Stockholm University

Introduction to Natural Language Processing (NLP)

Hercules Dalianis

Department of Computer and Systems Sciences (DSV)

hercules@dsv.su.se

Stockholm University

Introduction of lecturers

Prof Hercules Dalianis (Lecturer and Lab assistant)

Dr. Martin Duneld (Lecturer)

Associate professor Aron Henriksson (Lecturer on parental leave)

Phd student Thomas Vakili (Lecturer and Lab assistant)

Phd student Yongchao Wu (Lab assistant and Guest lecturer)

Dr Jussi Karlgren, (Guest lecturer, Spotify)

Content of course

- What is natural language (NL)?
- Tokenisation, morphology, tagging, parsing, generation
- Modeling of NL, word embeddings
- Evaluation
- Classic machine learning (ML) and neural (deep) learning methods (BERT)
- Classification and Named Entity Recognition
- Applications of NLP

Structure of the course

- Nine lectures (One recorded)
- Two guest lectures
- Three lab exercises (groups of two)
- One research topic report that may be developed to a master thesis (groups of two)
- Written exam March 17, 2023
- Exam retake April 26, 2023

Three laboration exercises

- Lab 1. Basic Natural language processing
- Lab 2. Text classification
- Lab 3. Named entity recognition
- Oral presentation of your lab results
- If you get approved of all the lab exercises before the deadlines you get bonus and you don't need to do the last exam question.
- Deadlines for bonus, see Ilearn.

Stockholm University

Literature

- Dan Jurafsky and James H. Martin (2021)
  Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed. draft)
  https://web.stanford.edu/~jurafsky/slp3/
- Steven Bird, Ewan Klein, and Edward Loper (2009)
  Natural Language Processing with Python
  ISBN: 9780596516499
  O'Reilly Media

Code of Honor and Regulations

- Plagiarism is under no circumstances acceptable. Please read DSV's Code Of Honor and Regulations.
- Plagiarism will be reported to the Vice-Chancellor of the university and will be punished!
- Fall 2016, 3 students have been reported for cheating during the exam. They were convicted guilty for cheating. Their presence at Stockholm University were suspended for 2 months, also CSN support suspended, the suspension protocol is available as a public document.

Curiginal

77%	MATCHING BLOCK 12/80	W	https://www.researchgate.net/publication/34191 ...

aimed for giving a complete tweet sentiment analysis based on ordinal regression with machine learning algorithms. The recommended model included pre-processing tweets as

the start

100%	MATCHING BLOCK 13/80	W	https://www.researchgate.net/publication/34191 ...

step and with the feature extraction model, an effective feature was generated. The methods such as

support vector regression (SVR), random forest (RF), Multinomial logistic regression (SoftMax), and Decision Trees (DTs) were employed for classifying the sentiment analysis. Moreover,

the

81%	MATCHING BLOCK 14/80	W	https://www.researchgate.net/publication/34191 ...

Twitter dataset was used for experimenting with the recommended model. The test results have shown that the recommended model has attained the best accuracy, and also DTs were performed well when compared

to other methods. In 2018, the author [3]

87%	MATCHING BLOCK 15/80	W	https://www.researchgate.net/publication/34191 ...

have suggested multi- strategy sentiment analysis models using semantic fuzziness for resolving the issues. The result has demonstrated that the proposed model has attained high efficiency. In 2020,

Stockholm University

Schedule

- Introduction Natural Language Processing
- Basic Natural Language Processing I: Tokenisation, Normalisation, Tagging
- Basic Natural Language Processing II: Parsing and Generation
  - Basic Natural language processing (Lab)
- Word Embeddings and Language Models
- Evaluation
- Classification and Sentiment analysis
  - Text classification (Lab)
- Named entity recognition, rule- and machine learning based

Schedule (cont)

- Deep learning
  - Named Entity Recognition (Lab)
- Applications in Natural Language Processing
- Guest lecturer Jussi Karlgren, Spotify
- Guest lecturer Yongchao Wu, DSV, *Automated essay scoring, the current state-of-the-art and the future*
- Research topics in NLP
- Presentation research topics by students I
- Presentation research topics by students II

