Latent Entity Clustering Using Entity Oriented Topic Models

Adam Slack*
Nottingham Trent University

Caroline S. Langensiepen **
Nottingham Trent University

We present a novel approach to extracting latent semantic relationships between entities across a corpus, by considering how the context surrounding an entity can be retained in bag-of-word statistical models, particularly, when a corpus of longer written narratives is used. Latent Entities, or entity stereotypes are extracted by applying clustering methods to entity-oriented topic models which are derived from a Gaussian-filtered entity-term co-occurrence matrix. Using this approach, it is demonstrated that the proposed methods can provide semantically meaningful latent entity models, which in turn can be interpreted by a human. The use of a modified co-occurrence matrix allows for the production of entity topic models with distinguishing topic distributions at an intra-document and inter-document level, highlighting the model's ability to capture surrounding contextual similarities and differences between entities within a document and across a corpus. The quality of resulting models are explored using a blend of qualitative and quantitative methods, using consistency metrics to compare models extracted from two corpora, as well as visualisations exploring the qualitative aspects of extracted models.

1. Introduction

This document applies to version 3 of CL class file. Prior style files such as "`cl.sty`" and "`coli.sty`" do not have all of the features described here. It is assumed that the user has a basic knowledge of LaTeX typesetting commands.

2. Class File Options

There are several options that can be used to switch the mode of MIT2 from normal article to manuscript style, or to different layout styles. This is specified in the usual LaTeX way by declaring:

```
\documentclass[bookreview,manuscript]{clv3}
```

bookreview: Sets the article layout for Book Review.

brief: Sets the article layout for Briefly Noted.

discussion: Sets the article layout for Squibs and Discussions.

pubrec: Sets the article layout for Publication Received.

shortpaper: Sets the article layout for Short Paper.

* School of Science and Technology, Nottingham Trent University. Email: adam.slack2013@ntu.ac.uk
** School of Science and Technology, Nottingham Trent University. Email: caroline.langensiepen@ntu.ac.uk

© 2018 Association for Computational Linguistics

manuscript: Sets the baseline spacing to double space. This option can be used in combination with other options.

By not declaring any option in the `\documentclass` command the class file will automatically set to standard article layout.

3. Title Page

The title page is created using the standard LaTeX command `\maketitle`. Before this command is declared, the author must declare all the data which are to appear in the title area.[1]

3.1 Volume, Number and Year

The command `\issue{vv}{nn}{yyyy}` is used in declaring the volume, number and year of the article. The first argument is for the volume, the second argument is for the issue number. Volume and Issue number will appear on the even page running head opposite the journal name. The third argument is for the Year which will appear in the copyright line at the bottom of the title page.

3.2 Document Head

Document head is produced with the command `\dochead{Document Head}`. Doc head will output differently, or may not appear at all, depending on the option used in the documentclass.

3.3 Paper Title

The paper title is declared like: `\title{Computer Linguistic Article}` in the usual LaTeXmanner. Line breaks may be inserted with (`\\`) to equalize the length of the title lines.

3.4 Authors

The name and related information for authors is declared with the `\author{}` command.

The `\thanks{}` command produces the "first footnotes.". LaTeX `\thanks` cannot accommodate multiple paragraphs, author will have to use a separate `\thanks` for each paragraph.

The `\affil{}` command produces the author affiliations that appears right under the author's name.

3.5 Running Headers

The running heads are declared with the `\runningtitle{Running Title}` for the journal name and `\runningauthor{Author's Surname}` for author. These information will appear on the odd pages. For `bookreview` option, odd page running

1 \maketitle is the command to execute all the title page information.

head is automatically set to "Book Reviews". Even page running head is default to Computational Linguistics opposite volume and issue number.

3.6 History Dates

History dates are declared with `\historydates{Submission received:...}`. This data should contain Submission, Revised and Accepted date of the article. History dates appear at the footnote area of title age.

4. Abstract

Abstract is the first part of a paper after `\maketitle`. Abstract text is placed within the abstract environment:

```
\begin{abstract}
This is the abstract text . . .
\end{abstract}
```

5. Section Headings

Section headings are declared in the usual LaTeX way via `\section{}`, `\subsection{}`, `\subsubsection{}`, and `\paragraph{}`. The first 3 levels of section head will have Arabic numbering separated by period. The `\paragraph{}` section will have the title head in Italics and at the same line with the first line of succeeding paragraph.

6. Citations

Citations in parentheses are declared using the `\cite{}` command, and appear in the text as follows: This technique is widely used (**?**). The command `\citep{}` (cite parenthetical) is a synonym of `\cite{}`.

