Neural Architectures and Evaluation Protocols for Open Information Extraction

Thesis submitted by

Samarth Aggarwal
2016CS10395

under the guidance of

Prof. Mausam

in partial fulfilment of the requirements
for the award of the degree of

Bachelor of Technology

Department Of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY DELHI

July 2020

THESIS CERTIFICATE

This is to certify that the thesis titled **Neural Architectures and Evaluation Protocols for Open Information Extraction**, submitted by **Samarth Aggarwal**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Bachelor of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Mausam
Professor
Dept. of Physics
IIT-Delhi, 110 016

Place: New Delhi
Date: 10th July 2020

ACKNOWLEDGEMENTS

TO BE ADDED

I thank IIT Delhi HPC Facility for compute resources.

ABSTRACT

Open Information Extraction refers to the task of obtaining relation tuples from a sentence. For eg. the sentence “Donald Trump is the president of United States.” yields (Donald Trump ; is the president of ; United States) as its OpenIE tuple.

The Open IE paradigm is a useful intermediary for a variety of down-stream tasks such as sentence similarity, event schema induction, text comprehension, knowledge base completion, and more. There have been several attempts at building OpenIE systems that explored rule-based such as OllIE, OpenIE-4 and OpenIE-5. Another wave of OpenIE systems that followed, comprised of neural approaches such as RnnOIE and Cui et al. (2018). However, the existing openie systems suffer from a wide range of problems. The rule-based systems suffered from cascading errors from a large number of components in succession. The existing neural OpenIE systems, although were able to solve some of these issues to a certain extent, were still far from ideal. Infact, they introduced other problems such as redundancy in their outputs. Together these factors solicit an OpenIE system that is able to overcome the issues pertaining to OpenIE.

Although human inspection revealed that the existing systems were not upto the mark, yet these systems scored high on the existing state-of-the-art OpenIE benchmarks such as OIE2016 (Stanovsky and Dagan, 2016). This means that the existing benchmarks do not correlate well with how humans evaluate OpenIE. In response, we contribute CaRB (Bhardwaj et al., 2019), with a high-quality crowdsourced gold dataset and intuitive evaluation policies that correlate well with human judgement of OpenIE. CaRB establishes itself as the new state of the art OpenIE benchmark.

CaRB evaluation of the Cui et al. (2018), then state of the art OpenIE systems, confirms its inept performance. We contribute IMoJIE (Kolluru et al., 2020), a neural OpenIE model that outperforms the previous state of the art by about 18 F1 points. It reduces the redundancy in output extractions significantly. Along with it, IMoJIE also presents a novel approach that can be used to generation high-quality training data from multiple low quality datasets.

Although IMoJIE improves the quality of OpenIE tuples significantly, this improvement comes at the cost of speed of extraction. We design a MLIL architecture to overcome the issue of speed of extraction and also obtain further performance nudges from it. This approach also yields a coordination analyzer that significantly improves the yield of the MLIL model.

In the end, we analyse the milestones covered in the world of OpenIE and contribute some ideas for future research.

Contents

List of Tables

List of Figures

ABBREVIATIONS

IITD	Indian Institute of Technology, Delhi
RTFM	Read the Fine Manual

NOTATION

r	Radius, m
α	Angle of thesis in degrees
β	Flight path in degrees

Chapter 1

Sample Chapter

This document provides a simple template of how the provided iitddiss.cls LaTeX class is to be used. Also provided are several useful tips to do various things that might be of use when you write your thesis.

To compile your sources run the following from the command line:

```
% pdflatex thesis.tex
% bibtex thesis
% pdflatex thesis.tex
% pdflatex thesis.tex
```

Modify this suitably for your sources.

To generate PDF's with the links from the hyperref package use the following command:

```
% dvipdfm -o thesis.pdf thesis.dvi
```

1.1 Package Options

Use this thesis as a basic template to format your thesis. The iitddiss class can be used by simply using something like this:

```
\documentclass[PhD]{iitddiss}
```

To change the title page for different degrees just change the option from PhD to one of MS, MTech or BTech. The dual degree pages are not supported yet but should be quite easy to add. The title page formatting really depends on how large or small your thesis title is. Consequently it might require some hand tuning. Edit your version of iitddiss.cls suitably to do this. I recommend that this be done once your title is final.

