Retrieving Compositional Documents Using Position-Sensitive Word Mover's Distance

Martin Trapp, Marcin Skowron, Dietmar Schabus

[1]Austrian Research Institute for Artificial Intelligence, Vienna, Austria
[2] SPSC Lab., Graz University of Technology, Graz, Austria
[3] Dept. of Computational Perception, Johannes Kepler University Linz, Linz, Austria

Overview

Motivation

Retrieving of compositional documents which consist of ranked sub-documents, e.g. threads of healthcare fora, news articles

Current methods do not leverage the effective generalization properties of semantic embeddings

Propose a semantic document distance which takes the characteristics of compositional documents into account

Approach

We present an extension of the Word Mover's Distance that takes the position and presentation bias, present in compositional documents, into account

We introduce a novel benchmark dataset for the task of retrieving compositional datasets

We show that incorporating semantic relation between words and sensitivity to the position and presentation bias improves retrieval performance

Method

Word Mover's Distance [1]

Minimal cost required to "transport" words from one document to another, where the cost is influenced by the semantic difference of the words

Position-Sensitive Word Mover's Distance

Incorporates position and presentation bias of sub-documents into the transportation problem

$$b_p = \left(\frac{1}{1 + r_p}\right)^{\gamma} \qquad \hat{d}_i = \frac{1}{z} \sum_{p=1}^{P} b_p c_{pi}$$

$$\underset{\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{N \times N}}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{N} T_{ij} ||x_i - x_j||_2$$

$$\text{subject to} \sum_{j=1}^{N} T_{ij} = \hat{d}_i, \ \sum_{i=1}^{N} T_{ij} = \hat{d}'_j \ \forall i, j$$

Twin Film Dataset

Retrieval of films which have the same or very similar plot but were produced by two different studios around the same time

Dataset consists of 111 twin film pairs with 221 unique films, extracted from Wikipedia

Films range over several genres and cover a wide range of production dates (1938 - 2016)

Table 1: Examples from twin films dataset.

First Film	Second Film
Oscar Wilde (1960)	The Trials of Oscar Wilde (1960)
Prefontaine (1997)	Without Limits (1998)
Kundun (1997)	Seven Years in Tibet (1997)
A Hijacking (2012)	Captain Phillips (2013)

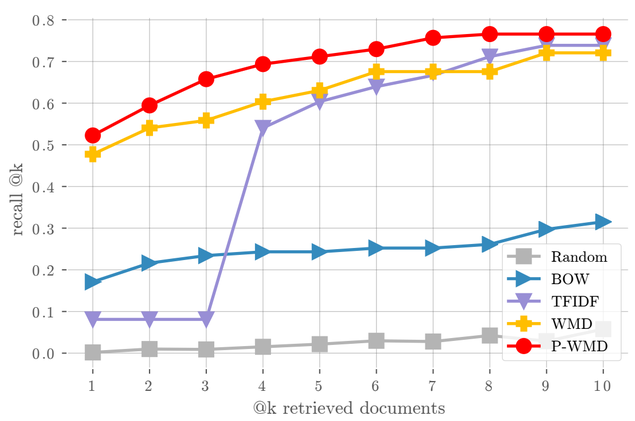
IMDB plot keywords and their position (number of up-votes) is used to compare films

Table 1: Top five plot keywords of examples from twin films dataset.

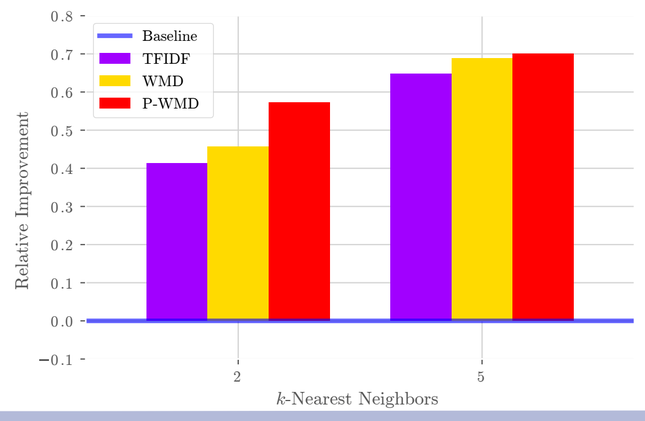
Film	Plot Keywords (Top 5)
Oscar Wilde (1960)	homosexual-history, grapes, playwright, grape, london-fog
The Trials of Oscar Wilde (1960)	gay-husband, gay-interest, homosexuality, homosexual, gay
Prefontaine (1997)	oregon, long-distance-runner, runner, olympics, watching-television
Without Limits (1998)	oregon, car-crash, death, university-of-oregon, coach
Kundun (1997)	tibet, chinese, dalai-lama, lama, tibetan
Seven Years in Tibet (1997)	dalai-lama, tibet, austria, mountain, himalaya
A Hijacking (2012)	somali-pirate, pirate, cargo-ship, ransom, ceo
Captain Phillips (2013)	ship, hostage, lifeboat, somalian-pirate, leader

Experiments

Twin Film Dataset



MovieLense [2] - Genre Detection



ACM International Conference on the Theory of Information Retrieval
October 1-4 2017
Amsterdam, Netherlands

[1] M. J Kusner, Y. Sun, N. I Kolkin, and K. Q Weinberger. 2015. From Word Embeddings To Document Distances.. In Proceedings of ICML, Vol. 15. 957–966.
[2] F. M Harper and J. A Konstan. 2016. The Movielens Satasets: History and Context. ACM TiiS 5, 4 (2016), 19.

Austrian Research Institute Artificial Intelligence
OFAI