THE UNIVERSITY OF MELBOURNE

Comp90042 Workshop Week 10

Table of Contents

1. Machine translation

2. Information extraction

1. Machine Translation

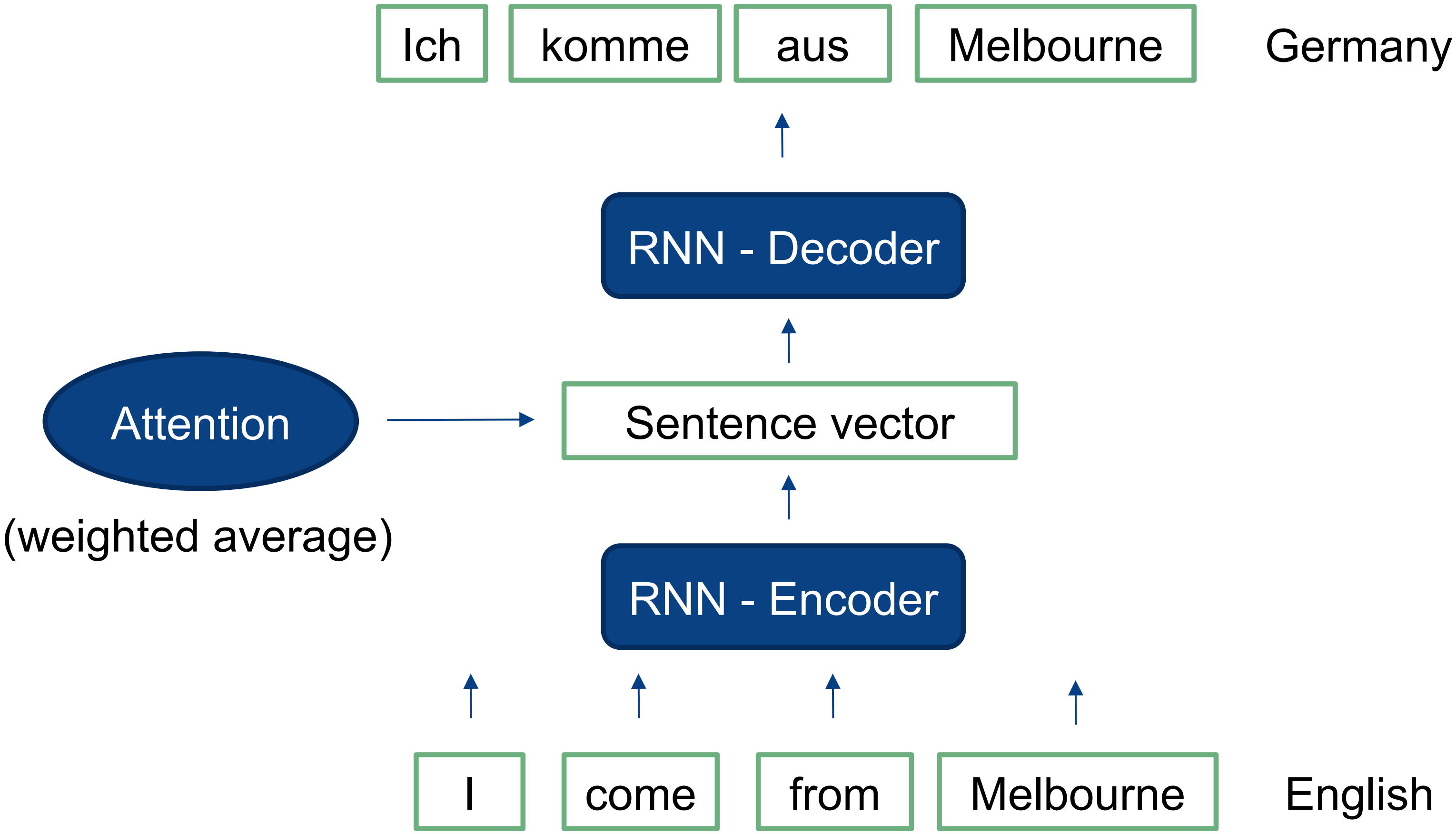
MT

- Statistical
  - Pre deep learning era.
  - Complex, lot of feature engineering.
  - Word-based(phrase-based) approach.

