Explaining Pairwise Relationships Between Documents

Nils Witt(✉)

ZBW-Leibniz Information Centre for Economics, Kiel, Germany
n.witt@zbw.eu

Abstract. Current methods in automatic text summarization only take a single document into account. In contrast, the proposed research in this paper aims at summarizing the discrepancy between two documents. We approach this by transforming documents into a representation where mathematical operations have a semantical correspondence. This allows us to recombine documents and draw a summary from the recombined vector. A discrepancy summary can briefly convey what a reader can additionally learn from reading an unknown document with respect to a document that the reader is already familiar with.

1 Introduction

In natural language processing(NLP) applications documents are usually transformed into a new reference system which allows applying mathematical measure to determine their differences. We want to investigate these differences between two documents with the goal of acquiring human-interpretable knowledge. To achieve this goal we developed an abstract model that captures and formalizes automatic dissimilarity summarization(ADS). This model will be accompanied by an evaluation framework that objectively compares different implementations of this model. Afterwards we will evaluate the performance of different implementations and investigate enhancements. To the best of our knowledge this question has not been studied, albeit the potential benefits in NLP applications. Word2Vec [4,7] and its increments (e.g. [4,9]) are promising candidates for a suitable data representation for ADS, as they comprise a shallow understanding of language. Summaries generated by ADS are more beneficial to users than plain summaries as they take the history of documents that a user has already read into account. These information can support users by choosing the next document to read with respect to previously read documents.

2 Research Questions

There is a vast amount of text representations available to determine the difference of two documents mathematically. These representations have important practical applications but fall short on depicting their insights to humans directly, which is the motivation for the proposed research. We have identified the following research questions:

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 648–651, 2017.
DOI: 10.1007/978-3-319-67008-9_62

- What kind of text representation is most appropriate for automatic discrepancy summarization?
- Is the information that is contained in the order of words better encoded by a sequence of word representations or by a document-wide representation?
- What is the most suitable way to convey the discrepancy between documents (text-based, graphical or other representations)?
- How can we determine the quality of a summary?

These questions are governed by two areas of research in natural language processing: *(1) Information Representation:* Often simple word count models like bag of words [3] and tf-idf [5] are used to represent text. In contrast, topic models like LDA [2] represent entire documents by uncovering their latent topic structure and encoding each document in a collection as a composition of those topics. Word embedding models [1] map each word in a vocabulary onto a vectors of real numbers. These vectors are arranged such that they encode semantical and syntactical properties of the language. *(2) Automatic Text Summarization:* Automatic text summarization is the process of shortening a text document while preserving salient aspects of it. There are two different approaches to the problem. *(1) Extraction:* Extraction methods try to find a subset of sentences from the original document that cover salient aspects of the original document. Which is what TextRank [6] does, by constructing a graph where every sentences is a vertex that has an edge to sentences that are sufficiently equal to the sentence. Subsequently, PageRank is used to determine the importance of each sentence. *(2) Abstraction:* Abstraction methods build semantical representation of text documents an derive a summary from that representation, which mimics what people do when they summarize a document.

3 Approach

The proposed research wants to represent the content differences of two documents such that they are human-interpretable, easily comprehensible and quickly to compute. At first an estimator (e.g. Recurrent Neural Network (RNN)) is trained to recover a text that generated the corresponding data representation. Secondly, a recombination takes place that generates a new data representation. And lastly, the trained estimator is used to find a coherent interpretation of this newly created data representation (see Fig. 1). The model is required to learn the language of the documents to a degree that allows it to generate correct language which is plausible as Sutskever et al. [8] have been able to create a RNN that generates English text despite the fact that it was trained on character level. Since there is no proper baseline available that we can compare against, the initial step is to create a simple implementation of the model described in Fig. 1. This includes a sentence-wise tf-idf encoding and a simple Gaussian naive Bayes classifier. Afterwards we will create and apply an evaluation framework that reflects the desired properties:

- **Correctness**: The summary produced is coherent given the two documents that it was derived from. A simple approach to test this is to consider the

