TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

JAAKKO PASANEN
NLP FOR CUSTOMER SUPPORT AGENT

Master of Science thesis

Examiner: Prof. Ari Visa
Examiner and topic approved by the
Faculty Council of the Faculty of
xxxx
on 30th July 2014

ABSTRACT

JAAKKO PASANEN: NLP for Customer Support Agent
Tampere University of Technology
Master of Science thesis, xx pages, x Appendix pages
xxxxxx 201x
Master's Degree Programme in xxx Technology
Major:
Examiner: Prof. Ari Visa
Keywords:

The abstract is a concise 1-page description of the work: what was the problem, what was done, and what are the results. Do not include charts or tables in the abstract.

Put the abstract in the primary language of your thesis first and then the translation (when that is needed).

TIIVISTELMÄ

JAAKKO PASANEN: Luonnollisen kielen ymmärrys asiakapalveluagentilla
Tampereen teknillinen yliopisto
Diplomityö, xx sivua, x liitesivua
xxxkuu 201x
xxx koulutusohjelma
Pääaine:
Tarkastajat: Prof. Ari Visa
Avainsanat:

The abstract in Finnish. Foreign students do not need this page.

Suomenkieliseen diplomityöhön kirjoitetaan tiivistelmä sekä suomeksi että englanniksi.

Kandidaatintyön tiivistelmä kirjoitetaan ainoastaan kerran, samalla kielellä kuin työ. Kuitenkin myös suomenkielisillä kandidaatintöillä pitää olla englanninkielinen otsikko arkistointia varten.

PREFACE

This document template conforms to Guide to Writing a Thesis at Tampere University of Technology (2014) and is based on the previous template. The main purpose is to show how the theses are formatted using LaTeX (or LaTeX to be extra fancy) .

The thesis text is written into file `d_tyo.tex`, whereas `tutthesis.cls` contains the formatting instructions. Both files include lots of comments (start with %) that should help in using LaTeX. TUT specific formatting is done by additional settings on top of the original `report.cls` class file. This example needs few additional files: TUT logo, example figure, example code, as well as example bibliography and its formatting (`.bst`) An example makefile is provided for those preferring command line. You are encouraged to comment your work and to keep the length of lines moderate, e.g. <80 characters. In Emacs, you can use `Alt-Q` to break long lines in a paragraph and `Tab` to indent commands (e.g. inside figure and table environments). Moreover, tex files are well suited for versioning systems, such as Subversion or Git.

Acknowledgements to those who contributed to the thesis are generally presented in the preface. It is not appropriate to criticize anyone in the preface, even though the preface will not affect your grade. The preface must fit on one page. Add the date, after which you have not made any revisions to the text, at the end of the preface.

Tampere, 11.8.2014

On behalf of the working group, Erno Salminen

CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS

ANN	Artificial Neural Network
LAS	Labelled Attachment Score
LDA	Latent Dirilecht Allocation
LSA	Latent Semantic Analysis
LSTM	Long short term memory; type of RNN with short term memory.
NER	Named entity recognition
NLP	Natural Language Processing
POS	Part-of-speech; also called lexical category
RNN	Recurrent neural network
S-LSTM	Stack long short term memory
TUT	Tampere University of Technology
UAS	Unlabelled Attachment Score

