Summarization Based on Embedding Distributions

Hayato Kobayashi, Masaki Noguchi, Taichi Yatsuka, Yahoo Japan Corporation

Introduction

Background

- Document summarization aims to rephrase a document in a short form called a summary while keeping its "meaning."
  - We aim to characterize the meaning of a document using embeddings or distributed representations of words in the document
- Recent work (Kågbäck et al., 2014) considered a summarization method using embeddings, which maximizes a submodular function defined by the summation of cosine similarities based on sentence embeddings.
  - This method essentially assumes linear meanings, since the objective function is characterized by the summation of sentence-level similarities.

Purpose

- Consider a summarization method based on document-level similarity, where we assume the non-linearity of meanings.

Contributions

- Proved **a cosine similarity of document embeddings is not submodular**.
- Proposed **an objective function based on embedding distributions**, each of which represents a set of word embeddings in a document, and proved
  - It has the **monotone submodularity**.
  - It is **asymptotically related to KL-divergence**.

Preliminaries

Submodularity

- Submodularity is a property of set functions, which is similar to the convexity or concavity of continuous functions.
  - If a set function f is monotone submodular, we can approximate the optimal solution efficiently by a simple greedy algorithm.

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

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

Embedding

- An embedding of a word is a real valued vector in an m-dimensional Euclidean space $\mathbb{R}^m$, which expresses the "meaning" of the word.
  - 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.

Proposed Method

- Focus on a summarization task as sentence selection in a document.
- The optimization framework of our task is formalized in Algorithm 1, which is the same as in the previous study.
- 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).

Algorithm 1: Modified greedy algorithm.

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

```
C ← ∅;  U ← D;
while U ≠ ∅ do
    s* ← argmax_{s∈U} f_C(s)/(w_s)^r;
    if Σ_{s∈C} w_s + w_{s*} ≤ ℓ then C ← C ∪ {s*};
    U ← U \ {s*};
s* ← argmax_{s∈D:w_s≤ℓ} f({s});
return C ← argmax_{C'∈{C,{s*}}} f(C');
```

Similarity Based on Document Embedding

- Define an objective function based on a cosine similarity of document embeddings as follows.

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

  - where $\boldsymbol{v}_D := \sum_{s \in D} \sum_{w \in s} \vec{w}$.
- Next theorem shows that **a solution of f^{Cos} by Algorithm 1 is not guaranteed to be near optimal**.

Theorem 1. $f^{Cos}(x)$ is not submodular.

Similarity Based on Embedding Distributions

- Propose an objective function 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.
- Formalize this idea as **the negative summation of the nearest neighbors' distances on embedding distributions**.

$$f^{NN}(C) := -\sum_{s \in D} \sum_{w \in s} g(N(w, C)) \quad \text{s.t.} \quad N(w, C) := \min_{\substack{v \in s : s \in C \\ \vec{w} \neq \vec{v}}} d(\vec{w}, \vec{v}),$$

  - where g is a non-decreasing scaling function.
- Next theorem shows that **a solution of f^{NN} by Algorithm 1 is guaranteed to be near optimal**.

Theorem 2. $f^{NN}(x)$ is monotone submodular.

- 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**.

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., $D \sim p$, $C_1 \sim q$, $C_2 \sim r$, 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 $D_{KL}(p \parallel r)$ and $D_{KL}(p \parallel q)$, i.e.,

$$\mathbb{E}[f^{NN}(C_2)] - \mathbb{E}[f^{NN}(C_1)] > 0 \Leftrightarrow D_{KL}(p \parallel q) - D_{KL}(p \parallel q) > 0,$$

as $|C_1| \to \infty$, $|C_2| \to \infty$, and $|D| \to \infty$.

Experiment

- Compared ROUGE-N of the following methods:
  - DocEmb: Algorithm1 with f^{Cos},
  - EmbDist: Algorithm1 with f^{NN} s.t. $g(x) = \ln(x), x, e^x$,
  - SemEmb: [Kageback et al. 2014],
  - TfIdf: [Lin and Bilmes, 2011],
  - ApxOpt: Algorithm1 with ROUGE-1 calculated by human references.
- Used Opinosis dataset (Ganesan et al., 2010) also used in the previous study.
  - Collection of user reviews in 51 different topics such as hotels, cars, and products, where 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.

	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

- **EmbDist (e^x) performed the best for ROUGE-1**, which is the best metric in terms of correlation with human summaries.
  - DocEmb also performed better than SenEmb and TfIdf.
  - These results imply that **the document-level similarity can capture more complex meanings than the sentence-level similarity**.
- TfIdf performed the worst for ROUGE-1.
  - This suggests that embedding-based methods naturally have robustness for user-generated content.

Conclusion

- Proposed simple but powerful summarization methods using the document-level similarity based on embeddings.
- Future research includes exploring other scaling functions suitable for our problem or different problems.
  - According to (Cusner et al., 2015), **our function f^{NN} with $g(x) = x$ is the same as a tight lower bound of Earth Mover's Distance (EMD)** developed in the image processing field (Rubner et al., 1998; Rubner et al., 2000).