COMP4650/6490 Document Analysis

An Overview of NLP Tasks

Dawei Chen

Software Innovation Institute

ANU School of Computing

Australian National University

Software Innovation Institute

Administrative Matters

COMP4650 representative nomination

- Email the convener by 4 August

Quiz 0

- Self-assessment, ungraded
- Closes on Thursday 31 August (semester 2 census date)

ALTA Workshop

- Details are on the last slide

Outline

1. Text Preprocessing
2. Language Formalisation & Understanding
3. NLP Applications

Outline

1. Text Preprocessing
2. Language Formalisation & Understanding
3. NLP Applications

Text Preprocessing

NLP tasks for preprocessing text:

- Tokenisation
- Stopwords removal
- Token normalisation: Stemming, Lemmatisation, etc.
- Sentence splitting

Tokenisation

- The task of dividing text into tokens (e.g., words, numbers, punctuation marks, and other symbols)
- Example
  - Input: `My sister didn't call me.`
  - Output: `my`, `sister`, `did`, `not`, `call`, `me`

Stopwords Removal

- Stop words usually refer to the most common words in a language e.g., `the, a, an, and, or, will, would, could`
- These words are not very useful in keyword search
- Removing stopwords reduces the number of postings that an IR system has to store

Stemming

- The task of turning tokens into stems
- Example: {`run`, `runs`, `running`} ⇒ `run`
- Stemming is usually a crude heuristic process that strips off suffixes,
  e.g., `studies` ⇒ stem: `studi`, suffix: `es`
- Stems need not be real words

Lemmatisation

- The task of turning words into lemmas (i.e., entries in a dictionary)
- Example: `better` ⇒ good
- Requires knowledge of the context (typically the intended Part-of-Speech of a word in the context), e.g.,
  - `meeting` ⇒ `meet` (Verb)
  - `meeting` ⇒ `meeting` (Noun)

Additional Normalisation Tasks

- Keep equivalence class of terms:
  `U.S.A = USA = united states`
- Synonym list: `car = automobile`
- Capitalisation: `ferrari` $\Rightarrow$ `Ferrari`
- Case-folding: `Automobile` $\Rightarrow$ `automobile`, `ACT` $\neq$ `act`

Sentence Splitting

- The task of segmenting text into sentences
- Involves detecting boundaries of sentences, e.g., using punctuation such as period (`.`), question mark (`?`), and exclamation point (`!`)
- Example:
  `Dr. Watson stroked his moustache.`
  `“You reminded me,” he remarked, “of your mother.”`

Outline

1. Text Preprocessing
2. Language Formalisation & Understanding
3. NLP Applications

Language Formalisation & Understanding

NLP Tasks that help with language formalisation and undertanding:

- Part-of-Speech (POS) tagging
- Named entity recognition (NER)
- Parsing: Syntactic, Semantic, Discourse
- Natural language inference (NLI)
- Relation extraction
- Coreference resolution
- Word sense disambiguation (WSD)
- Language modelling

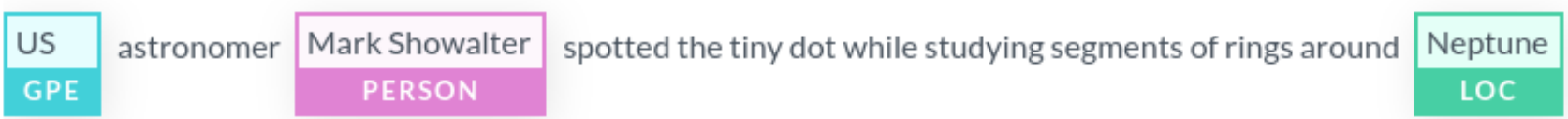
Part-of-Speech (POS) Tagging

- The task of assigning grammatical categories (POS-Tags) to tokens
- Part-of-Speech refers to the syntactic role of each token in a sentence
- Example (using Penn Treebank POS Tags):

She	eats	like	a	vegetarian.
Pronoun	Verb	Preposition	Determiner	Noun
PRP	VBZ	IN	DT	NN