TERMS AND NOTES

1-of-V Coding	Representing words as sparse binary vectors which have 1 at the word's vocabulary index and 0 all others. With vocabulary dog, cat, mouse, dog becomes [1, 0, 0], cat becomes [0, 1, 0] and mouse [0, 0, 1].
Bag-of-words	
Constituent	In syntactic analysis, a constituent is a word or a group of words that function(s) as a single unit within a hierarchical structure. Many constituents are phrases. *Yesterday I saw* ***an orange bird with a white neck***
Feature	Data representation that can be effectively exploited in machine learning tasks. E.g. Word occurrence frequencies.
Feature Vector	
Lemmatisation	Process of finding the base form of a word, e.g. flew -> fly
n-gram	
Parsing	Within computational linguistics the term is used to refer to the formal analysis by a computer of a sentence or other string of words into its constituents, resulting in a parse tree showing their syntactic relation to each other, which may also contain semantic and other information.
POS-tagging	Process of marking up a word to particular part-of-speech (nouns, verbs, etc...) based on both its definition and its context.
Skip-gram	
Structured Prediction	Predicting structured objects, rather than scalar discrete or real values. Translating a natural language sentence into a syntactic representation such as a parse tree can be seen as a structured prediction problem in which the structured output domain is the set of all possible parse trees.
Tree bank	Parsed text corpus that annotates syntactic or semantic sentence structure. Contains trees for sentences where phrases in a sentence are structured in a tree of syntactic or semantic relations. Very useful for training POS-taggers etc...
Tri-Training	Parsing unlabeled data with two different parses and selecting only the sentences for which the two parsers produce the same trees Weiss et al. 2015
Word vector	N-dimensional vector representation of a word with interesting properties such as: vector('Paris') - vector('France') + Vector('Italy') ->

vector('Rome')

1. INTRODUCTION

Testing citation Andor et al. 2016

2. NATURAL LANGUAGE PROCESSING

2.1 Feature Engingeering in NLP

2.1.1 Word Embeddings

- Traditionally words have been represented by indices. Mikolov, Corrado, et al. 2013
- Index representation is simple as computationally cheap, making use of huge datasets possible. Simple models with huge data outperform complex models with less data. Mikolov, Corrado, et al. 2013
- Word embeddings represent words as n-dimensional vectors. Mikolov, Corrado, et al. 2013
- see section 1.2 of Mikolov, Corrado, et al. 2013 for previous work and history of word embeddings
- See LSA and LDA for previous systems. Neural networks significantly outperform LSA in preserving linearities. LDA doesn't scale for large datasets. Mikolov, Corrado, et al. 2013
- Word embeddings try to map words with semantic similarities close to each other. Words may have several types of similarities such as France and Italy are countries but dogs and triangles are both in plural form. Mikolov, Yih, et al. 2013

2.1.2 Word2vec

- Mikolov, Corrado, et al. 2013
- Can be used with datasets of billions of words
- Has two models: Continuous bag-of-words and continuous skip-gram

- Continuous bag-of-words predicts current word from the context (surrounding words)
- Continuous skip-gram predicts context (surrounding words) from current word.
- Continuous Bag-of-Words is better for small datasets, continuous skip-gram is better for large datasets.
- CBOW is better for syntax, Skip-gram is better for semantics.
- Can be used to find semantic relationships like vector('biggest') - vector('big') + vector('small') => vector('smallest')
- State of the art (as of 2013)

2.2 POS Tagging

- Started from rule based taggers
- Tagger by Brill 1992 (known as Brill tagger) learns the rules and as such can be considered as a hybrid approach
- Contemporary research is focused on statistical and ANN based taggers
- Rest of this section focuses on statistical parsers
- Ling et al. 2015 introduced S-LSTM based State-of-the-art tagger
- Andor et al. 2016 Improved accuracy with transition based tagger
- Chen and Manning 2014 were first to represent POS-tag and arc labels as embeddings
- Andor et al. 2016 and Weiss et al. 2015 built their solutions based on Chen and Manning 2014
- Nivre 2004 introduced system for transition based taggers known as arc-starndard system Chen and Manning 2014

2.2.1 Turku Dependecy Treebank

- Treebanks are needed in computational linguistics.
- First Finnish treebank.