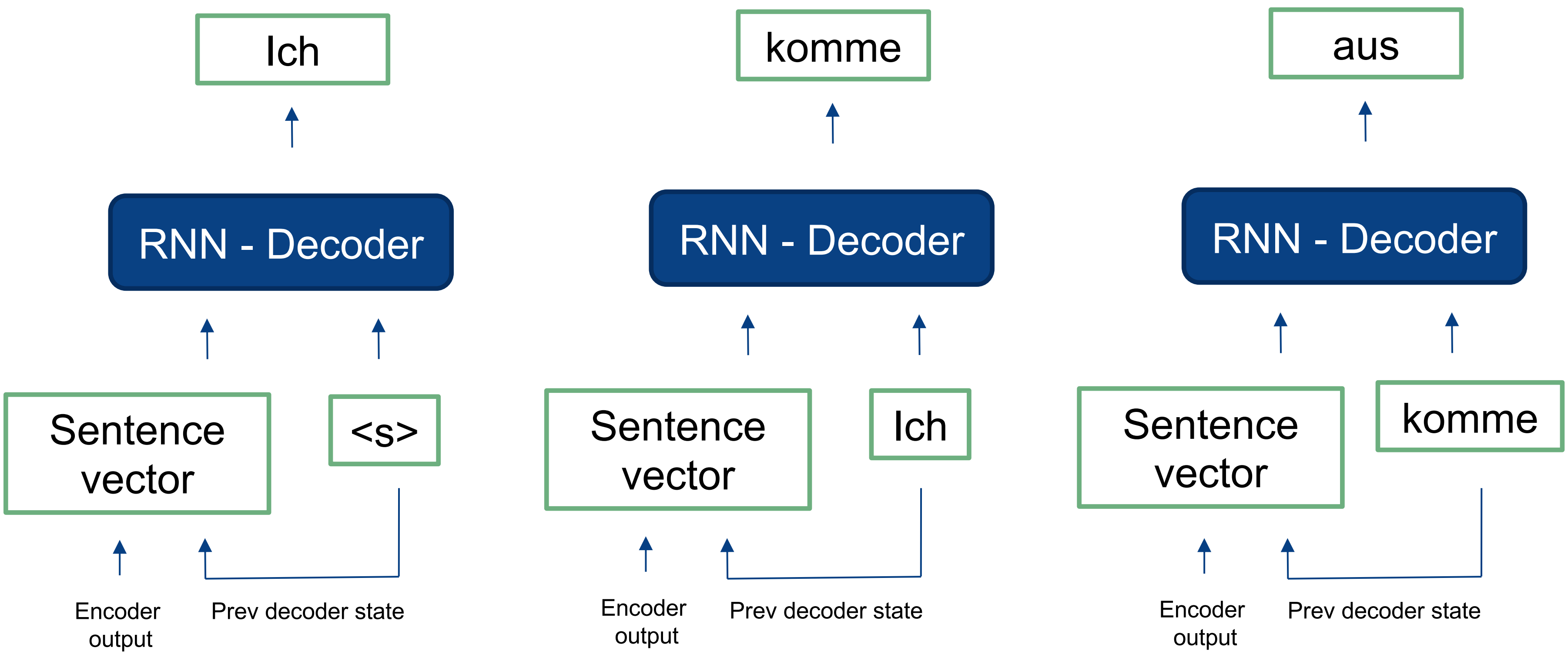
    $$\text{argmax}_e P(e \mid f) \blacktriangleright \text{argmax}_e P(f \mid e) P(e)$$

- Neural Network
  - Cleaner, less feature engineering
  - Encoder-Decoder framework
  - End-to-End
  - Use RNN / Transformer