To write a synopsis simply use the synopsis.tex file as a simple template. The synopsis option turns this on and can be used as shown below.

```
\documentclass[PhD,synopsis]{iitddiss}
```

Once again the title page may require some small amount of fine tuning. This is again easily done by editing the class file.

This sample file uses the `hyperref` package that makes all labels and references clickable in both the generated DVI and PDF files. These are very useful when reading the document online and do not affect the output when the files are printed.

1.2 Example Figures and tables

Fig. 2.1 shows a simple figure for illustration along with a long caption. The formatting of the caption text is automatically single spaced and indented. Table 2.1 shows a sample table with the caption placed correctly. The caption for this should always be placed before the table as shown in the example.

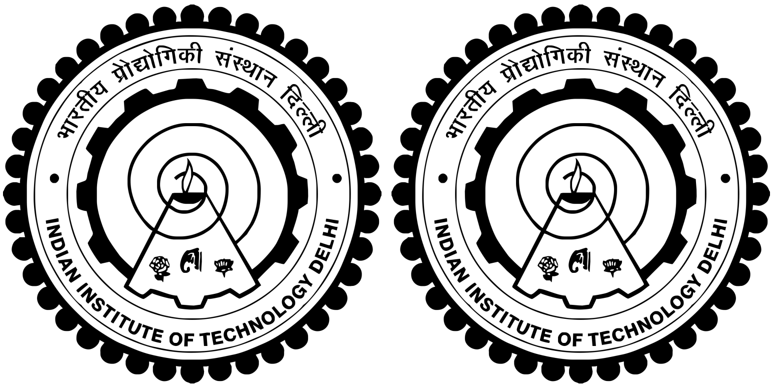
भारतीय प्रौद्योगिकी संस्थान दिल्ली
INDIAN INSTITUTE OF TECHNOLOGY DELHI
भारतीय प्रौद्योगिकी संस्थान दिल्ली
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Figure 1.1: Two IITD logos in a row. This is also an illustration of a very long figure caption that wraps around two two lines. Notice that the caption is single-spaced.

Table 1.1: A sample table with a table caption placed appropriately. This caption is also very long and is single-spaced. Also notice how the text is aligned.

x	x^2
1	1
2	4
3	9
4	16
5	25
6	36
7	49
8	64

1.3 Bibliography with BIBTEX

I strongly recommend that you use BIBTEX to automatically generate your bibliography. It makes managing your references much easier. It is an excellent way to organize your references and reuse them. You can use one set of entries for your references and cite them in your thesis, papers and reports. If you haven't used it anytime before please invest some time learning how to use it.

I've included a simple example BIBTEX file along in this directory called `refs.bib`. The `iitddiss.cls` class package which is used in this thesis and for the synopsis uses the `natbib` package to format the references along with a customized bibliography style provided as the `iitd.bst` file in the directory containing `thesis.tex`. Documentation for the `natbib` package should be available in your distribution of LATEX. Basically, to cite the author along with the author name and year use `\cite{key}` where `key` is the citation key for your bibliography entry. You can also use `\citet{key}` to get the same effect. To make the citation without the author name in the main text but inside the parenthesis use `\citep{key}`. The following paragraph shows how citations can be used in text effectively.

More information on BIBTEX is available in the book by Lamport (1986). There are many references (Lamport, 1986; K, 2016) that explain how to use BIBTEX. Read the `natbib` package documentation for more details on how to cite things differently.

Here are other references for example. Ramachandran (2001) presents a Python based visualization system called MayaVi in a conference paper. Ramachandran et al. (2003) illustrates a journal article with multiple authors. Python (van Rossum et al., 1991–) is a programming language and is cited here to show how to cite something that is best identified with a URL.

1.4 Other useful LATEX packages

The following packages might be useful when writing your thesis.

