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
December 2016
Master's Degree Programme in Automation Technology
Major: Learning and Intelligent Systems
Examiner: Prof. Ari Visa
Keywords: Hype

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
Joulukuu 2016
Automaatiotekniikan koulutusohjelma
Pääaine: Oppivat ja älykkäät järjestelmät
Tarkastajat: Prof. Ari Visa
Avainsanat: Hype

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

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

NOTES

0.1 Terms for Computational Linguistics

Bag-of-words	
Clausal Complement	
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***
Corpus	A collection of texts with linguistic annotations.
Lemmatisation	Process of finding the base form of a word, e.g. flew -> fly
Lexeme	A basic lexical unit of a language consisting of one word or several words, the elements of which do not separately convey the meaning of the whole.
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	
Token	A structure representing a lexeme that explicitly indicates its categorization for the purpose of parsing.
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	vector('Paris') - vector('France') + Vector('Italy') -> vector('Rome')

0.2 Universal Dependecies

0.2.1 CoNNL-U format

Universal dependencies use CoNNL-U format for treebanks, CoNNL-U is revised version of CoNNL-X. Annotations are encoded in text files with word lines, blank lines for sentence boundaries and comments starting with hash (#).

Word lines consist of following columns:

ID	Word ID in sentence
FORM	Word form or punctuation symbol
LEMMA	Lemma or stem of word form
UPOSTAG	Universal part-of-speech tag
XPOSTAG	Language specific part-of-speech tag
FEATS	List of morphological features
HEAD	Head of the curren token, value of ID or zero (0)
DEPREL	Universal dependecy relation to the HEAD
DEPS	List of secondary dependencies
MISC	Any other annotation

Example in Finnish: Jäällä kävely avaa aina hauskoja ja erikoisia näkökulmia kaupunkiin

ID	FORM	LEMMA	UPOSTAG	XPOSTAG
1	Jäällä	jää	NOUN	N
2	kävely	kävely	NOUN	N
3	avaa	avata	VERB	V
4	aina	aina	ADV	Adv
5	hauskoja	hauska	ADJ	A
6	ja	ja	CONJ	C
7	erikoisia	erikoinen	ADJ	A
8	näkökulmia	näkö#kulma	NOUN	N
9	kaupunkiin	kaupunki	NOUN	N
10	.	.	PUNCT	Punct

FEATS
Case=Ade|Number=Sing
Case=Nom|Number=Sing
Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act
—
Case=Par|Degree=Pos|Number=Plur
—
Case=Par|Degree=Pos|Number=Plur
Case=Par|Number=Plur
Case=Ill|Number=Sing
—

HEAD	DEPREL	DEPS	MISC
2	nmod	_	_
3	nsubj	_	_
0	root	_	_
3	advmod	_	_
8	amod	_	_
5	cc	_	_
5	conj	8:amod	_
3	dobj	_	_
8	nmod	_	SpaceAfter=No

0.2.2 Universal POS tags

ADJ Adjective. Describing word qualifying noun or noun phrase. **deep**, **intelligent**

ADP Adposition. Word expressing spatial or temporal relations **under**, **around**, **before** or mark various semantic roles **of**, **for**

ADV Adverb. Modifies another word. Typically express manner, place, time, frequency etc. She sang **loudly**. You are **quite** right.

AUX Auxiliary verb. A verb used in forming the tenses, moods, and voices of other verbs. **Do** you want tea?. He **has** given his all.

CONJ Coordinating conjunction. Conjunction placed between words, phrases, clauses or sentences of equal rank. **and**, **but**, **or**.

DET Determiner. Expresses reference of a noun (group). **The** girl is **a** student. **Which** book is that?

INTJ Interjection. Shows emotion or feeling of the author, includes exclamations, curses, greetings and such. **Ouch!**, **hey**, **huh?**.

NOUN Noun. Denotes a person, animal, place thing or idea. The **cat** sat on a **mat**.

NUM Numeral. Number, written with digits or letters. **12**, **eleven**.

PART Particle. Cannot be inflected. Interjections and conjunctions. In finnish also **että**, **jotta**, **koska**, **kun** etc...

PRON Pronoun. Replaces (often previously introduced) noun. Joe saw Jill, and **he** waved at **her**.

PUNCT Punctuation. Full stop, comma, bracket etc.

SCONJ Subordinating conjunction. A conjunction that introduces a subordinating clause, e.g. **although**, **because**, **whenever**.

SYM Symbol.

VERB Verb. Conveys an action **bring**, **read**, an occurrence **happen**, **become**, or a state of being **be**, **exist**.

X	Other

1. INTRODUCTION

Testing citation Andor et al. 2016

2. NATURAL LANGUAGE PROCESSING

2.1 Pre-processing

2.2 POS Tagging

2.3 Dependecy Parsing

2.4 Co-Reference Parsing

2.5 Sentence Segmentation

2.6 Lemmatisation

2.7 Synonym recognition

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.