1. Machine Translation (Encoder-Decoder)

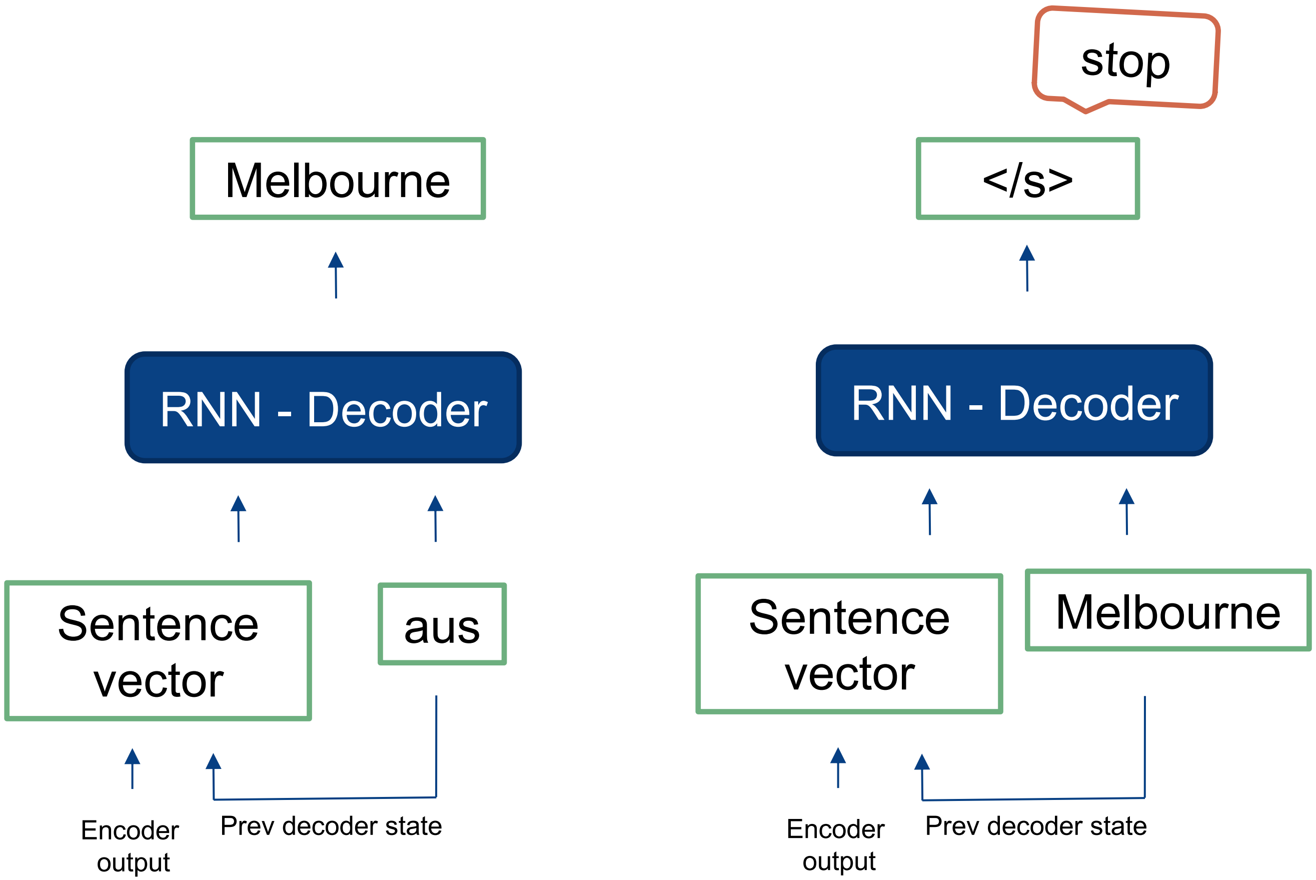


1. Machine Translation (Generation)



Note: Sentence vector is dynamic if you use Attention!

1. Machine Translation (Generation)



Note: Sentence vector is dynamic if you use Attention!