- Open licence, including for text annotated
- 204339 tokens, 15126 sentences
- Based on Stanford Dependency scheme with minor modifications to exclude phenomena not present in Finnish and to include new annotations not present in English.
- Transposed to CoNNL-U scheme by universal dependencies project
- Connexor Machinese Syntax is the only currenty available Finnish full dependency parser.
- Texts from 10 different categories ranging from news and legal text to blog entries and fiction.
- Dependency parsing is done manually with full double annotation process.
- Uses Omorfi for morphological analysis. Ambiguous tokens are handled partly manually, partly rule based and partly with machine learning.
- FTB uses 3 different taggers for morphology, check them out!
- FTB is 97% grammar examples, meant for rule based POS tagger development

2.2.2 Transition Based Parsers

- Good balance between efficiency and accuracy Weiss et al. 2015
- Parsed left to right; at each position the parses chooses action from a set of possible actions.
- Greedy models are fast but error prone and need hand engineered features Weiss et al. 2015
- Actions can be chosen by ANN to avoid hand engineering Chen and Manning 2014, Weiss et al. 2015

2.2.3 Syntaxnet

- Transition based
- Locally and globally normalized
- Backpropagation through entire net
- State-of-the-Art
- Andor et al. 2016

2.3 Dependecy Parsing

2.4 Co-Reference Parsing

2.5 Sentence Segmentation

2.6 Lemmatisation

- See OMorFi

BIBLIOGRAPHY

Andor, D. et al. (2016). "Globally Normalized Transition-Based Neural Networks". In: *Acl 2016*, pp. 2442–2452. DOI: 10.18653/v1/P16-1231. arXiv: arXiv:1603.06042v2.

Brill, E. (1992). "A Simple Rule-Based Part of Speech Tagger". In: *Applied natural language*, p. 3. ISSN: 00992399. DOI: 10.3115/1075527.1075553. arXiv: 9406010 [cmp-lg].

Chen, D. and C. D. Manning (2014). "A Fast and Accurate Dependency Parser using Neural Networks". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* i, pp. 740–750. ISSN: 9781937284961. URL: https://cs.stanford.edu/%7B~%7Ddanqi/papers/emnlp2014.pdf.

Ling, W. et al. (2015). "Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* September, pp. 1520–1530. DOI: 10.18653/v1/D15-1176. arXiv: 1508.02096. URL: http://dx.doi.org/10.18653/v1/d15-1176$%5Cbackslash$nfile:///Files/68/6810072d-e133-426e-807f-445df2840420.pdf$%5Cbackslash$npapers3://publication/doi/10.18653/v1/d15-1176$%5Cbackslash$nhttp://arxiv.org/abs/1508.02096.

Mikolov, T., G. Corrado, et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12. ISSN: 15324435. DOI: 10.1162/153244303322533223. arXiv: arXiv:1301.3781v3. URL: http://arxiv.org/pdf/1301.3781v3.pdf.

Mikolov, T., W.-t. Yih, and G. Zweig (2013). "Linguistic regularities in continuous space word representations". In: *Proceedings of NAACL-HLT* June, pp. 746–751. URL: http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:Linguistic+Regularities+in+Continuous+Space+Word+Representations%7B%5C#%7D0$%5Cbackslash$nhttps://www.aclweb.org/anthology/N/N13/N13-1090.pdf.

Nivre, J. (2004). "Incrementality in deterministic dependency parsing". In: *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, pp. 50–57. DOI: 10.3115/1613148.1613156. URL: http://dl.acm.org/citation.cfm?id=1613156.

Weiss, D. et al. (2015). "Structured Training for Neural Network Transition-Based Parsing". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Lan-*

guage Processing (Volume 1: Long Papers) 2012, pp. 323–333. DOI: 10.3115/v1/P15-1032. arXiv: 1506.06158. URL: http://www.aclweb.org/anthology/P15-1032.

APPENDIX A. SOMETHING EXTRA

Appendices are purely optional. All appendices must be referred to in the body text

APPENDIX B. SOMETHING COMPLETELY DIFFERENT

You can append to your thesis, for example, lengthy mathematical derivations, an important algorithm in a programming language, input and output listings, an extract of a standard relating to your thesis, a user manual, empirical knowledge produced while preparing the thesis, the results of a survey, lists, pictures, drawings, maps, complex charts (conceptual schema, circuit diagrams, structure charts) and so on.