TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

JAAKKO PASANEN

NATURAL LANGUAGE UNDERSTANDING FOR INFORMATION RETRIEVAL IN CHAT BASED CUSTOMER SUPPORT

Master of Science thesis

Examiner: Prof. Ari Visa
Examiner and topic approved by the Faculty Council of the Faculty of Engineering Sciences
on 31st December 2016

late

ABSTRACT

JAAKKO PASANEN: Natural Language Understanding for Information Retrieval in Chat Based Customer Support
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 tiedon hakua varten chat-pohjaisessa asiakaspalvelussa
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

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

NOTES

0.1 Terms for Computational Linguistics

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 Multiset of words appearing in a text with occurrence counts for each word. Used as a tool for feature generation. Does not preserve word order or grammar. Can implemented as a dictionary (or associative array) where words are the keys and counts are the values.

Conditional Random Field

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.

Dimensionality When discussing word embeddings and word vector spaces the dimensionality refers to definition in linear algebra. Dimensionality of arrays in computing means the number of indices required to specify an element in the array. Word vector in 50 dimensional vector space $\mathbb{R}^{50}$ would be represented in computing as one dimensional array of length 50 [d1, d2, d3, ..., d50]

Feature Numeric data representation that can be effectively exploited in machine learning tasks. E.g. Word occurrence frequencies.

Feature Vector Vector containing all the features. For an image a feature vector could be all the raw values of pixels as a single sequence. For a trigram model with 300 dimensional word embeddings a feature vector would be a 900 dimensional vector formed by concatenating all the separate word embedding vectors.

Language Model Probability distribution over sequences of words. Given such a sequence, say of length m, it assigns a probability $P(w_1, \ldots, w_m)$ to the whole sequence. Problems caused by growing vocabulary can be addressed with continuous language models such as neural net language models (NNML). Word2Vec by Mikolov, Corrado, et al. 2013 addresses this problem with Continuous Bag-of-words and Skip-gram models.

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.

n-gram Probabilistic language model where probability of current word is the joint probability of previous n words. Bigram example: $P(I, saw, the, red, house) \approx P(I|{\wedge})P(saw|I)P(red|the)P(house|red)P(\$|house)$. The words unigram, bigram and trigram language model denote n-gram model language models with n = 1, n = 2 and n = 3, respectively.

One-hot Group of bits which the legal comibations of values are only those with a single high (1) and all the others low (0). See also 1-of-V Coding.

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 Language model which predicts the context (previous and next n words) of a current word from the current word instead of traditional way of predicting current word from the context.

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.

Token A structure representing a lexeme that explicitly indicates its categorization for the purpose of parsing. In plain words tokens are instances of words in a text. Not to be confused with word type.

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 Lookup Table Matrix $\mathbf{P} \in \mathbb{R}^{d \times |V|}$ of d rows and |V| columns, where d is the word vector dimensionality and |V| is the size of vocalbulary. Word lookup tables are unable to generate representations for previoulsy unseen words, as is required for morphology. Ling et al. 2015

Word Type Unique words in a text. *Good wine is good* has 4 tokens but only 3 word types.

Word vector N-dimensional vector representation of a word with interesting properties such as: 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

- Most of the NLP work has been for english.
- Cross-linguistic annotation and parsing has been a reality only after introduction of The Universal Dependencies project and SyntaxNet.
- Similarly Finnish parsing has been unreachable until the first Finnish corpus Turku Dependency Treebank Haverinen et al. 2014 and cross-linguistic parsers.

2.1 Pre-processing

2.2 Feature Engingeering in NLP

2.2.1 Word Embeddings

- Traditionally words have been represented by indices. Mikolov, Corrado, et al. 2013
- Next step was to use 1-of-V coding.
- 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
- Word embeddings can be generated from character sequences with significantly better performance for morphological languages. Ling et al. 2015

2.2.2 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)

2.2.3 GloVe

Pennington et al. 2014

2.2.4 Charater to Word

Ling et al. 2015

2.3 Annotations

- Stanford Dependencies De Marneffe et al. 2006
- Stanford Dependencies emerged as de facto annotation scheme for english, but has been adapted to several other languages including Finnish. Nivre et al. 2016, Haverinen et al. 2014.
- Turku Dependency Treebank has been tranformed into universal dependencies. Pyysalo et al. 2015
- Unified annotation scheme reduces need for cross-language adaptations in downstream development. Petrov et al. 2012
- Universal Dependencies project started from the requirement for cross-linguistically consistent treebank annotations even for morphological languages. Nivre et al. 2016.
- Universal Dependencies project was born from merging several previous attempts to form a cross-linguistically sound dependency annotation schemes. Nivre et al. 2016
- UD data has been encoded in the CoNLL-U format, a revision of the popular CoNLL-X format. Nivre et al. 2016
- UD treebanks released in November 2015. Nivre et al. 2016

2.4 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

2.4.1 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

2.4.2 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

2.4.3 Syntaxnet

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

2.5 Dependecy Parsing

2.6 Co-Reference Parsing

2.7 Sentence Segmentation

This is not relevant in our system?