1.Transformer

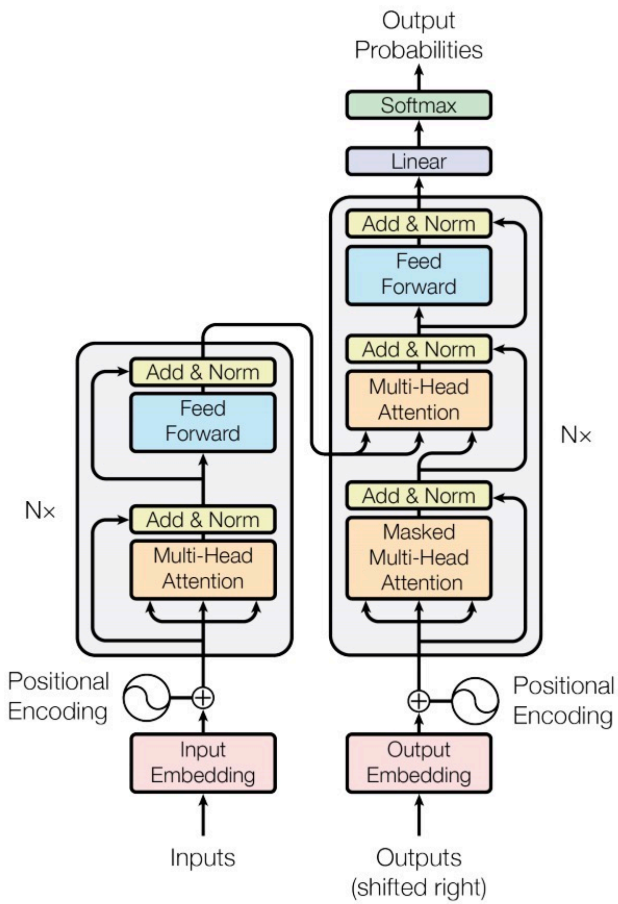


Figure from: Attention is all you need

1. Machine Translation

Q1 What aspects of human language make automatic translation difficult?

1. Lexical complexity
2. Morphology (E.g. English – Turkish, English - German)
3. Syntax (E.g. English - Japanese)
4. Semantics
5. Requires parallel corpus which is hard to obtain, especially for low-resource language pairs.

