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

1. Introduction . . . . . 2

2. Natural Language Processing . . . . . 3

2.1 POS Tagging . . . . . 3

2.2 Dependecy Parsing . . . . . 3

2.3 Co-Reference Parsing . . . . . 3

2.4 Sentence Segmentation . . . . . 3

2.5 Lemmatization . . . . . 3

Bibliography . . . . . 4

APPENDIX A. Something extra . . . . . 5

APPENDIX B. Something completely different . . . . . 6

# LIST OF FIGURES

# LIST OF TABLES

# 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 |

# 1. INTRODUCTION

Testing citation Andor et al. 2016

# 2. NATURAL LANGUAGE PROCESSING

## 2.1 POS Tagging

POS (part-of-speech) is the process of marking up a word in a text corpus to particular prat of speech based on both its definition and its context.

POS, also called lexical category, is a category of words which have similar grammatical properties. English parts of speech are noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, and sometimes numeral, article or determiner.

POS tagging may also use additional finer lexical categories to a word for additional information on word's role. Fundamental problem in coarse and fine POS tagging is the lack of standardized terminology for POS tags. Universal dependencies project tries to address this issue.

POS tagging is often the first step in Natural language processing (NLP) pipeling. Currently (October 2016) the best part-of-speech tagger is described by Andor et. al (2016) and implemented as Google's SyntaxNet. Other notable POS tagging software include Stanford University's POS tagger, spaCy - python library for NLP, Python Natural Language Toolkit.

## 2.2 Dependecy Parsing

## 2.3 Co-Reference Parsing

## 2.4 Sentence Segmentation

## 2.5 Lemmatization

# 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.