PhD Thesis

Koustuv Sinha

Acknowledgements

Abstract

Abstract in French

Contributions to Original Knowledge

Contributions of Authors

List of Figures

List of Tables

Contents

Part I

Introduction

Central Theme of the thesis : Understanding systematicity in pre-trained language models through semantic and syntactic generalization.

Part II

Background

Chapter 1

Early methods for text representation

Chapter 2

Neural Inductive bias of text representation

2.1 Feed Forward Neural Networks

2.2 Recurrent Neural Networks

2.3 Transformer Models

Large Language Models (LLMs) are the state-of-the-art in language models, which are based on Transformers.

Chapter 3

Pre-training and the advent of Large Language Models

Success of pre-training and scale

Chapter 4

Systematicity and Generalization

4.1 Definitions

1. Productivity

2. Word Order Sensitivity

4.2 Tasks

Part III

Understanding semantic generalization through productivity

Chapter 5

Technical Background

Chapter 6

CLUTRR: A Diagnostic Benchmark for Inductive Reasoning in Text

Paper: [1]

6.1 Dataset construction

6.2 Productivity and reasoning

Chapter 7

Results

Chapter 8

Discussion

Chapter 9

Follow-up findings in the community

Chapter 10

Related Work

Part IV

Quantifying syntactic generalization using word order

Paper [2]

Chapter 11

Technical Background

Chapter 12

Word Order in Natural Language Inference

12.1 Probe Construction

Chapter 13

Experiments & Results

Chapter 14

Discussion

Chapter 15

Follow-up findings in the community

Chapter 16

Related Work

Part V

Probing syntax understanding through distributional hypothesis

Paper: [3]

Chapter 17

Technical Background

Chapter 18

Dataset construction and pre-training

Chapter 19

Experiments

19.1 Downstream reasoning tasks

19.2 Evaluating the effectiveness of probing syntax

Chapter 20

Discussion

Chapter 21

Follow-up findings in the community

Chapter 22

Related Work

Part VI

Measuring systematic generalization by exploiting absolute positions

Chapter 23

Technical Background

Chapter 24

Systematic understanding of absolute position embeddings

Chapter 25

Experiments

Chapter 26

Discussion

Chapter 27

Related Work

Part VII

Conclusion

Chapter 28

Summary

Chapter 29

Limitations

Chapter 30

Future Work

Part VIII

Bibliography

Bibliography

[1] Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. In *Empirical Methods in Natural Language Processing (EMNLP) 2019*, September 2019.

[2] Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. UnNatural Language Inference. In *Association for Computational Linguistics (ACL) 2021*, June 2021.

[3] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. In *Empirical Methods in Natural Language Processing (EMNLP)*, April 2021.

Glossary

Transformers A class of models first derived by Vaswani et al. 2017. 15

Acronyms

LLMs Large Language Models. 15

Part IX

Appendix

Chapter 31

Org mode auto save

Run the following snippet to auto save and compile in org mode.

```
(defun kdm/org-save-and-export ()
(interactive)
(if (and (eq major-mode 'org-mode)
    (ido-local-file-exists-p (concat (file-name-sans-extension (buffer-name
  (org-latex-export-to-pdf)))

(add-hook 'after-save-hook 'kdm/org-save-and-export)
```

Chapter 32

Add newpage before a heading

```
(defun org/get-headline-string-element  (headline backend info)
  (let ((prop-point (next-property-change 0 headline)))
    (if prop-point (plist-get (text-properties-at prop-point headline) :pa

(defun org/ensure-latex-clearpage (headline backend info)
  (when (org-export-derived-backend-p backend 'latex)
    (let ((elmnt (org/get-headline-string-element headline backend info)))
      (when (member "newpage" (org-element-property :tags elmnt))
        (concat "\\clearpage\n" headline)))))

(add-to-list 'org-export-filter-headline-functions
             'org/ensure-latex-clearpage)
```

Chapter 33

Glossary and Acronym build using Latexmk

Add the following snippet in the file "~/.latexmkrc": (Source: `https://tex.stackexchange.com/a/44316`)

```
add_cus_dep('glo', 'gls', 0, 'run_makeglossaries');
add_cus_dep('acn', 'acr', 0, 'run_makeglossaries');

sub run_makeglossaries {
    my ($base_name, $path) = fileparse( $_[0] ); #handle -outdir param by s
    pushd $path; # ... cd-ing into folder first, then running makeglossarie

    if ( $silent ) {
        system "makeglossaries -q '$base_name'"; #unix
        # system "makeglossaries", "-q", "$base_name"; #windows
    }
    else {
        system "makeglossaries '$base_name'"; #unix
```

```
        # system "makeglossaries", "$base_name"; #windows
    };

    popd; # ... and cd-ing back again
}

push @generated_exts, 'glo', 'gls', 'glg';
push @generated_exts, 'acn', 'acr', 'alg';
$clean_ext .= ' %R.ist %R.xdy';
```