an_2018_model_free_context_aware_word_composition

Year

2018

Author(s)

An, Bo and Han, Xianpei and Sun, Le

Title

Model-Free Context-Aware Word Composition

Venue

COLING

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

(Presumed) manual labeling

Topic labeling parameters

\

Label generation

(Presumed) manual labeling on the 50 topics

Motivation

\

Topic modeling

GibbsLDA++ (Phan and Nguyen, 2007)

Topic modeling parameters

α = 0.5

β = 0.1

Number of topics: 50

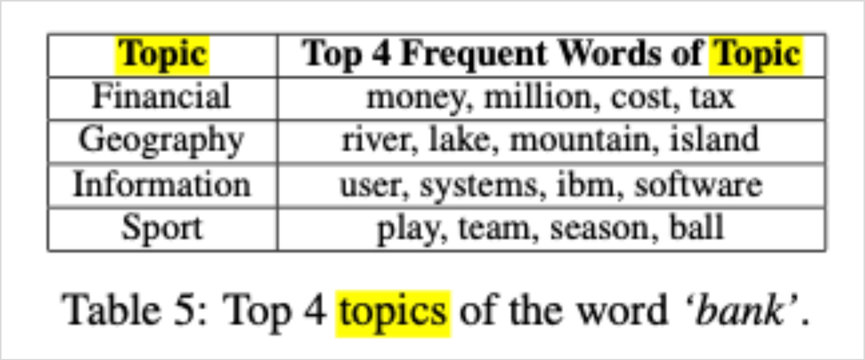
Number of iterations: 400

Nr. of topics

50

Label

Single-word label of the 50 generated topics



Topic	Top 4 Frequent Words of Topic
Financial	money, million, cost, tax
Geography	river, lake, mountain, island
Information	user, systems, ibm, software
Sport	play, team, season, ball

Table 5: Top 4 topics of the word '*bank*'.

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Domain (paper): Word composition
Domain (corpus): Miscellaneous

Problem statement

Proposing a model-free context-aware word composition model, which employs the latent semantic information as global context for learning representations.
The proposed model attempts to resolve the word sense disambiguation and word composition in a unified framework.

(Inspired by Topical Word Embeddings (TWE)) this paper utilises topic distribution as the global context of a linguistic unit.
Each topic is utilized to derive accurate meanings for all the word occurrences in the linguistic unit and to learn the topic-specific representation of the unit.
After that, the context-aware representation of the linguistic unit is inferred by summarizing all its representations under different topics based on the topic distribution.
In this way, the method can make use of the topic information of a word to learn its accurate topic-specific representation, and the topic distribution of the a unit is employed as a cue to guide the process of word composition to learn meaningful representation.

Corpus

Origin: British National Corpus (BNC)
Content: 93 million terms
Details: Samples of written language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century,

Origin: Wikipedia
Content: 990 million tokens
Details: Snapshot of the English Wikipedia corpus

Document

British National Corpus (BNC)
The **written part** of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text.

Pre-processing

No mention of pre-processing steps

```
@inproceedings{an_2018_model_free_context_aware_word_composition,
    title = "Model-Free Context-Aware Word Composition",
    author = "An, Bo  and
      Han, Xianpei  and
      Sun, Le",
    booktitle = "Proceedings of the 27th International Conference on
Computational Linguistics",
    month = aug,
    year = "2018",
    address = "Santa Fe, New Mexico, USA",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/C18-1240",
    pages = "2834--2845",
    abstract = "Word composition is a promising technique for representation
learning of large linguistic units (e.g., phrases, sentences and documents).
However, most of the current composition models do not take the ambiguity of
words and the context outside of a linguistic unit into consideration for
learning representations, and consequently suffer from the inaccurate
representation of semantics. To address this issue, we propose a model-free
context-aware word composition model, which employs the latent semantic
information as global context for learning representations. The proposed model
attempts to resolve the word sense disambiguation and word composition in a
unified framework. Extensive evaluation shows consistent improvements over
various strong word representation/composition models at different
granularities (including word, phrase and sentence), demonstrating the
effectiveness of our proposed method.",
}
```

#Thesis/Papers/Initial