Summarization Based on Embedding Distributions

Hayato Kobayashi **Masaki Noguchi** **Taichi Yatsuka**
Yahoo Japan Corporation
9-7-1 Akasaka, Minato-ku, Tokyo 107-6211, Japan
{hakobaya, manoguch, tyatsuka}@yahoo-corp.jp

Abstract

In this study, we consider a summarization method using the document level similarity based on embeddings, or distributed representations of words, where we assume that an embedding of each word can represent its "meaning." We formalize our task as the problem of maximizing a submodular function defined by the negative summation of the nearest neighbors' distances on *embedding distributions*, each of which represents a set of word embeddings in a document. We proved the submodularity of our objective function and that our problem is asymptotically related to the KL-divergence between the probability density functions that correspond to a document and its summary in a continuous space. An experiment using a real dataset demonstrated that our method performed better than the existing method based on sentence-level similarity.

1 Introduction

Document summarization aims to rephrase a document in a short form called a summary while keeping its "meaning." In the present study, we aim to characterize the meaning of a document using embeddings or distributed representations of words in the document, where an embedding of each word is represented as a real valued vector in a Euclidean space that corresponds to the word (Mikolov et al., 2013a; Mikolov et al., 2013b).

Many previous studies have investigated summarization (Lin and Bilmes, 2010; Lin and Bilmes, 2011; Lin and Bilmes, 2012; Sipos et al., 2012; Morita et al., 2013), but to the best of our knowledge, only one (Kågebäck et al., 2014) considered a direct summarization method using embeddings, where the summarization problem was formalized as maximizing a submodular function defined by the summation of cosine similarities based on sentence embeddings. Essentially, this method assumes linear meanings since the objective function is characterized by the summation of sentence-level similarities. However, this assumption is not always valid in real documents, and thus there may be a better combination of two other sentences than the best and second best sentences in terms of similarity in a document.

In this study, we consider a summarization method based on document-level similarity, where we assume the non-linearity of meanings. First, we examine an objective function defined by a cosine similarity based on document embeddings instead of sentence embeddings. Unfortunately, in contrast to our intuition, this similarity is not submodular, which we disprove later. Thus, we propose a valid submodular function based on *embedding distributions*, each of which represents a set of word embeddings in a document, as the document-level similarity. Our objective function is calculated based on the nearest neighbors' distances on embedding distributions, which can be proved to be asymptotically related to KL-divergence in a continuous space. Several studies (Lerman and McDonald, 2009; Haghighi and Vanderwende, 2009) have addressed summarization using KL-divergence, but they calculated KL-divergence based on word distributions in a discrete space. In other words, our study is the first attempt to summarize by asymptotically estimating KL-divergence based on embedding distributions in a continuous space. In addition, they involved the inference of complex models, whereas our method is quite simple but still powerful.

2 Preliminaries

We treat a document as a bag-of-sentences and a sentence as a bag-of-words. Formally, let D be a document, and we refer to an element

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1984–1989,
Lisbon, Portugal, 17-21 September 2015.

$s \in D$ of a sentence and $w \in s$ of a word. We denote the size of a set S by $|S|$. Note that D and s are defined as multisets. For example, we can define a document such as $D := \{s_1, s_2\}$ with $s_1 := \{\texttt{just}, \texttt{do}, \texttt{it}\}$ and $s_2 := \{\texttt{never}, \texttt{say}, \texttt{never}\}$, which correspond to two sentences "Just do it" and "Never say never," respectively. From the definition, we have $|s_1| = 3$ and $|s_2| = 3$.

2.1 Submodularity

Submodularity is a property of set functions, which is similar to the convexity or concavity of continuous functions.

We formally define submodularity as follows.