- It is very useful to include line numbers in your document. That way, it is very easy for people to suggest corrections to your text. I recommend the use of the `lineno` package for this purpose. This is not a standard package but can be obtained on the internet. The directory containing this file should contain a lineno directory that includes the package along with documentation for it.
- The `listings` package should be available with your distribution of LATEX. This package is very useful when one needs to list source code or pseudo-code.
- For special figure captions the `ccaption` package may be useful. This is specially useful if one has a figure that spans more than two pages and you need to use the same figure number.

- The notation page can be entered manually or automatically generated using the `nomencl` package.

More details on how to use these specific packages are available along with the documentation of the respective packages.

Chapter 2

Introduction

This document provides a simple template of how the provided `iitddiss.cls` LaTeX class is to be used. Also provided are several useful tips to do various things that might be of use when you write your thesis.

To compile your sources run the following from the command line:

```
% pdflatex thesis.tex
% bibtex thesis
% pdflatex thesis.tex
% pdflatex thesis.tex
```

Modify this suitably for your sources.

To generate PDF's with the links from the `hyperref` package use the following command:

```
% dvipdfm -o thesis.pdf thesis.dvi
```

2.1 Package Options

Use this thesis as a basic template to format your thesis. The `iitddiss` class can be used by simply using something like this:

```
\documentclass[PhD]{iitddiss}
```

To change the title page for different degrees just change the option from `PhD` to one of `MS`, `MTech` or `BTech`. The dual degree pages are not supported yet but should be quite easy to add. The title page formatting really depends on how large or small your thesis title is. Consequently it might require some hand tuning. Edit your version of `iitddiss.cls` suitably to do this. I recommend that this be done once your title is final.

To write a synopsis simply use the `synopsis.tex` file as a simple template. The synopsis option turns this on and can be used as shown below.

```
\documentclass[PhD,synopsis]{iitddiss}
```

Once again the title page may require some small amount of fine tuning. This is again easily done by editing the class file.

This sample file uses the `hyperref` package that makes all labels and references clickable in both the generated DVI and PDF files. These are very useful when reading the document online and do not affect the output when the files are printed.

2.2 Example Figures and tables

Fig. 2.1 shows a simple figure for illustration along with a long caption. The formatting of the caption text is automatically single spaced and indented. Table 2.1 shows a sample table with the caption placed correctly. The caption for this should always be placed before the table as shown in the example.

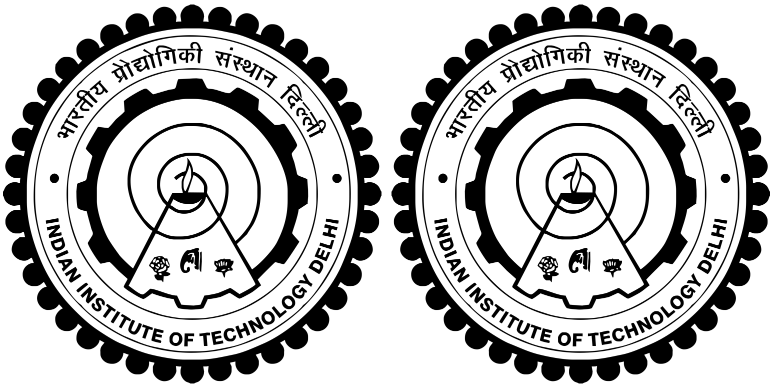
भारतीय प्रौद्योगिकी संस्थान दिल्ली
INDIAN INSTITUTE OF TECHNOLOGY DELHI
भारतीय प्रौद्योगिकी संस्थान दिल्ली
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Figure 2.1: Two IITD logos in a row. This is also an illustration of a very long figure caption that wraps around two two lines. Notice that the caption is single-spaced.