Intro lecture – Hercules

- What is natural language?
- What is text? What is speech?
- Natural language understanding
- Natural language generation
- Methods for NLP
- Applications
- Examples on master thesis topics
- Some quizzes

Martin Duneld

Stockholm University

- **Basic Natural Language Processing I**
  - Common NLP tasks, with a focus on pre-processing text
  - Tokenisation
  - Term normalisation
  - Morphology
  - Tagging
  - Stop words
- **Evaluation**
  - Extrinsic vs. intrinsic evaluation
  - Manual vs. semi-automatic vs. automatic evaluation
  - Common evaluation metrics in NLP evaluation
  - The importance of baselines and inter-assessor agreement
  - Significance testing
  - Getting the most out of you (annotated) data

Hercules Dalianis

Stockholm University

- **Basic Natural Language Processing II**
  - Properties of natural languages
  - Grammars
  - Parsing
  - Generation
- **Applications in Natural Language Processing**
  - Machine translation
  - Natural language interfaces
  - Text generation – Robot journalism
  - Speech recognition and synthesis
  - Spell and grammar checking
  - Information extraction – NER etc.
  - Text summarisation
  - De-identification and pseudonymisation
  - News monitoring / Market intelligence

Word Embeddings and Language Models – Aron Henriksson (Recorded)

Stockholm University

How do we create representations of language?

Distributional semantics

- *The distributional hypothesis*

An evolution of models

- Latent semantic analysis
- Random indexing
- Word2Vec
- ELMo

Language models

- BERT, GPT, XLNet
- Pre-training & fine-tuning

Text classification and sentiment analysis – Thomas Vakili

- From: *Speech and Language Processing (3rd ed. draft)* Dan Jurafsky and James H. Martin
  - **Chapter 4:** Text classification with bag-of-word models, Naive Bayes' for sentiment analysis
  - **Chapter 20:** How to organize lexicons for sentiment/affect/connotation/emotion. How to use such lexicons for sentiment analysis
- + examples from research on applications of sentiment analysis and other types of text classification

Deep Learning and the Future of Language Models – Thomas Vakili

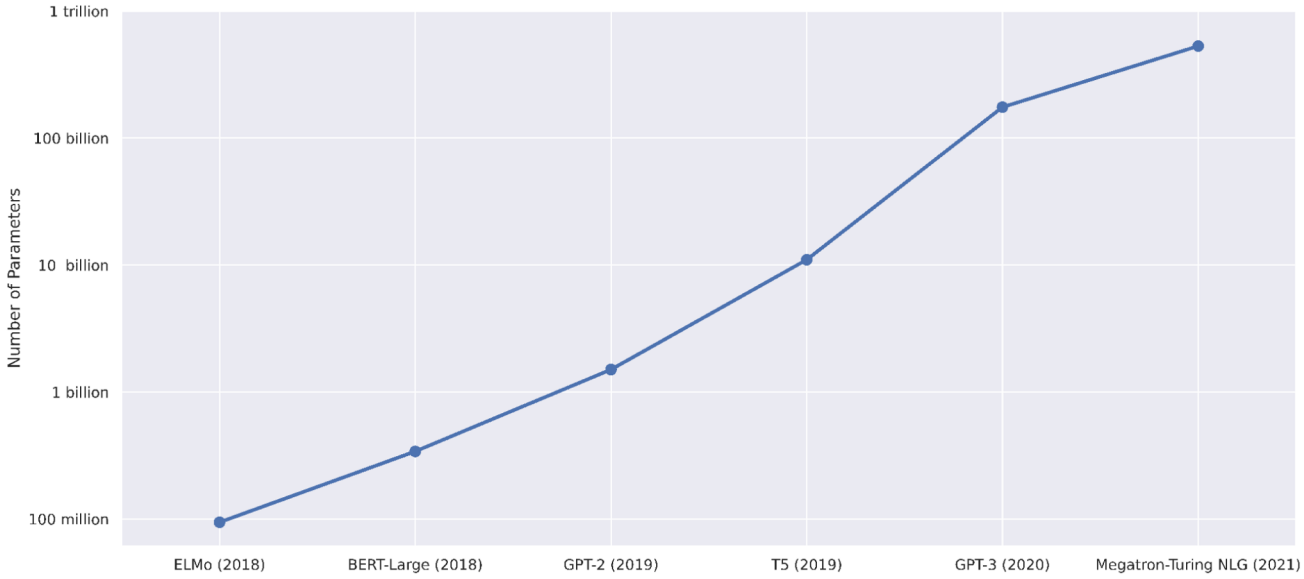
Stockholm University

Learning to model sequences:

