



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 L^AT_EX 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

1. Introduction	2
2. Natural Language Processing	3
2.1 POS Tagging	3
2.2 Dependency Parsing	3
2.3 Co-Reference Parsing	3
2.4 Sentence Segmentation	3
2.5 Lemmatisation	3
Bibliography	4
APPENDIX A. Something extra	5
APPENDIX B. Something completely different	6

LIST OF ABBREVIATIONS AND SYMBOLS

NER	Named entity recognition
NLP	Natural Language Processing
POS	Part-of-speech also called lexical category
TUT	Tampere University of Technology

TERMS AND NOTES

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***

Lemmatisation Process of finding the base form of a word, e.g. flew -> fly

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

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...

Word vector $\text{vector('Paris')} - \text{vector('France')} + \text{Vector('Italy')} \rightarrow \text{vector('Rome')}$

1. INTRODUCTION

Testing citation Andor et al. 2016

2. NATURAL LANGUAGE PROCESSING

2.1 POS Tagging

2.2 Dependency Parsing

2.3 Co-Reference Parsing

2.4 Sentence Segmentation

2.5 Lemmatisation

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.

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.