2.8 Machine Translation

See section 4.1.1 for more info on our implementation.

2.9 Lemmatisation

2.10 Synonym recognition

- See OMorFi

3. CHAT BASED CUSTOMER SUPPORT

Customer support problem domain. History?

3.1 Chat Revolution

3.2 Automation with Virtual Agents

3.3 Existing Systems

4. OUR SYSTEM

- This thesis only considers customer message understanding.
- Greetings, Chatting, information retrieval and response generation are important part of a dialogue system but have been left out of scope of this thesis.
- Large portion of the research on the conversational models focus on chatter systems where utterance responses are retrieved or generated from response corpus. These systems do not aim to solve customer problem by incorporating relevant information retrieved from a specific knowledge base.
- Utterance systems are not suitable for customer support out side of simple frequently asked questions.
- Furthermore large portion of responses generated by neural conversation models are safe responses e.g. *I don't know* Li et al. 2015.
- Conversational models trained on customer support history incorporate data from long time period and as such data may contain outdated information which cannot be corrected unless sentences containing outdated information are replaced. Finding and replacing such information from customer support dialogues requires reading much of the dialogues by humans and rewriting conversation history; task both time consuming, error prone and expensive.
- Also incorporating new information into such dataset requires manual conversation synthesis where humans write new simulated conversations into a dialogue history used for training the system.
- Other ways for updating old and adding new information are required for real world customer support. Our system draws it's information from a knowledge base with traditional user interface for updating and adding new knowledge.

4.1 Pipeline

4.1.1 Translating Colloquial Speech

Mä -> Minä

4.2 Evaluation

- Recently end-to-end dialogue system have adopted metrics from amchine translation and text summarization. These don't work so well. Liu et al. 2016
- Liu et al. 2016 considers unsupervised utterance systems where response is generated or selected from a set of possible responses. This is irrelevant since we are building a problem solving dialogue system instead of chattering system.

5. CUSTOMER BUSINESS VALUE AND SATISFACTION

6. CONCLUSIONS AND FUTURE WORK

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.

De Marneffe, M.-C., B. MacCartney, and C. D. Manning (2006). "Generating typed dependency parses from phrase structure parses". In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 449–454. DOI: 10.1.1.74.3875. URL: http://nlp.stanford.edu/pubs/LREC06%7B%5C_%7Ddependencies.pdf.

Haverinen, K. et al. (2014). "Building the essential resources for Finnish: the Turku Dependency Treebank". In: *Language Resources and Evaluation* 48.3, pp. 493–531. ISSN: 15728412. DOI: 10.1007/s10579-013-9244-1.

Li, J. et al. (2015). "A Diversity-Promoting Objective Function for Neural Conversation Models". In: *Arxiv* Mmi, pp. 110–119. arXiv: 1510.03055. URL: http://arxiv.org/abs/1510.03055.

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%7B%%7D5Cnfile:///Files/68/6810072d-e133-426e-807f-445df2840420.pdf%7B%%7D5Cnpapers3://publication/doi/10.18653/v1/d15-1176%7B%%7D5Cnhttp://arxiv.org/abs/1508.02096.

Liu, C.-W. et al. (2016). "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation". In: *Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 13. arXiv: 1603.08023. URL: http://arxiv.org/abs/1603.08023.

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%7B%%7D5Cnhttps://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.

Nivre, J. et al. (2016). "Universal Dependencies v1: A Multilingual Treebank Collection". In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1659–1666.

Pennington, J., R. Socher, and C. D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. ISSN: 10495258. DOI: 10.3115/v1/D14-1162. arXiv: 1504.06654.

Petrov, S., D. Das, and R. Mcdonald (2012). "A Universal Part-of-Speech Tagset". In: arXiv: 1104.2086.

Pyysalo, S. et al. (2015). "Universal Dependencies for Finnish". In: *Nordic Conference of Computational Linguistics NODALIDA 2015* Nodalida, p. 163.

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 Language 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.