From RNNs to BiLSTMs to Transformers

Unsolved challenges in language modelling:

- Biased models and datasets
- Privacy leakage
- Ever-increasing resource use



Stockholm University

Introduction to Natural Language Processing (NLP)

Intro lecture – Hercules

- What is natural language?
- What is text? What is speech?
- Natural language understanding
- Natural language generation
- Methods for NLP
- Applications
- Examples on master thesis topics
- Some quizzes

Stockholm University

What is natural language (NL)?

- NL is used in communication between living beings - humans, animals - even plants?
- In contrast to formal language, logic, mathematics, programming languages
- Ambiguous / non ambiguous

  I saw a man on a hill with a telescope

Stockholm University

What is natural language processing?

- Contains parts of computer science, linguistics, (and phonetics), logic, mathematics and psychology, but also philosophy and signal processing and of course - human computer interaction.

Natural language processing

- Language Technology
- Human Language Technology
- Computational Linguistics

Swedish

- Språkteknologi
- Datorlingvistik (datalingvistik means something different)

Stockholm University

Text and speech

- Text contains written tokens, with spaces
- Speech signals, frequencies, continuously
- In ancient times written text did not have spaces between words
  - Chinese writing do not have spaces between characters/words

Natural Language Processing (NLP)

Stockholm University

- Computers can communicate with humans and vice versa in natural language
- Assist the human with translation
- Summarisation
- Retrieve information
- Extract information from the text.
- Decide on content on large amounts of information in text.
- Generate text with new information

Morphologi and syntax

- Words have morphology - inflections
  - *word, words* - singular plural determiners
  - *have, has, had* – tense
- Sentences have syntax. The order of symbols/words
  - *This sentence is written in the correct syntax.*
  - *This sentence written is syntax wrong.*

Semantics and pragmatics

- *Sentences should have – semantics – meaning.*

  The meaning of symbols.

  - *This sentence is written in the correct syntax.*
  - *Colorless green ideas sleep furiously - semantically nonsense*

- A sentence may have pragmatics. How to use the sentence

  - *Can you pass the salt?*

Prolog and DCG for syntax

- The programming language Prolog in DCG (Definite Clause Grammar) format can be easily used to define and execute a grammar.

```
% grammar
s --> np, vp.
vp --> verb.
vp --> verb, np.
np --> article, noun.
np --> article, adjective, noun.
%np --> np, conj, np.
```

```
% dictionary
verb --> [is].
article --> [the].
noun --> [flower].
adjective --> [red].
conj --> [and].
```

Stockholm University

Natural language parsing – understanding
Natural language generation

- Two steps for Machine translation
  - To parse – understand (syntactically) input language
  - To generate – create comprehensible text.
- 1959-1983 Rule based systems – very costly to create
  - Grammar or grammar rules – thousands handcrafted for the syntax
  - Dictionaries - hundred thousands words

Stockholm University

Natural language generation

- Fewer rules, but still complicated.
  - Automatic text summarisation
  - Robot journalism etc.

Text versus structured data

- Machine learning (ML) mostly structured data
- Text considered scary or difficult
- Make text to numbers and run relatively easy
- Playing with words
  - Common words
  - Uncommon words
  - Order of words
  - Length of words
  - Length of sentences, etc.

Stockholm University

Playing with words

- Common words in one text
- In several documents
- Term frequency (tf)
- Inverse document frequency (IDF)
- Tf * IDF importance of one word in one text
- Stopwords

Stockholm University

Word features

- POS tag – *noun, verb, adj, etc.*
- Initial capital letter – *Hercules Yes, books, No*
- Capital letters – upper case letters – *DSV, Yes*
- Lower case letters – *Hercules, No books, Yes*
- Lemma – book, read
- Compounds – *icecream, sunflower*
- *De-compounding – ice cream, sun flower*
- Word – word before word or after and features of it
- Etc.

