S., Ahn, C.M., Kim, D.: Automatic generic document summarization based on non-negative matrix factorization. *Inf. Process. Manage.* **45**(1), 20–34 (2009)
5. Peyrard, M.: A simple theoretical model of importance for summarization. In: Proceedings of the 57th Annual Meeting of the ACL (2019)