Definition 1 (Submodularity)**.** Given a set X, a set function $f : 2^X \rightarrow \mathbb{R}$ is called *submodular* if for any two sets S_1 and S_2 such that $S_1 \subset S_2 \subset X$ and element $x \in X \setminus S_2$,

$$f(S_1 \cup \{x\}) - f(S_1) \geq f(S_2 \cup \{x\}) - f(S_2).$$

For simplicity, we define $f_S(x) := f(S \cup \{x\}) - f(S)$, which is called *the marginal value of x with respect to S*. A set function f is called *monotone* if $f_S(x) \geq 0$ for any set $S \subset X$ and element $x \in X \setminus S$.

If a set function f is monotone submodular, we can approximate the optimal solution efficiently by a simple greedy algorithm, which iteratively selects $x^* = \mathrm{argmax}_{x \in X \setminus S_i} f_{S_i}(x)$ where ties are broken arbitrarily, and we substitute $S_{i+1} = S_i \cup \{x^*\}$ in the i-th iteration beginning with $S_0 = \emptyset$. This algorithm is quite simple but it is guaranteed to find a near optimal solution within $1 - 1/\mathrm{e} \approx 0.63$ (Calinescu et al., 2007).

2.2 Embedding

An embedding or distributed representation of a word is a real valued vector in an m-dimensional Euclidean space $\mathbb{R}^m$, which expresses the "meaning" of the word. We denote an embedding of a word w by $\vec{w} \in \mathbb{R}^m$. If for any two words w_1 and w_2, the meaning of w_1 is similar to that of w_2, then $\vec{w}_1$ is expected to be near to $\vec{w}_2$.

A recent study (Mikolov et al., 2013a) showed that a simple log-bilinear model can learn high quality embeddings to obtain a better result than recurrent neural networks, where the concept of embeddings was originally proposed in studies of neural language models (Bengio et al., 2003). In the present study, we use the CW Vector[1] and W2V Vector[2] which are also used in the previous study (Kågebäck et al., 2014).

3 Proposed Method

In this study, we focus on a summarization task as sentence selection in a document. The optimization framework in our task is the same as in the previous study and formalized in Algorithm 1, where w_s represents the pre-defined weight or cost of a sentence s, e.g., sentence length, and r is its scaling factor. This algorithm, called modified greedy, was proposed in (Lin and Bilmes, 2010) and interestingly performed better than the state-of-the-art abstractive approach as shown in (Lin and Bilmes, 2011). Note that we have omitted the notation of D from f for simplicity because D is fixed in an optimization process.

Algorithm 1: Modified greedy algorithm.

Data: Document D, objective function f, and summary size ℓ.
Result: Summary $C \subset D$.

$C \leftarrow \emptyset$; $U \leftarrow D$;
while $U \neq \emptyset$ **do**
  $s^* \leftarrow \mathrm{argmax}_{s \in U} f_C(s)/(w_s)^r$;
  if $\sum_{s \in C} w_s + w_{s^*} \leq \ell$ **then** $C \leftarrow C \cup \{s^*\}$;
  $U \leftarrow U \setminus \{s^*\}$;
$s^* \leftarrow \mathrm{argmax}_{s \in D : w_s \leq \ell} f(\{s\})$;
return $C \leftarrow \mathrm{argmax}_{C' \in \{C, \{s^*\}\}} f(C')$;

Similarity Based on Document Embeddings

First, we examine an objective function f^{Cos} defined by a cosine similarity based on document embeddings. An embedding of a document D is defined as $\mathbf{v}_D := \sum_{s \in D} \sum_{w \in s} \vec{w}$. We formalize the objective function f^{Cos} as follows.

$$f^{Cos}(C) := \frac{\mathbf{v}_C \cdot \mathbf{v}_D}{\|\mathbf{v}_C\| \, \|\mathbf{v}_D\|}.$$

Note that the optimal solution does not change, if we use an average embedding $\mathbf{v}_D / \sum_{s \in D} |s|$ instead of $\mathbf{v}_D$. The next theorem shows that a solution of f^{Cos} by Algorithm 1 is not guaranteed to be near optimal.