Cryptography and information retrieval

- Stop words – common words in all texts
  - *For example: and, or in, under, with, between…* around 200 stop words or 40% of total amount of words in a text
- Word frequencies
  - In one text compared to all texts

    IDF – Inverse Document Frequency
  - tf x IDF in search engines.

Vector space

- Each word in a text is represented by a multi dimensional vector
- Similarity between words
- No position in text

Tokens - Words

- Tokens - all content incl period, question mark, etc
- Words - just alphanumeric content

Stockholm University

Tokenisation

- Read a string and tokenize it
  - Space is one delimiter
  - But also **.,?!**
  - How do you tokenize 20.4 mg?

Tokenisation

- Read a string and tokenize it
  - Space is one delimiter
  - But also **.,?!**
  - How do you tokenize 20.4 mg?

  “The patient took 20.4 mg Prednisolon.”

  =>

  ”The” “patient” ”took “20.4” “mg” “Prednisolon” “.”

Tokenisation

- Read a string and tokenize it
  - Space is one delimiter
  - But also **.,?!**
  - How do you tokenize 20.4 mg?

  “The patient took 20.4 mg Prednisolon.”

  =>

  ”The” “patient” ”took” “20.4” “mg” “Prednisolon” “.” OR

  ”The” “patient” ”took” “20” ”.” “4” “mg” “Prednisolon” “.”

Stockholm University

POS - Part Of Speech Tagging

- Read the tokenised text and tag it

  “The” “patient” ”took “20.4” “mg” “Prednisolon” “.”

  ⇒POS - Part of Speech tagging

  ('The', 'DT'), ('patient', 'NN'), ('took', 'VBD'), ('20.4', 'CD'), ('mg', 'NN'), ('Prednisolon', 'NNP'), ('.', '.')

Tag set

- NN noun, singular ‘patient’..
- NNP proper noun, singular ‘'Prednisolon’..
- RB adverb ’very, silently’,..
- VB verb, base form take
- VBD verb, past tense took
- JJ adjective ’ill’..
- CC coordinating conjunction ’and’
- CD cardinal digit ’20.4’…
- DT determiner ’the’..

+ more

Stockholm University

Parsing

- Read the tokenised text and parse it

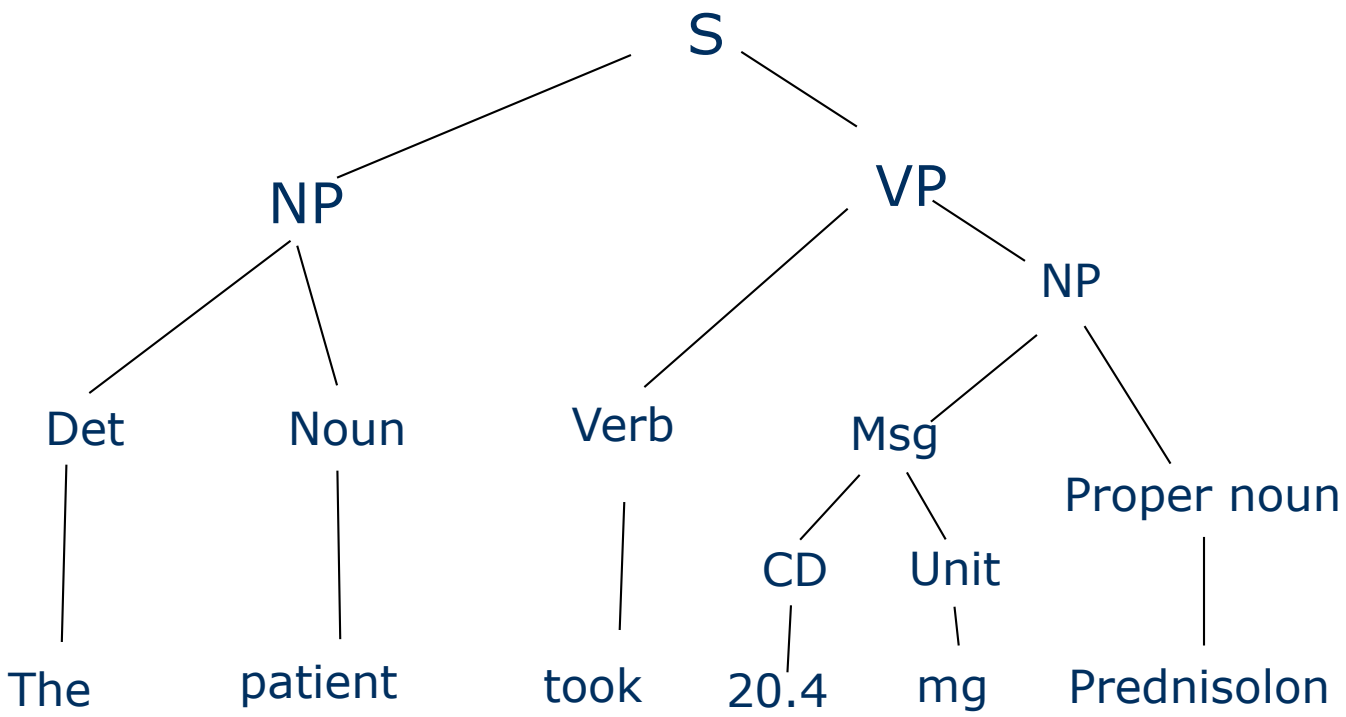
  “The” “patient” ”took “20.4” “mg” “Prednisolon” “.”