Morphology

Gender inflectional

- Translating from genderless language to gendered language

English The doctor asked the nurse to help her in the procedure

doctora

Spanish El doctor le pidio a la enfermera que le ayudara con el procedimiento

Example from Stanovsky et al., ACL 2019

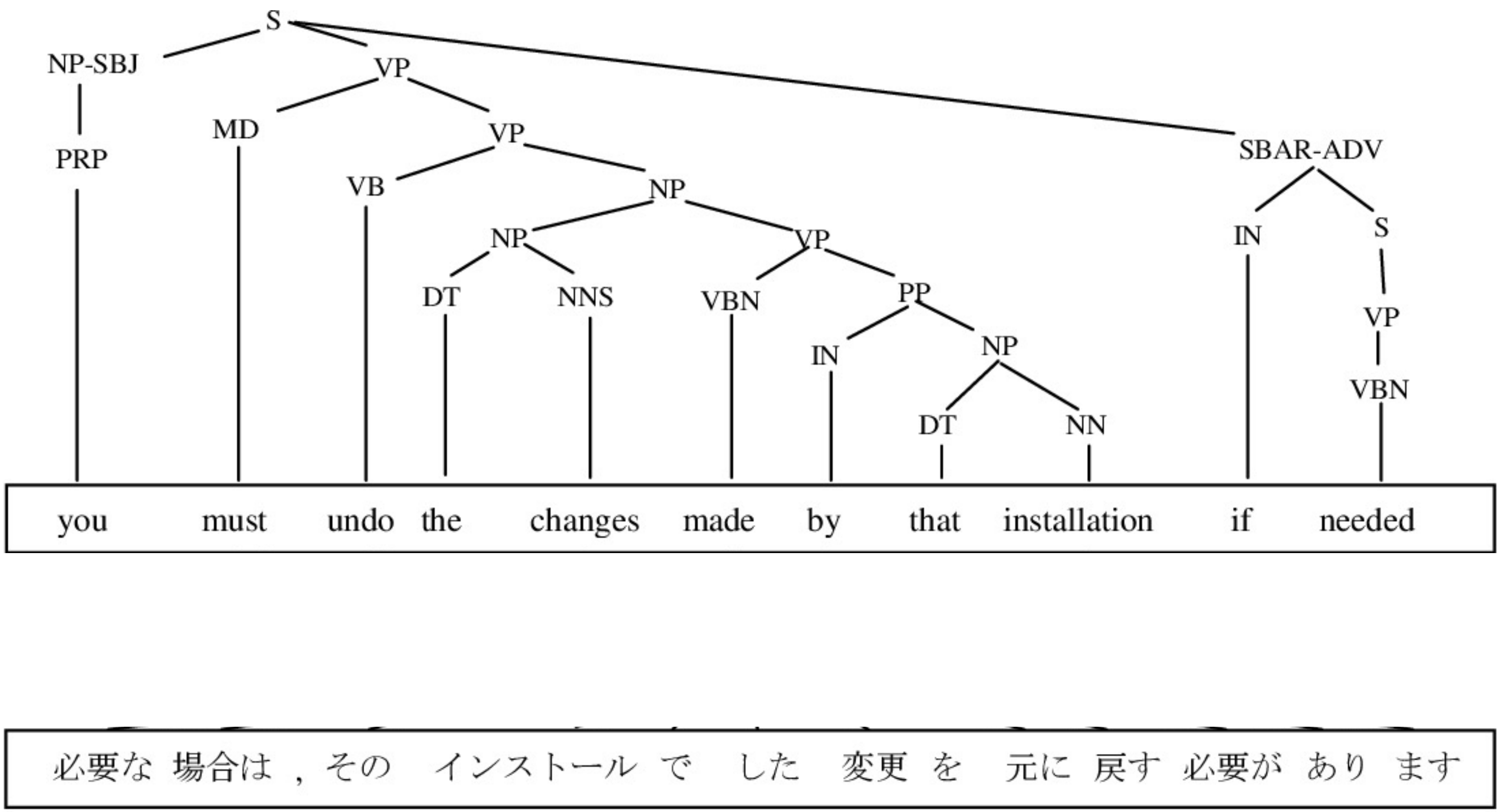
English Marie Curie was a physicist.

German Marie Curie war Physiker.

Physikerin

Example from Wang et al., ACL 2022

Syntax


S
NP-SBJ
VP
PRP
MD
VP
SBAR-ADV
VB
NP
NP
VP
IN
S
DT
NNS
VBN
PP
VP
IN
NP
VBN
DT
NN
you must undo the changes made by that installation if needed
必要な 場合は ， その インストール で した 変更 を 元に 戻す 必要が あり ます

Low-resource language pairs

2. Information Extraction

Q2	What is Information Extraction?

Extract information from a (generally unstructured) document, into a structured format

Example:
The University of Tokyo, abbreviated as Todai, is a public research university located in Bunkyo, Tokyo, Japan.

2. Information Extraction

Abbreviation

The University of Tokyo, abbreviated as Todai, is a public research university located in Bunkyo, Tokyo, Japan.

Capital

Location

1st step:
Named Entity
Recognition

- Find the name entities
- Sequence Model: RNN, HMM, CRF

2nd step:
Relation
Extraction

- Find relations between two entities.
  E.g. "Tokyo" vs "Japan”
- Mostly classifiers

2. Information Extraction

Q2	What is Information Extraction?