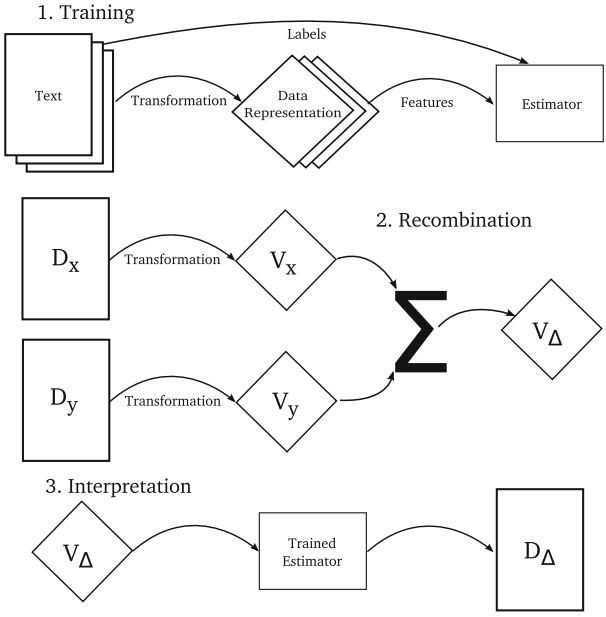

Fig. 1. Abstract model of the automatic discrepancy summarization. During the training phase, an estimator learns to recreate a text from its data representation. In the recombination phase, a new vectors gets created by applying some function (e.g. subtraction) onto those two vectors (The *D*s refer to a text unit whereas the *V*s are the corresponding data representations). The Interpretation phase transforms the newly created vector into a human-understandable form.

extremes: What is the result of comparing a document with itself? Since there is no discrepancy, the summary should be empty. What happens when two topically orthogonal documents are examined? The summary should be dominated by one of the documents.
- **Interpretable**: Is a user able to comprehend the results produced and draw conclusion from them? For example, if the summary is a sentence, the sentence must be syntactically, grammatically and semantically correct.
- **Generalizable**: The technique must work on text documents across different domains. Moreover, the model must produce coherent results on domains that were not covered in the set of training documents.
- **Deterministic**: The system always produces the same output when it is given the same inputs.

Subsequently, more sophisticated implementations will be investigated and compared against the baseline results. Currently, two experiments are projected:

1. **Topic Models**: A topic model will be used to represent the text units. Since topic models preserve more structure of the text, it is expected to obtain better results.

2. **Text Embeddings**: In this experiment we want to investigate the question whether the sequential nature of text documents should be captured by the representation of the text or by the estimator. This leads to two different settings: (1) Sequence-aware model: Sentences will be transformed into a sequence of word vectors. The estimator will be a RNN. (2) Sequence-aware data representation: Sentences will be transformed into a single paragraph vector. The estimator will be a support vector machine.

4 Preliminary Results

In [10] we have derived words describing the difference between two documents from their document embeddings. That is, given two document vectors d_0 and d_1, we compiled a set of word vectors $w_0 \dots w_n$ such that $d_0 \approx d_1 \circ w_0 \circ w_1 \circ \dots \circ w_n$ ($\circ$ was either the summation or the subtraction operator). We approached that task with an iterative process that generated a word at each iteration. It began by finding the word that maximizes $similarity(d_0, d_1 \circ w_0)$ and continued by maximizing $similarity(d_0 \circ w_0, d_1 \circ w_0 \circ w_1)$ etc. We found that this leads to a list of words that is descriptive of the content difference of d_0 and d_1. However, we realized that words are not the correct text unit for the given task, as they provide to few information and, lacking context, lead to ambiguous and false conclusions. Hence, we want to investigate alternative methods that may overcome the drawbacks while preserving the beneficial properties of our previous solution.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Harris, Z.S.: Distributional structure. Word **10**(2–3), 146–162 (1954)
4. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. arXiv preprint arXiv:1405.4053 (2014)
5. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM J. Res. Dev. **1**(4), 309–317 (1957)
6. Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: AFCL (2004)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013)
8. Sutskever, I., Martens, J., Hinton, G.E.: Generating text with recurrent neural networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 1017–1024 (2011)
9. Trask, A., Michalak, P., Liu, J.: sense2vec-a fast and accurate method for word sense disambiguation. arXiv preprint arXiv:1511.06388 (2015)
10. Witt, N., Seifert, C., Granitzer, M.: Explaining topical distances using word embeddings. In: Database and Expert Systems Applications, 2016 27th International Workshop on Text-based Information Retrieval, pp. 212–217. IEEE (2016)