  ⇒ parsing

  (((“The”, DET), (“patient”, NN),
  NP)

  ((”took”, Verb),
  ((“20.4”, CD) (“mg”, MS) (“Prednisolon”, NNP), NP)
  VP)
  , S)

Syntactic tree



Morphology

- Morphology of words or tokens

  sunflowers

  sun+flower+s

  prefix+stem+plural-s

- Lemmatisation remove “s”

  sunflower

  stemming may produce *sunflow* in this case

  (Lemmatisation linguistically more correct than stemming, stemming more practical)

Compound splitting

sunflowers

=> compound splitting or decompounding

sun flowers

Spelling correction

arplanse

Use spelling correction with

Levenshtein-Damerau Edit distance.

Four operations on one character (two on Transposition)

- Insertion
- Deletion
- Replace
- Transposition

Spelling correction (Cont.)

Stockholm University

Levenshtein distance (or edit distance) between two strings is the number of deletions, insertions, or substitutions required to transform source string into target string.

arplanse => Levenshtein distance = 2 to airplanes

Insert ”i”

airplanse => Levenshtein distance = 1 to airplanes

Transposition ”se”

airplanes => Levenshtein distance = 0 to airplanes

Stockholm University

Ready Python libraries for spell checking.

Peter Norvig spelling corrector

https://github.com/pirate/spellchecker/

\+ dictionary

Free dictionaries in Swedish and Danish

http://runeberg.org/words/

Free dictionary in English

https://www.bragitoff.com/2016/03/english-dictionary-in-csv-format/

Not grammar checking

Machine learning

- Supervised
  - Annotated data sets
    - Named Entity Tagging
    - Classification
- Unsupervised
    - Clustering
- Semi supervised
  - Some annotated data and lots of un-annotated data.
  - Active learning
  - BERT Deep Learning

Stockholm University

Feature engineering

- Feature engineering is used for traditional ML
  - SVM - Support Vector Machine, Random Forest, Naïve Bayes, CRF - Conditional Random fields, etc.
- Word
- Tags (POS word class, lemma, length of word, word before, word after, vector, etc) are features adding meaning

Stockholm University

Deep learning are neural methods

Deep neural learning methods

- Deep learning methods are artificial neural networks with representation learning, for example.
  - Long short-term memory (LSTM),
  - Recurrent neural network (RNN)
  - Generative Pre-Trained Transformer (GPT-2, GPT-3)
  - Bidirectional Encoder Representations BERT
- Neural methods - old methods that are now used because of high computer performance.

Stockholm University

Deep learning methods

- Arranges the feature engineering
- The engineer has to do parameter tuning
- Train, develop, evaluate

Stockholm University

BERT language models

- Pre trained on Gb of texts
- BERT English
- Multilingual BERT on several European languages
- KB BERT (Kungliga Biblioteket/National Library)

=>

- Fine tuned on annotated text for a specific task

Applications of NLP

- Machine translation- translation between languages
- Automatic text summarisation – Extract- Abstract
  - Single document summary
  - Multi document summary
- Robot journalism - create new articles – rule based
  - Stock reports
  - Weather reports
  - Earth quake reports,
  - Sports results.