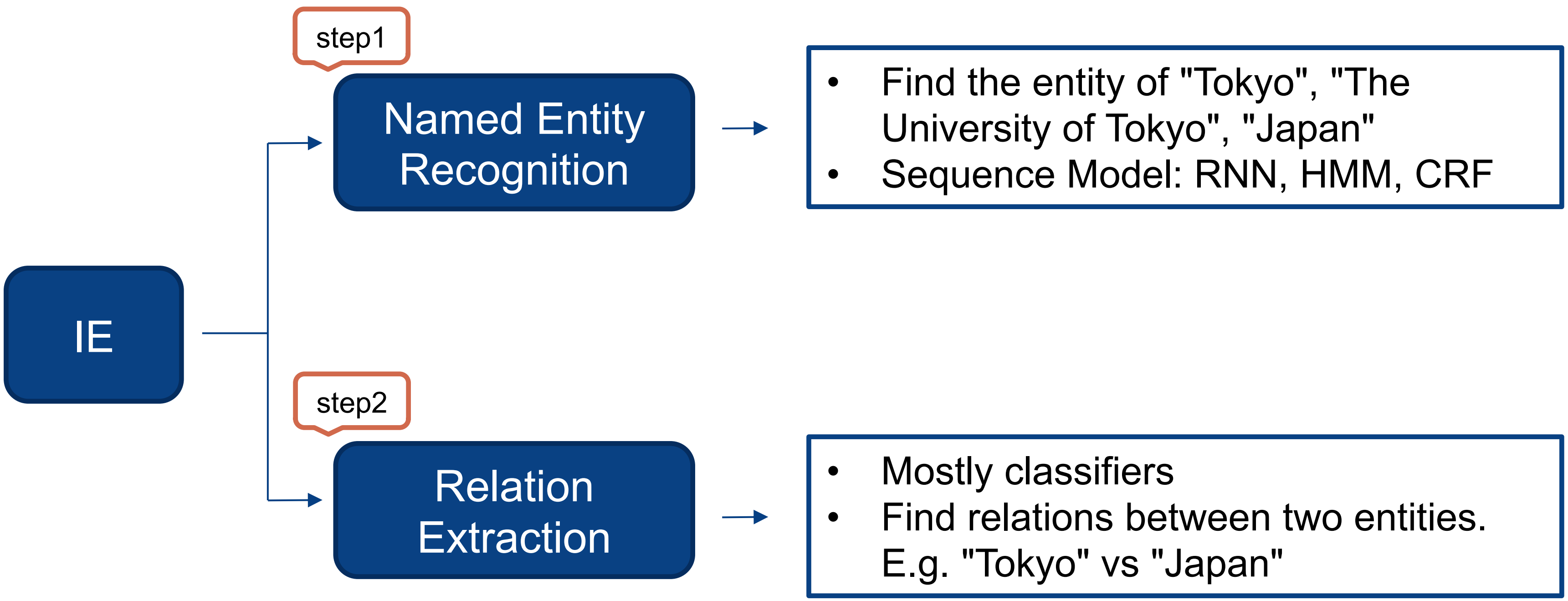
Extract information from a (generally unstructured) document, into a structured format

Example:
The University of Tokyo, abbreviated as Todai, is a public research university located in Bunkyo, Tokyo, Japan.

Result:
- Capital(Japan, Tokyo)
- Location(The University of Tokyo, Japan)
- Location(Todai, Japan)

2. Information Extraction

How to perform IE?



2. Information Extraction

Q2A	What is NER? Why is it difficult?

NER = Named Entity Recognition

- Find named entities within a document.
- Some types of named entities:
  1. proper nouns (Names, Location, Organization)
  2. Times
  3. Numerical values
- Ambiguous
  - Common nouns vs proper nouns —> apple vs Apple
  - People's name vs location —> Philip is in Philip Island
  - Organization vs location —> New York Times vs New York

2. Information Extraction

Q2A What is NER? Why is it difficult?

NER = Named Entity Recognition

- Find named entities within a document.
- Some types of named entities:
  1. proper nouns (Names, Location, Organization)
  2. Times
  3. Numerical values
- Ambiguous
  - Common nouns vs proper nouns —> apple vs Apple
  - People's name vs location —> Philip is in Philip Island
  - Organization vs location —> New York Times vs New York
- Solve location ambiguous
  - **gazetter -** create a (somewhat) exhaustive list of names of places.
  - List of names of people? – can’t, constantly changing

2. Information Extraction

Q2B	What is **IOB** trick in a sequence labelling? Why is it important?

Motivation: Named entities can consist of **more than** 1 token

Ex: [The University of Melbourne] is the best Australian University.

4 letters.