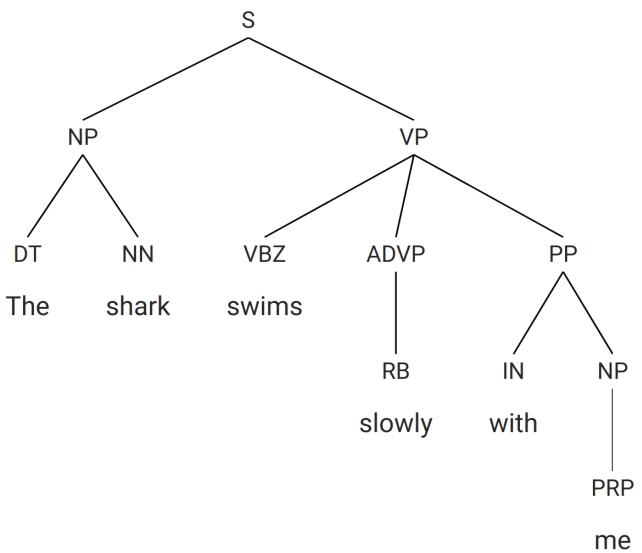
Named Entity Recognition (NER)

- The task of finding and classifying named entities in sentences
- Help other NLP tasks, e.g., syntax parsing, relation extraction, machine translation
- Example

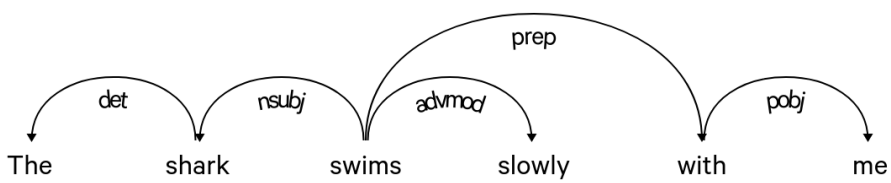


Syntactic Parsing

- Extracting syntactic structure (tree or forest) from text
- Constituency parsing: Phrases represented as nodes in a tree
- Dependency parsing: Dependencies between words
- Example: `The shark swims slowly with me.`



(a) Constituency tree



(b) Dependency tree

Semantic Parsing

- The task of transforming sentence into a meaning representation, e.g., a logical formula
- Example
  - Input: `show me flights tomorrow from ci0 to ci1`
  - Output: `(lambda $0 e(and (flight $0 (from $0 ci0) (to $0 ci1) (tomorrow $0)))`

Semantic Role Labelling (Shallow Semantic Parsing)

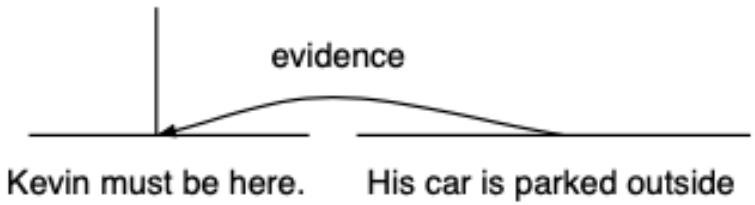
- The task of finding the semantic roles of the arguments of each predicate in a sentence
- A predicate with *n* arguments:

  `predicate(arg1, arg2, ···, argn)`
- Example

  - AGENT: The volitional causer of an event
  - GOAL: The destination of an object of a transfer event
  - THEME: The participant most directly affected by an event

Discourse Parsing

- The task of identifying the rhetorical structure (tree or graph) of documents or sentences
- A discourse is a coherent structured group of sentences (not a random collection of sentences)
- Discourse relations: background, contrast, evidence, purpose, etc.
- Example



Natural Language Inference (NLI)

- NLI (textual entailment) is the task of predicting if the premise sentence entails the hypothesis sentence
- 3-class classification: Entailment, Contradiction, Neutral
- Example (Entailment)
  - Premise: `I have never seen a hummingbird.`
  - Hypothesis: `I have never seen a hummingbird not flying.`

Relation Extraction

- The task of finding and classifying semantic relations among entities
- Create new knowledge, augment current knowledge bases, support other NLP tasks (e.g., question answering)
- Example

Relations	Types	Examples
Physical-Located	PER-GPE	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**XYZ**, the parent company of **ABC**
Person-Social-Family	PER-PER	**Yoko**'s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...