Theorem 1. *f^{Cos} is not submodular.*

[1] http://metaoptimize.com/projects/wordreprs

[2] https://code.google.com/p/word2vec

Proof. A simple counterexample is sufficient to prove the theorem. Let us consider $D := \{s_1 := \{w_1\},\ s_2 := \{w_2\},\ s_3 := \{w_3\},\ s_4 := \{w_4\}\}$ with corresponding vectors $\vec{w}_1 := (1,1)$, $\vec{w}_2 := (1,2)$, $\vec{w}_3 := (1,-1)$, and $\vec{w}_4 := (1,-2)$, respectively. In this case, the document embedding $\mathbf{v}_D$ is $(4,0)$. We set $C_1 := \{s_1\}$ and $C_2 := \{s_1, s_2\}$. Clearly, $C_1 \subset C_2$. However, we obtain $f^{Cos}_{C_1}(s_4) = f^{Cos}(\{s_1, s_4\}) - f^{Cos}(\{s_1\}) \approx 0.187$ and $f^{Cos}_{C_2}(s_4) = f^{Cos}(\{s_1, s_2, s_4\}) - f^{Cos}(\{s_1, s_2\}) \approx 0.394$. Therefore, we have $f^{Cos}_{C_2}(s_4) > f^{Cos}_{C_1}(s_4)$. □

Similarity Based on Embedding Distributions

We propose a valid submodular objective function f^{NN} based on embedding distributions. The key observation is that for any two embedding distributions A and B, when A is similar to B, each embedding in A should be near to some embedding in B. In order to formalize this idea, we define *the nearest neighbor of a word w in a summary C* as $n(w, C) := \mathrm{argmin}_{v \in s : s \in C, \vec{w} \neq \vec{v}}\, d(\vec{w}, \vec{v})$, where d is the Euclidian distance in the embedding space, i.e., $d(\vec{w}, \vec{v}) := \|\vec{w} - \vec{v}\|$. We denote the distance of w to its nearest neighbor $n := n(w, C)$ by $N(w, C) := d(\vec{w}, \vec{n})$. Finally, we define f^{NN} as follows:

$$f^{NN}(C) := -\sum_{s \in D} \sum_{w \in s} g(N(w, C)),$$

where g is a non-decreasing scaling function. The function f^{NN} represents the negative value $-\delta$ of dissimilarity δ between a document and summary based on embedding distributions. Note that we can use sentence embeddings instead of word embeddings as embedding distributions, although we focus on word embeddings in this section.

The next theorem shows the monotone submodularity of our objective function, which means that a solution of f^{NN} by Algorithm 1 is guaranteed to be near optimal.

Theorem 2. *f^{NN} is monotone submodular.*

Proof. (Monotonicity) First, we prove the monotonicity. For simplicity, we use the following two abbreviations: $C^s := C \cup \{s\}$ and $\sum_w^D := \sum_{s \in D} \sum_{w \in s}$. For any set $C \subset D$ of sentences and sentence $s \in D \setminus C$, we have $f^{NN}_C(s) = f^{NN}(C^s) - f^{NN}(C) = \sum_w^D (g(N(w, C)) - g(N(w, C^s)))$. Since $C \subset C^s$, obviously $N(w, C) \geq N(w, C^s)$ holds. Therefore, we obtain $f^{NN}_C(s) \geq 0$ from the non-decreasing property of g.

(Submodularity) Next, we prove the submodularity. For any two sets C_1 and C_2 of sentences such that $C_1 \subset C_2 \subset D$, and sentence $s \in D \setminus C_2$, we have $f^{NN}_{C_1}(s) - f^{NN}_{C_2}(s) = f^{NN}(C^s_1) - f^{NN}(C_1) - (f^{NN}(C^s_2) - f^{NN}(C_2)) = \sum_w^D (g(N(w, C_1)) - g(N(w, C^s_1)) - g(N(w, C_2)) + g(N(w, C^s_2)))$. Let $\alpha := g(N(w, C_1)) - g(N(w, C^s_1)) - g(N(w, C_2)) + g(N(w, C^s_2))$.