Stockholm University

User modelling

- Who is the reader?
- What does he or she know already?
- Did s/he ask something?
- Analyse the question

Search engines – information retrieval

- Index contains document - and their representation
  Inverse document index
- User statistics
  - Common queries
  - Autocomplete
- Lemmatisation – search on inflected form
- Spell checkers – the index is dictionary!
- Extracts - snippets

Stockholm University

Named Entity Recognition

- Recognise
  - personal names, locations, organisations, products time points, dates, measures, etc

https://explosion.ai/demos/displacy-ent

Train on manually annotated texts

Morobito Consultants ORG said the company was engaged in June of 2020 DATE to prepare a “repair and restoration plan” for fixes needed under the state recertification requirements. At the time of the collapse this week DATE , the company said, roof repairs were underway but concrete restoration, which was to be handled by another firm, had not begun.

The collapse has stunned industry experts in the Miami GPE area, including John Pistorino PERSON , a consulting engineer who designed the 40-year DATE reinspection program when he was consulting for the county in the 1970s DATE .

- A couple of thousand annotated instances give good results using CRF machine learning
- Using Deep learning gives better results with same amount of training instances.
- Deep learning though pre-trained on millions of texts.

HB Deid - De-identification of electronic patient records in Swedish

- https://hbdeid.dsv.su.se
- Deid = NER + Pseudo (or surrogates)
- Pseudonymisation of patient records

HB Deid

- 9 classes
  - First Name
  - Last Name
  - Phone Number
  - Age
  - Dates / Date part / Full date
  - Health Care Unit
  - Location
  - Personal Number

PHI Class	Instances
First Name	923
Last Name	931
Phone Number	137
Age	55
Full Date	457
Date Part	709
Health Care Unit	1,414
Location	95
Organisation	43
Total	4,764

Table 1: Stockholm EPR PHI Corpus.

Stockholm University

Other health NLP-applications

- Automatic ICD-10 coding
- Detect health care associated infections
- Detect side effects of drug– adverse events
- Detect cancer by detecting early symptoms of cancer
- Spell and grammar checking in the patient record
- Summarise the patient record to a discharge letter

Stockholm University

Opinion mining

- Find if a text is positive or negative
- Positive or negative review of a film, book, camera, car, or other product
- Train on already manually classified texts
- SVM is a well performing traditional algorithm
- Predict if a text is positive or negative

Stockholm University

Manually classified texts

- *The book arrived as expected and was in great shape. Thanks.* POS
- *I am nearly finished with this book - I haven't been this mesmerized by a book in forever! I would certainly recommend it*. POS
- *This seemed too long and too drawn out*. NEG
- *Because you wait and wait, and then three don't turn up at once.* NEG

Authorship detection and plagiarism detection

- Who wrote this text? Same author as another text?
- Similarity between small parts of text
- 4 – 6 words distance
  - We use such programs at the courses at SU (Urkund) and DSV (comparing text chunks between exam answers)