We indicate whether a given token is Beginning a named entity, Inside a named entity, or Outside a named entity.

The-B-LOC University-I-LOC of-I-LOC Melbourne-I-LOC is-O the-O best-O Australian-O University-O .-O

2. Information Extraction

Chelsea denied Pep Guardiola's Manchester City in Porto.

Preprocessing:

1. ~~Lowercase?~~
2. ~~Remove stop words?~~
3. Tokenize?

apple vs. Apple

The University of Melbourne

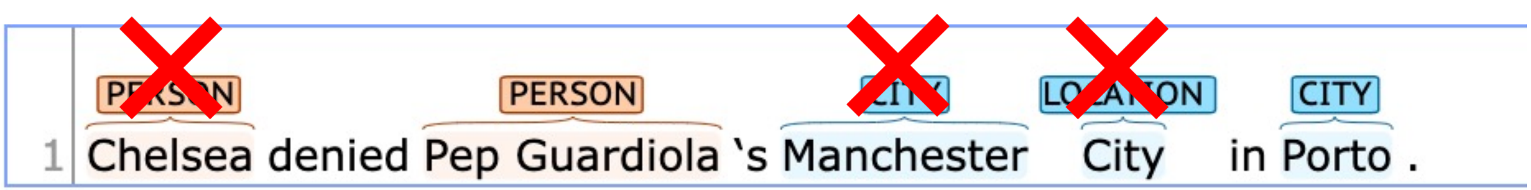
2. Information Extraction

PER: people, characters
ORG: companies, sports teams
LOC: regions, mountains, seas

IO tagging

IOB tagging

IO tagging	I-ORG	O	I-PER	I-PER	O	I-ORG	I-ORG	O	I-LOC	O
	Chelsea	denied	Pep	Guardiola	‘s	Manchester	City	in	Porto	.
IOB tagging	B-ORG	O	B-PER	I-PER	O	B-ORG	I-ORG	O	B-LOC	O



from https://corenlp.run/

2. Information Extraction

	我	住在	北京市	朝阳区	。
IO Tagging	O	O	I-LOC	I-LOC	O
IOB Tagging	O	O	B-LOC	B-LOC	O

2. Information Extraction

Q2C What is **Relation Extraction?**

- Attempt to find relationships between entities in a text

  Ex: Harry Potter vs JK.Rowling —> relation Author

- It is done after obtaining entities (the NER tags)

2. Information Extraction

Q2D | Why are hand–written patterns generally inadequate for IE, and what other approaches can we take?

Why?

- Too many different ways of expressing the same information.
- High precision ☺
- But low recall ☹

I visited The University of Melbourne in Melbourne
I visited Melbourne in Australia
The Victorian Library is located in Melbourne

Rule: A in B(LOC) -> Location(A,B)

Location(The University of Melbourne, Melbourne) ✓
Location(Melbourne, Australia) ✓

Location(The Victorian Library, Melbourne) MISSED

Precision = 2/2 = 1.0
Found two, two are correct

Recall = 2/3
Total three, only found two

2. Information Extraction

Q2D	Why are hand–written patterns generally inadequate for IE, and what other approaches can we take?

Other approach:

- **Parsing** the sentence might lead a more systematic method of deriving all of the relations, but language variations mean that it's still quite difficult.

- Frame the problem as **supervised machine learning**, with general features (like POS tags, NE tags, etc.)

- **Bootstrapping patterns** — using known relations to derive sentence structures that describe the relationship.
  - Rule: A in B(LOC) -> Location(A,B)
  - Location(Melbourne, Australia)
  - Search sentences contain Melbourne and Australia, store them as new rules

Next

iPython `11-machine-translation`

Whats inside?

- Encoder-Decoder for machine translation (Char-level)
- Use Colab -- faster

To do?

- Modify to use GRU
- Modify to do translation at the word-level:
    - French & English tokenizer (SpaCy)
    - Create vocab
    - Replace low frequency with UNK
    - Change the corpus reading function
    - Updating training and inference to use the vocab to look up words
    - Apply attention mechanism (if you have time)