If $n(w, C^s_2) \in s$, then $n(w, C^s_1) \in s$ holds, since $C^s_1 \subset C^s_2$. This means that $N(w, C^s_2) = N(w, C^s_1) = N(w, \{s\})$. Clearly, $N(w, C_1) \geq N(w, C_2)$, since $C_1 \subset C_2$. Therefore, we obtain $\alpha \geq 0$ from the non-decreasing property of g.

If $n(w, C^s_2) \notin s$ and $n(w, C^s_1) \notin s$, we have $N(w, C^s_1) = N(w, C_1)$ and $N(w, C^s_2) = N(w, C_2)$. This indicates that $\alpha = 0$.

If $n(w, C^s_2) \notin s$ and $n(w, C^s_1) \in s$, so similarly $N(w, C^s_1) \leq N(w, C_1)$ and $N(w, C^s_2) = N(w, C_2)$ hold. Therefore, we obtain $\alpha \geq 0$. □

The objective function f^{NN} is simply heuristic for small documents, but the next theorem shows that f^{NN} is asymptotically related to an approximation of KL-divergence in a continuous space, if g is a logarithmic function. This result implies that we can use mathematical techniques of a continuous space for different NLP tasks, by mapping a document into a continuous space based on word embeddings.

Theorem 3. *Suppose that we have a document D and two summaries C_1 and C_2 such that $|C_1| = |C_2|$, which are samples drawn from some probability density functions p, q, and r, i.e., $D \sim p$, $C_1 \sim q$, and $C_2 \sim r$, respectively. If the scaling function g of f^{NN} is a logarithmic function, the order relation of the expectations of $f^{NN}(C_1)$ and $f^{NN}(C_2)$ is asymptotically the same as that of the KL-divergences $\mathcal{D}_{KL}(p \,||\, r)$ and $\mathcal{D}_{KL}(p \,||\, q)$, i.e., $\mathbb{E}[f^{NN}(C_1)] - \mathbb{E}[f^{NN}(C_2)] > 0 \Leftrightarrow \mathcal{D}_{KL}(p \,||\, r) - \mathcal{D}_{KL}(p \,||\, q) > 0$, as $|D| \to \infty$, $|C_1| \to \infty$, and $|C_2| \to \infty$.*

Proof. Let m be the dimension on embeddings. Using a divergence estimator based on nearest neighbor distances in (Pérez-Cruz, 2009; Wang et al., 2009), we can approximate $\mathcal{D}_{KL}(p \,||\, q)$ by $\widehat{\mathcal{D}}_{KL}(D, C_1) := \frac{m}{|D|} \sum_w^D \ln \frac{N(w,C_1)}{N(w,D)} + \ln \frac{|C_1|}{|D|-1}$. Therefore, we obtain $\widehat{\mathcal{D}}_{KL}(D, C_2) - \widehat{\mathcal{D}}_{KL}(D, C_1) \propto \sum_w^D \ln N(w, C_2) - \sum_w^D \ln N(w, C_1)$. Since $g(x) = \ln(x)$, we have $f^{NN}(C_1) - f^{NN}(C_2) > 0$ if

and only if $\widehat{\mathcal{D}}_{KL}(D, C_2) - \widehat{\mathcal{D}}_{KL}(D, C_1) > 0$ holds. The fact that $\mathbb{E}[\widehat{\mathcal{D}}_{KL}(D, C_1)] \rightarrow \mathcal{D}_{KL}(p \mid\mid q)$ as $|C_1| \rightarrow \infty$ and $|D| \rightarrow \infty$ concludes the theorem. □

4 Experiments

We compared our two proposed methods DocEmb and EmbDist with two state-of-the-art methods SenEmb and TfIdf. The first two methods DocEmb and EmbDist represent Algorithm 1 with our proposed objective functions f^{Cos} and f^{NN}, respectively. TfIdf represents Algorithm 1 with an objective function based on the sum of cosine similarities of tf-idf vectors that correspond to sentences, which was proposed in (Lin and Bilmes, 2011). SenEmb uses a cosine similarity measure based on embeddings instead of tf-idf vectors in the same framework as TfIdf, which was proposed in (Kågebäck et al., 2014).