Table 2.1: A sample table with a table caption placed appropriately. This caption is also very long and is single-spaced. Also notice how the text is aligned.

x	x^2
1	1
2	4
3	9
4	16
5	25
6	36
7	49
8	64

2.3 Bibliography with BIBTEX

I strongly recommend that you use BIBTEX to automatically generate your bibliography. It makes managing your references much easier. It is an excellent way to organize your references and reuse them. You can use one set of entries for your references and cite them in your thesis, papers and reports. If you haven't used it anytime before please invest some time learning how to use it.

I've included a simple example BIBTEX file along in this directory called `refs.bib`. The `iitddiss.cls` class package which is used in this thesis and for the synopsis uses the `natbib` package to format the references along with a customized bibliography style provided as the `iitd.bst` file in the directory containing `thesis.tex`. Documentation for the `natbib` package should be available in your distribution of LATEX. Basically, to cite the author along with the author name and year use `\cite{key}` where `key` is the citation key for your bibliography entry. You can also use `\citet{key}` to get the same effect. To make the citation without the author name in the main text but inside the parenthesis use `\citep{key}`. The following paragraph shows how citations can be used in text effectively.

More information on BIBTEX is available in the book by Lamport (1986). There are many references (Lamport, 1986; K, 2016) that explain how to use BIBTEX. Read the `natbib` package documentation for more details on how to cite things differently.

Here are other references for example. Ramachandran (2001) presents a Python based visualization system called MayaVi in a conference paper. Ramachandran et al. (2003) illustrates a journal article with multiple authors. Python (van Rossum et al., 1991–) is a programming language and is cited here to show how to cite something that is best identified with a URL.

2.4 Other useful LATEX packages

The following packages might be useful when writing your thesis.

- It is very useful to include line numbers in your document. That way, it is very easy for people to suggest corrections to your text. I recommend the use of the `lineno` package for this purpose. This is not a standard package but can be obtained on the internet. The directory containing this file should contain a lineno directory that includes the package along with documentation for it.
- The `listings` package should be available with your distribution of LATEX. This package is very useful when one needs to list source code or pseudo-code.
- For special figure captions the `ccaption` package may be useful. This is specially useful if one has a figure that spans more than two pages and you need to use the same figure number.

- The notation page can be entered manually or automatically generated using the `nomencl` package.

More details on how to use these specific packages are available along with the documentation of the respective packages.

Chapter 3

Literature Survey

Chapter 4

CaRB - A Crowdsourced Benchmark for Open IE

Chapter 5

IMoJIE - Iterative Memory Based Joint Open IE

Chapter 6

Remaining Problems

Chapter 7

Conjunction Splitting

Chapter 8

MLIL - Multi Level Iterative Labelling

Chapter 9

Milestones of OpenIE

Chapter 10

Future Ideas

Appendix A

A SAMPLE APPENDIX

Just put in text as you would into any chapter with sections and whatnot. Thats the end of it.

Bibliography

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam. 2019. CaRB: A Crowdsourced Benchmark for OpenIE. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019*, pages 6263–6268.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. In *Proceedings of Association for Computational Linguistics (ACL), 2018*, pages 407–413.

Sai Praneeth Reddy K. 2016. *LaTeX class for dissertations submitted to IIT-D*. Ph.D. thesis, Department of Computer Science and Engineering, IIT-Delhi, New Delhi – 110016.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020. IMoJIE: Iterative Memory-Based Joint Open Information Extraction. In *The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Seattle, U.S.A.

Leslie Lamport. 1986. *LaTeX: A document preparation system*. Addision-Wesley.

Prabhu Ramachandran. 2001. MayaVi: A free tool for CFD data visualization. In *4th Annual CFD Symposium*. Aeronautical Society of India. Software available at: http://mayavi.sf.net.

Prabhu Ramachandran, S. C. Rajan, and M. Ramakrishna. 2003. A fast, two-dimensional panel method. *SIAM Journal on Scientific Computing*, 24(6):1864–1878.

Guido van Rossum et al. 1991–. The Python programming language.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page (to appear), Austin, Texas. Association for Computational Linguistics.

LIST OF PAPERS BASED ON THESIS

1. Authors.... Title... *Journal*, Volume, Page, (year).
2. Authors.... Title... *Journal*, Volume, Page, (year).