Stockholm University

Methods

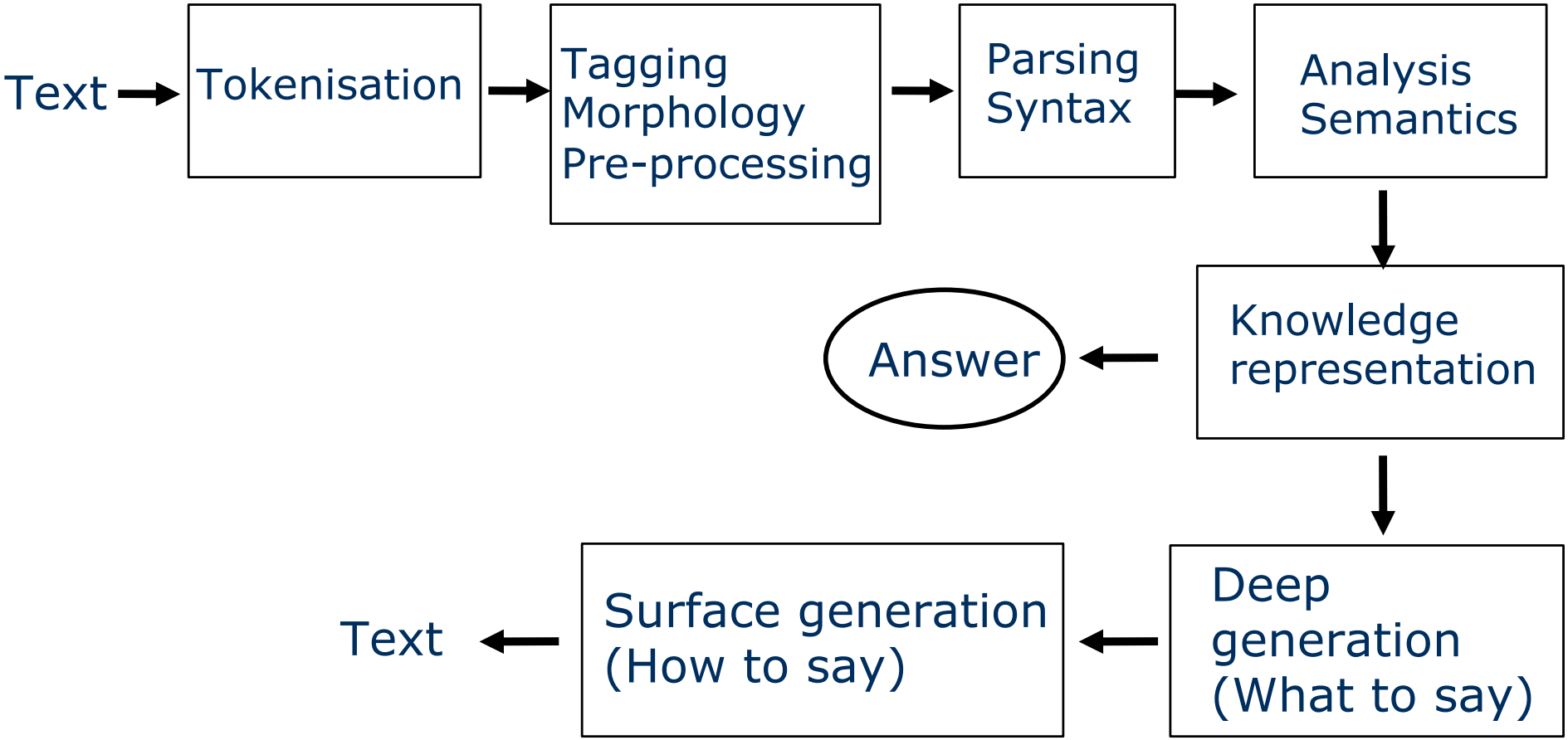
- Fingerprints – n-grams
- String matching
- Bag of words – Vector space
- Stylometrics – Statistical matching
- Citation analysis

Fake new detection

- Tricky to decide
- News should be objective, but are never really.
- News with responsible publishers / ansvarig utgivare
- Satire or fake news?
- Easier to see if trolls comment on the fake news

Pipeline NLP

Stockholm University



Master thesis proposals I

- Creating and evaluating an explanation based Computer-Assisted Coding (CAC) tool for ICD-10 for Swedish discharge summaries.
- Building and evaluating an automatic discharge summary system for Swedish patient record notes
- Generating synthetic training text from Swedish electronic patient records
- Creating a synthetic Health Bank database using machine learning methods

Stockholm University

Master thesis proposals II

- Build and evaluate a machine translation system based on Moses http://www.statmt.org/moses/ or ModernMT, https://github.com/modernmt/modernmt in one specific domain.
- Build and evaluate an automatic essay grading scoring system
- Build an automatic radiology report generation system based on an X-ray input.

Master thesis proposals III

- Prove that a clinical deep learning language model don’t leak the privacy of patients or - at least calculate the leakage.