We conducted an experiment with almost the same setting as in the previous study, where they used the Opinosis dataset (Ganesan et al., 2010). This dataset is a collection of user reviews in 51 different topics such as hotels, cars, and products; thus, it is more appropriate for evaluating summarization of user-generated content than well-known DUC datasets, which consist of formal news articles. Each topic in the collection comprises 50–575 sentences and includes four and five gold standard summaries created by human authors, each of which comprises 1–3 sentences.

We ran an optimization process to choose sentences within 100 words[3] by setting the summary size and weights as $\ell = 100$ and $w_s = |s|$ for any sentence s, respectively. As for TfIdf and SenEmb, we set a cluster size of k-means as $k = |D|/5$ and chose the best value for a threshold coefficient α, trade-off coefficient λ, and the scaling factor r, as in (Lin and Bilmes, 2011). Note that our functions DocEmb and EmbDist have only one parameter r, and we similarly chose the best value of r. Regarding DocEmb, EmbDist, and SenEmb, we used the best embeddings from the CW Vector and W2V Vector for each method, and created document and sentence embeddings by averaging word embeddings with tf-idf weights since it performed better in this experiment. In the case of EmbDist, we used a variant of f^{NN} based on distributions of sentence embeddings. In addition, we examined three scaling functions: logarithmic, linear, and exponential functions, i.e., $\ln x$, x, e^x, respectively.

[3]The previous work used a sentence-based constraint as $\ell = 2$ and $w_s = 1$, but we changed the setting since the variation in length has a noticeable impact on ROUGE scores as suggested in (Hong et al., 2014).

	R-1	R-2	R-3	R-4
ApxOpt	62.22	21.60	8.71	4.56
EmbDist ($\ln x$)	56.00	16.70	4.93	**1.89**
EmbDist (x)	55.70	15.73	4.59	1.84
EmbDist (e^x)	**56.29**	15.96	4.43	1.39
DocEmb	55.80	13.59	3.23	0.90
SenEmb	53.96	15.42	3.97	1.10
TfIdf	52.97	**17.24**	**5.40**	1.49

Table 1: ROUGE-N (R-N) metrics of DocEmb, EmbDist, SenEmb, and TfIdf.

We calculated the ROUGE-N metric (Lin, 2004)[4], which is a widely-used evaluation metric for summarization methods. ROUGE-N is based on the co-occurrence statistics of N-grams, and especially ROUGE-1 has been shown to have the highest correlation with human summaries (Lin and Hovy, 2003). ROUGE-N is similar to the BLEU metric for machine translation, but ROUGE-N is a recall-based metric while BLEU is a precision-based metric.

Table 1 shows the results obtained for ROUGE-N (N ≤ 4) using DocEmb, EmbDist, SenEmb, and TfIdf. ApxOpt represents the approximation results of the optimal solution in our problem, where we optimized ROUGE-1 with the gold standard summaries by Algorithm 1. The obtained results indicate that our proposed method EmbDist with exponential scaling performed the best for ROUGE-1, which is the best metric in terms of correlation with human summaries. The W2V Vector was the best choice for EmbDist. Furthermore, the other proposed method DocEmb performed better than the state-of-the-art methods SenEmb and TfIdf, although DocEmb is not theoretically guaranteed to obtain a near optimal solution. These results imply that our methods based on the document-level similarity can capture more complex meanings than the sentence-level similarity. On the other hand, TfIdf with tf-idf vectors performed the worst for ROUGE-1. A possible reason is that a wide variety of expressions by users made it difficult to calculate similarities. This also suggests that embedding-based methods

[4]We used their software ROUGE version 1.5.5 with the parameters: -n 4 -m -a -l 100 -x -c 95 -r 1000 -f A -p 0.5 -t 0.

naturally have robustness for user-generated content.