Coreference Resolution

- The task of determining if two mentions refer to the same entity
- Mention (or referring expression): text span that mention an entity
- Example:

```
The trucks shall treat the roads before they freeze.
```

Word Sense Disambiguation (WSD)

- The task of selecting the correct sense for a word in text
- Word sense: A discrete representation of one aspect of the meaning of a word
- Example: `bank`
  - A financial institution

    `The Commonwealth Bank of Australia`
  - The slope beside a body of water

    `The bank of the Murrumbidgee`

Language Modelling

- The task of modelling the probability distribution over sequences of words
- A language model can compute the probability of a sequence of words (or predict upcoming words from prior word context)
- Example: $P(\texttt{I want to eat Asian food})$
  Apply the chain rule of probability

$$
\begin{aligned}
&P(\texttt{I want to eat Asian food}) \\
&= P(\texttt{I})\, P(\texttt{want} \mid \texttt{I})\, P(\texttt{to} \mid \texttt{I want})\, P(\texttt{eat} \mid \texttt{I want to}) \\
&\quad P(\texttt{Asian} \mid \texttt{I want to eat})\, P(\texttt{food} \mid \texttt{I want to each Asian})
\end{aligned}
$$

- Large (neural) language models pre-trained on vast amounts of textual data can learn to perform a broad range of NLP tasks without explicit supervision

Outline

1. Text Preprocessing
2. Language Formalisation & Understanding
3. NLP Applications

NLP Applications

- Information retrieval & question answering
- Text categorisation
- Machine translation
- Text summarisation
- Text generation
- ...

Information Retrieval (IR) & Question Answering (QA)

- IR is the task of finding documents that satisfies an information need from within large collections[1]
- IR-based QA: using IR to find relevant text (on the web)
- Knowledge-based QA: through querying databases of facts (e.g., performing inference in a knowledge base)
- Querying a pre-trained language model (with/without using IR)

Text Categorisation

- Classification tasks that assign a label or category to an entire text or document
- Sentiment analysis: classifying text into three classes – POSITIVE, NEGATIVE or NEUTRAL
- Spam detection: assigning SPAM or NOT-SPAM to an email
- Language id: determining the language the text is written in
- Authorship attribution: determining the author of text

[1] Manning, Raghavan, and Schütze 2009.

Machine Translation (MT)

- The task of automatically translating text from one language to another
- Aiding human translators by producing a draft translation (computer-aided translation)

Text Summarisation

- The task of creating a summary (representing the main points) of text
- Can be approached by extracting content from the original text or formulated as a text generation task

Text Generation

- The task of producing text conditioning on some other text
- Many NLP applications involve text generation, e.g., question answering, machine translation, text summarisation, conversational dialogue systems (i.e., chatbots), image captioning, code generation, ...
- Text generation is one of the essential capabilities of generative artificial intelligence (AI) systems

Summary

We introduced a number of typical NLP tasks:

1. Text Preprocessing
   - Tokenisation, Stopwords removal
   - Token normalisation, Sentence splitting
2. Language Formalisation & Understanding
   - POS tagging, NER
   - Parsing: Syntactic, Semantic, Discourse
   - NLI, Relation extraction, Coreference resolution
   - WSD, Language modelling
3. NLP Applications
   - Information retrieval & question answering
   - Text categorisation, Machine translation
   - Text summarisation, Text generation

Some of these tasks will be covered in this course.

ALTA Workshop

The 21st Annual Workshop of the Australasian Language Technology Association (ALTA'23)

Dr. Gabriela Ferraro

https://alta2023.alta.asn.au/

Extra course marks for participants (a tradition of this course)

- 2 marks for accepted paper(s)
- 3 marks if ranked in the top-3 of shared-task (i.e., language technology programming competition)
- 3 marks maximum