Stockholm University

Quizzes for the lectures I

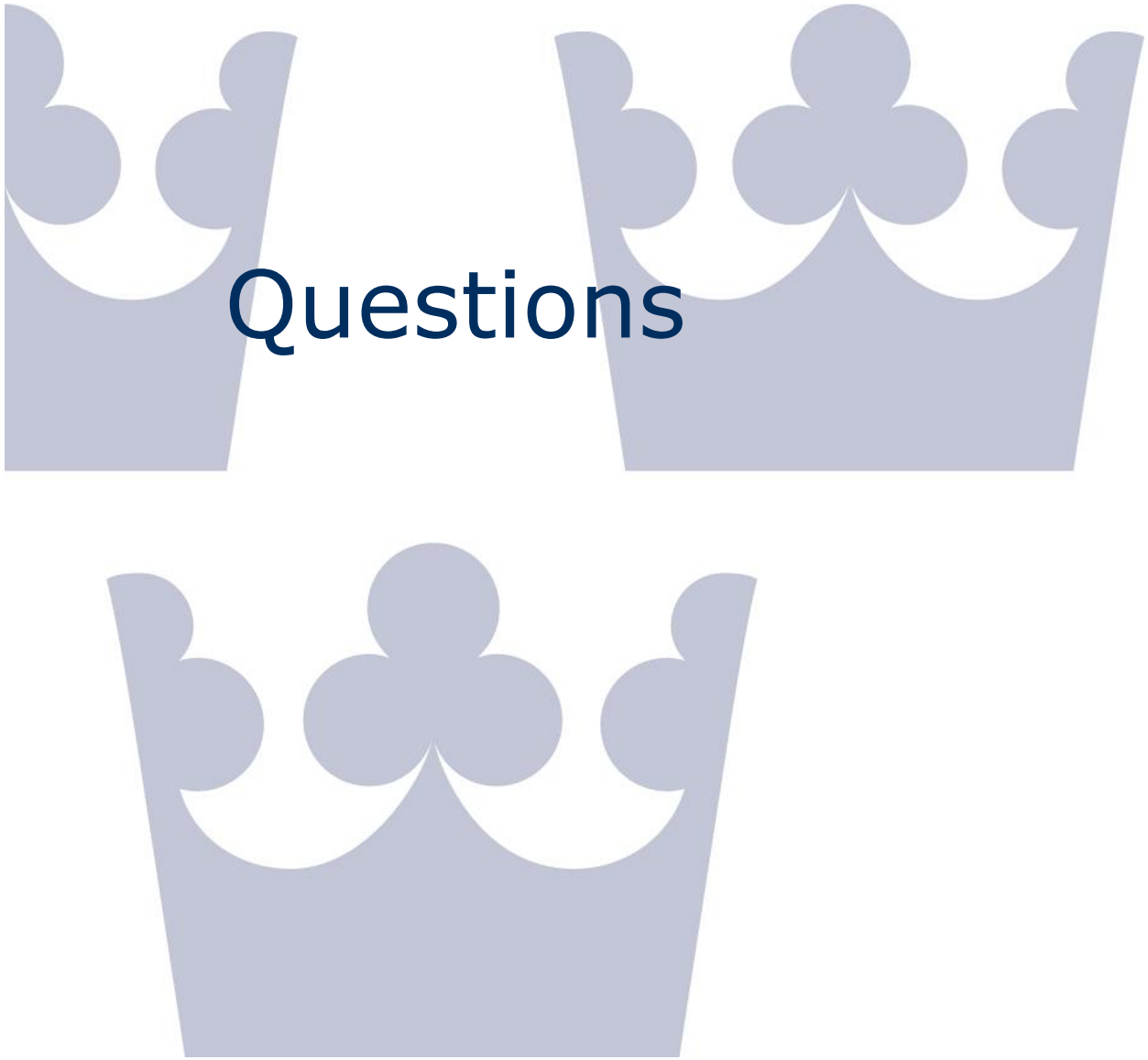
- Why is NLP so hard?
- Why are NLP systems so bad on some tasks
  - Think about different domains
- When does NLP system fail?
- What is stemming?
- Would you prefer a search engine with stemming or no stemming?

Stockholm University

Quizzes for the lectures II

- What are stop words?
- What is text summarisation?
- What is named entity recognition?
- What are features in machine learning?

Stockholm University



Questions