# huang_2018_phrasectm_correlated_topic_modeling_on_phrases_within_markov_random_fields

## Year

2018

## Author(s)

Huang, Weijing

## Title

PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields

## Venue

ACL

---

## Topic labeling

Manual

## Focus

Secondary

## Type of contribution

Established approach

## Underlying technique

Manual labeling

## Topic labeling parameters

\

## Label generation

We outputted the topics with top-10 words/phrases in each topic and the correlation of topics for 10 human annotators, and asked them to label the topics.

The annotators in Group A got the CTM result on Maths@Wiki and PhraseCTM's on Argentina@Wiki.
The annotators in Group B got the results in the opposite setting.

The labeling process was logged to calculate the accumulated time.
The labeling time to reach 50 accurate topics' labels on PhraseCTM is much less than the labeling time on CTM. In average, the annotators spent 7.1 minutes on PhraseCTM while 13.2 minutes on the others.

| | CTM | | PhraseCTM | |
|---|---|---|---|---|
| | Maths | Argentina | Maths | Argentina |
| Group A | 12.4 | - | - | 7.5 |
| Group B | - | 14.0 | 6.7 | - |
| In Average | 13.2 | | 7.1 | |

Table 2: Human time consumption on topic labeling for correlated topics generated by CTM and PhraseCTM, measured in minutes.

## Motivation

Using the labeling procedure as a proxy to measure the interpretability of the generated topics among the two models (CTM and PhraseCTM).
Using the time to label as a metric in this comparison.

---

## Topic modeling

(Plain) Correlated Topic Model (CTM) (Blei and Lafferty, 2005) and PhraseCTM

## Topic modeling parameters

Nr. of topics (k): 100

## Nr. of topics

100

## Label

Two (one per annotator's group) manually assigned single or multi-word labels

## Label selection

\

## Label quality evaluation

\

## Assessors

\

---

## Domain

Domain (paper): Phrase-level topic modeling
Domain (corpus): News, Mathematics, Chemistry, papers in the medical domain

## Problem statement

Recent emerged phrase-level topic models are able to provide topics of phrases (each topic is represented as a list of phrases, which are easy to read for humans).

For example, the topic represented in "grounding con- ductor, grounding wire, aluminum wiring, neutral ground, …" is easier to read than the topic with words "ground, wire, use, power, cable, wires, …", although they are both about the topic of house- hold electricity.

But these models are lack of the ability to capture the correlation structure among the discovered numerous topics.
We propose a novel topic model PhraseCTM and a two-stage method to find out the correlated topics at phrase level.
In the first stage, we train PhraseCTM, which models the generation of words and phrases simultaneously by linking the phrases and component words within Markov Random Fields when they are semantically coherent.

In the second stage, we generate the correlayion of topics from PhraseCTM.

## Corpus

Origin: 20Newsgroups
Nr. of documents:
Details:

Origin: Wikipedia
Nr. of documents:
Details: Mathematics, Chemistry, and Argentina subsets of English Wikipedia

Origin: PubMed
Nr. of documents:
Details:  Subset of PubMed Abstracts

## Document

## Pre-processing

For each corpus, we extract the phrases using AutoPhrase

---

```
@inproceedings{huang_2018_phrasectm_correlated_topic_modeling_on_phrases_within
_markov_random_fields,
    title = "{P}hrase{CTM}: Correlated Topic Modeling on Phrases within {M}
arkov Random Fields",
    author = "Huang, Weijing",
    booktitle = "Proceedings of the 56th Annual Meeting of the Association for
Computational Linguistics (Volume 2: Short Papers)",
    month = jul,
    year = "2018",
    address = "Melbourne, Australia",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/P18-2083",
    doi = "10.18653/v1/P18-2083",
```

```
    pages = "521--526",
    abstract = "Recent emerged phrase-level topic models are able to provide topics of phrases, which are easy to read for humans. But these models are lack of the ability to capture the correlation structure among the discovered numerous topics. We propose a novel topic model PhraseCTM and a two-stage method to find out the correlated topics at phrase level. In the first stage, we train PhraseCTM, which models the generation of words and phrases simultaneously by linking the phrases and component words within Markov Random Fields when they are semantically coherent. In the second stage, we generate the correlation of topics from PhraseCTM. We evaluate our method by a quantitative experiment and a human study, showing the correlated topic modeling on phrases is a good and practical way to interpret the underlying themes of a corpus.",
}
```

#Thesis/Papers/Initial