Citations used in the sentence are declared using the `\namecite{}` commands, and appear in the text as follows: **?** first described this technique. The command `\citet{}` (cite textual) is a synonym of `\namecite{}`.

This style file is designed to be used with the BibTeX style file `compling.bst`. Include the command `\bibliographystyle{compling}` in your source file.

Citation commands are based on the `natbib` package; for details on options and further variants of the commands, see the `natbib` documentation. In particular, options exist to add extra text and page numbers. For example, `\cite[cf.][ch.\ 1]{winograd}` yields: (cf. **?**, ch. 1).

The following examples illustrate how citations appear both in the text and in the references section at the end of this document.

1. Article in journal: **?**; **?**.
2. Book: **?**; **?**.
3. Article in edited collection/Chapter in book: **?**; **?**; **?**.
4. Technical report: **?**; **?**.
5. Thesis or dissertation: **?**; **?**; **?**.

6. Unpublished item: **?**.
7. Conference proceedings: **?**.
8. Paper published in conference proceedings: **?**; **?**.

7. Definition with Head

Definition with head is declared by using the environment:

```
\begin{definition}
Definition text. . .
\end{definition}
```

This environment will generate the word **"Definition 1"** in bold on separate line. The sequence number is generated for every definition environment. Definition data will have no indention on the first line while succeeding lines will have hang indention.

8. Lists

The usual LaTeX itemize, enumerate and definition list environments are used in CLV3 style.

To produce Numbered List use the environment:

```
\begin{enumerate}
\item First numbered list item
\item Second numbered list item
\item Third numbered list item
\end{enumerate}
```

To produce Bulleted List use the environment:

```
\begin{itemize}
\item First bulleted list item
\item Second bulleted list item
\item Third bulleted list item
\end{itemize}
```

To produce Definition List use the environment:

```
\begin{deflist}
\item[First]  Definition list item. . .
\item[Second] Definition list item. . .
\item[Third]  Definition list item. . .
\end{deflist}
```

Additional list environment were also defined such as Unnumbered, Arabic and Alpha lists.

Unnumbered List is the list where item labels are not generated. To produce Unnumbered List use the environment:

```
\begin{unenumerate}
\item First list item
\item Second list item
```

```
\item Third list item
\end{unenumerate}
```

To produce Arabic List use the environment:

```
\begin{arabiclist}
\item First arabic list item
\item Second arabic list item
\item Third arabic list item
\end{arabiclist}
```

To produce Alpha List use the environment:

```
\begin{alphalist}
\item First alpha list item
\item Second alpha list item
\item Third alpha list item
\end{alphalist}
```

All the list environments mentioned above can be nested with each other.

8.1 Other List Types
8.1.1 Outline List or Example List.

```
\begin{exlist}
\item First outline list item. . .
\item Second outline list item. . .
\item Third outline list item. . .
\end{exlist}
```

8.1.2 Output Formula or Algorithm.

```
\begin{algorithm}
\item[Step 1] First item. . .
\item[Step 2] Second item. . .
\end{algorithm}
```

See sample on the `COLI-template.pdf`.

9. Word Formula or Displayed Text

Word formula and displayed text are text that should be displayed in a separate line without indention. This are achieved by using the environment:

```
\begin{displaytext}
This is a sample of displayed text . . .
\end{displaytext}
```

10. Dialogue

Dialogue text are presentation of people's conversation. These will be presented on a separate line where each dialogue starts with the name of speaker, followed by colon. Succeeding lines will be hang indented. To produce Dialogue use the environment:

```
\begin{dialogue}
Speaker 1: dialogue. . .

Speaker 2: dialogue. . .
\end{dialogue}
```

Please make sure to insert an empty line between dialogues.

11. Extracts

Extract text acts like quote, where left and right margins are indented. To produce Extract use the environment:

```
\begin{extract}
This is an example of Extract text. . .
\end{extract}
```

See sample on the `COLI-template.pdf`.

12. Theorem-like Environments

There are several theorem-like environments defined in CLV3 class file. Theorem-like environments generate the name of the theorem as label, and counter number in bold.

12.1 Example

To produce Example use the environment:

```
\begin{example}
This is Example text. . .
\end{example}
```

12.2 Lemma

To produce Lemma use the environment:

```
\begin{lemma}
Lemma text. . .
\end{lemma}
```

This produces the following output:

Lemma 1
Lemma text.

A small vertical space separates the end of the lemma from the following text.

12.3 Theorem

To produce Theorem use the environment:

```
\begin{theorem}
Theorem text. . .
\end{theorem}
```

15. Others

Other items such as Equations, Figures, Tables and References are produced in the standard LaTeX typesetting.