In the case of N $\geq$ 2, `TfIdf` performed the best for ROUGE-2 and ROUGE-3, while `EmbDist` with logarithmic scaling is better than `TfIdf` for ROUGE-4. According to (Lin and Hovy, 2003), the higher order ROUGE-N is worse than ROUGE-1 since it tends to score grammaticality rather than content. Conversely, Rankel et al. (2013) reports that there is a dataset where the higher order ROUGE-N is correlated with human summaries well. We may need to conduct human judgments to decide which metric is the best in this dataset for more accurate comparison. However, it is still important that our simple objective functions can obtain good results competing with the state-of-the-art methods.

5 Conclusion

In this study, we proposed simple but powerful summarization methods using the document-level similarity based on embeddings, or distributed representations of words. Our experimental results demonstrated that the proposed methods performed better than the existing state-of-the-art methods based on the sentence-level similarity. This implies that the document-level similarity can capture more complex meanings than the sentence-level similarity.

Recently, Kusner et al. (2015) independently discovered a similar definition to our objective function f^{NN} through a different approach. They constructed a dissimilarity measure based on a framework using Earth Mover's Distance (EMD) developed in the image processing field (Rubner et al., 1998; Rubner et al., 2000). EMD is a consistent measure of distance between two distributions of points. Interestingly, their heuristic lower bound of EMD is exactly the same as $-f^{NN}$ with a linear scaling function, i.e., $g(x) = x$. Moreover, they showed that this bound appears to be tight in real datasets. This suggests that our intuitive framework can theoretically connect the two well-known measures, KL-divergence and EMD, based on the scaling of distance. Note that, to the best of our knowledge, there is currently no known study that considers such a theoretical relationship.

In future research, we will explore other scaling functions suitable for our problem or different problems. A promising direction is to consider a relative scaling function to extract a biased summary of a document. This direction should be useful for query-focused summarization tasks.

Acknowledgments

The authors would like to thank the reviewers for their helpful comments, especially about Earth Mover's Distance.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research (JMLR 2003)*, 3:1532–4435.

Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. 2007. Maximizing a Submodular Set Function Subject to a Matroid Constraint (Extended Abstract). In *Proceedings of the 12th International Conference on Integer Programming and Combinatorial Optimization (IPCO 2007)*, pages 182–196. Springer-Verlag.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A Graph-based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 340–348. Association for Computational Linguistics.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-document Summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2009)*, pages 362–370. Association for Computational Linguistics.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 26–31. European Language Resources Association (ELRA).

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. Extractive Summarization using Continuous Vector Space Models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC 2014)*, pages 31–39. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 957–966. JMLR.org.

Kevin Lerman and Ryan McDonald. 2009. Contrastive Summarization: An Experiment with Consumer Reviews. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2009)*, pages 113–116. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2010. Multi-document Summarization via Budgeted Maximization of Submodular Functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010)*, pages 912–920. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2011. A Class of Submodular Functions for Document Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 510–520. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2012. Learning Mixtures of Submodular Shells with Application to Document Summarization. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI 2012)*, pages 479–490. AUAI.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2003)*. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*. Association for Computational Linguistics.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013a. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119. Curran Associates, Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 746–751. Association for Computational Linguistics.

Hajime Morita, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2013. Subtree Extractive Summarization via Submodular Maximization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1023–1032. Association for Computational Linguistics.

Fernando Pérez-Cruz. 2009. Estimation of Information Theoretic Measures for Continuous Random Variables. In *Advances in Neural Information Processing Systems 21 (NIPS 2009)*, pages 1257–1264. Curran Associates, Inc.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 131–136. Association for Computational Linguistics.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 1998. A Metric for Distributions with Applications to Image Databases. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV 1998)*, pages 59–66. IEEE Computer Society.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Ruben Sipos, Adith Swaminathan, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Temporal Corpus Summarization Using Submodular Word Coverage. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*, pages 754–763. ACM.

Qing Wang, S.R. Kulkarni, and S. Verdu. 2009. Divergence Estimation for Multidimensional Densities Via k-Nearest-Neighbor Distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405.