Document Similarity for Texts of Varying Lengths via Hidden Topics

Hongyu Gong*, Tarek Sakakini*, Suma Bhat*, Jinjun Xiong[+]
*University of Illinois at Urbana-Champaign, [+]IBM Thomas J. Watson Research Center

ACL 2018

center for cognitive computing systems research

IBM | ILLINOIS

Motivation

Measuring the similarity between texts is an important task for many NLP applications. Available approaches to measure document similarity are inadequate for document pairs that have non-comparable lengths, i.e., a long document and its summary.

I've learned ***Newton's Law***, what are relevant projects?

User

Challenges:

- Small lexical overlap: different word usages in documents and summaries;
- Abstraction gaps: one is detailed while the other is concise

Hidden topics are what we use to bridge the gap and measure the document-summary similarity.

Problem Statement

Well, there is a project studying the relation between **gravity and car motion**

Matching System

Matching Text pairs of varying lengths:

- *Science projects – concepts:* recommend relevant projects given technical concepts or assign relevant concepts to science projects;
- *Articles – abstracts:* extract relevant articles given concise abstracts or categorize articles into different conceptual groups.

Model

Hidden topics in a document

- Long document's word vectors: $W = \{w_1, w_2, \dots, w_n : w_i \in \mathbb{R}^d\}$
- Hidden topics are vectors $h_1, h_2, \dots, h_K$, generating each word in document:

$$w_i \approx \sum_{k=1}^{K} \alpha_k^i h_k, \text{ where } \alpha_k^i \text{ is constant}$$

- Extract hidden topics so as to minimize reconstruction error:

$$h_1^*, h_2^*, \dots, h_K^* = \underset{h_1, h_2, \dots, h_K}{\operatorname{argmin}} \min_{\{\alpha_k^i\}} \sum_{i=1}^{n} \left| w_i - \sum_{k=1}^{K} \alpha_k^i h_k \right|$$

- Topic importance: $\bar{\iota}_k$ for the topic h_k^*

Summary reconstruction

- Summary word embeddings: $S = \{s_1, s_2, \dots, s_m\}$
- Relevance of topic h_k^* and the summary S

$$r(h_k^*, S) = \frac{1}{m} \sum_{j=1}^{m} \operatorname{cosSim}(s_j, h_k^* {h_k^*}^T s_j)$$

Document-Summary Relevance

- Weighted topic-summary relevance

$$r(W, S) = \sum_{k=1}^{K} \bar{\iota}_k \cdot r(h_k^*, S)$$

Interpretation of Hidden Topics

Figure 1: Topic words from papers on word sense disambiguation

- Hidden topics: vectors not corresponding to specific words.
- Topic words: interpretation of hidden topics. They are words in the document that can be reconstructed using hidden topics with minimal error.

Experiment: Project-Concept Matching

Task: given a pair of concept and project, decide whether it is a good match.

Dataset: 537 concept-project pairs annotated by UIUC engineering students.

Baselines:

(1) word mover's distance (wmd): embedding-based algorithm in measuring document similarity;

(2) Doc2vec : neural network model to represent documents as vectors. The cosine similarity between document representations is taken as the document similarity.

Table 1: Performance on project-concept Matching

Method	Precision	Recall	F1 score
Our system	**0.7579**	**0.8852**	**0.8155**
wmd	0.6426	0.7353	0.6793
doc2vec	0.6149	0.8432	0.6949

Experiment: Article-Abstract Matching

Task: given each human-generated summary of research topic, we rank 730 ACL papers according to their relevance to the summary. Calculate the precision of top ones.

Dataset: 730 ACL research papers divided into 50 topics, and each category consists of 11 papers on the same topic.

Figure 2: Precision@k for article-abstract matching