Topic 10: Machine Translation

Linguistics module

Teangacha Iar-bhunscoile Éireann
Post-Primary Languages Ireland

AILO
All Ireland Linguistics Olympiad

ADAPT
Engaging Content
Engaging People

PPLI delivering
Languages Connect

Supported by
An Roinn Oideachais
Department of Education

Overview of this linguistics module

- Topic 1 Introduction to areas of linguistics and problem-solving
- Topic 2 Historical Linguistics
- Topic 3 Phonetics
- Topic 4 Sociolinguistics
- Topic 5 Writing systems
- Topic 6 Language Acquisition
- Topic 7 Morphology
- Topic 8 Syntax
- Topic 9 Psycholinguistics / Neurolinguistics
- **Topic 10 Machine Translation**

What is Machine Translation (MT)?

- Machine Translation means converting text or speech in one language (***source, "s"***) to a text in another language (***target, "t"***).

- MT is a subfield of Natural Language Processing (NLP) which deals with the use of computers to model and process human language

"This translation app isn't working."

How does Machine Translation (MT) work?

- Early MT systems: rules and dictionaries
  - manual work carried out by language experts
  - It took months to develop a new system

- Modern systems: learn from parallel texts
  - based on the probability that a target text “t” is a translation of a source text “s” P(t|s)
  - no language knowledge is needed
  - “only” large amounts of parallel texts
  - it takes hours to develop a new system

  - Statistical machine translation (SMT) from 1990 till 2016
  - Neural machine translation (NMT) since 2016

How computers translate (Rule-based and SMT)

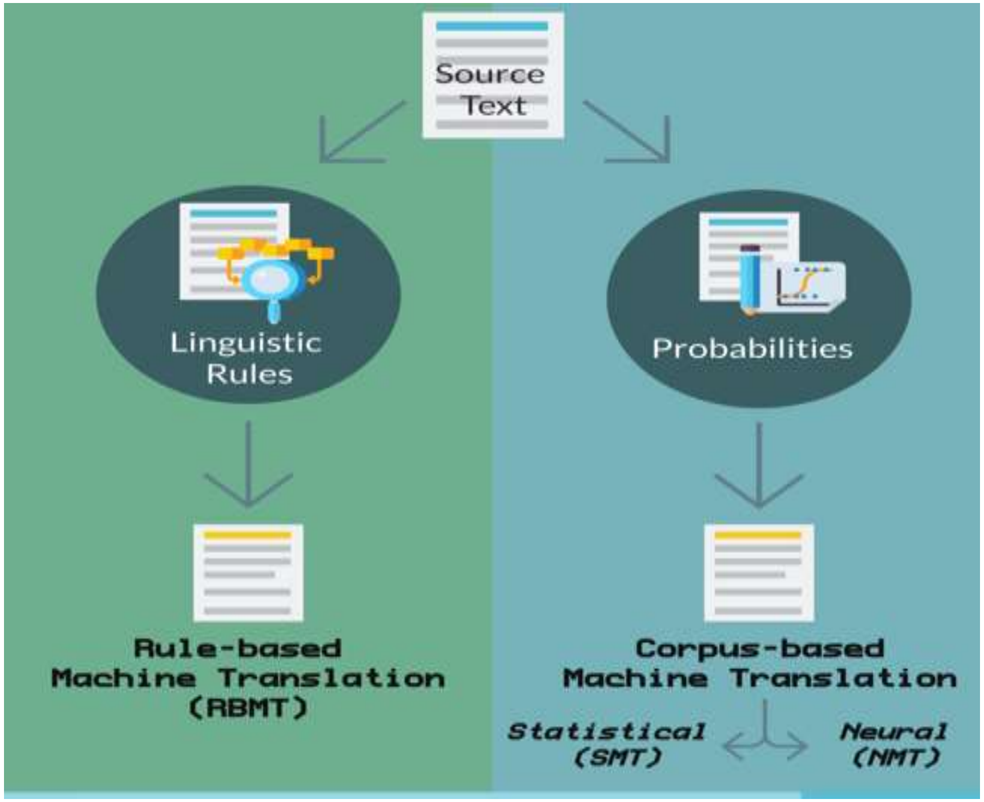
How modern machine translation systems work these days?

- Modern systems:
    - learn the probability P(t|s) from parallel texts

- Statistical machine translation (SMT)
    - Learns this probability from frequencies of words and word groups in data

- In the last 6 years: Neural Machine Translation (NMT)
    - learns this probability as a complex mathematical function provided by neural networks

Machine Translation



Task 10.1
Learning from parallel texts

What are “parallel texts”?

- Texts containing the same content in two (or more) different languages
- Sentence alignment is necessary for current MT systems (each line in one language corresponds to the line in another language)
- Large amounts of parallel texts are needed for MT systems to learn to translate (millions of sentences)

10.1.1 An example of English-Spanish parallel text (0n worksheet 10.1 also)

English	Spanish
Torres and associates.	Torres y asociados.
Carlos Torres has three associates.	Carlos Torres tiene tres asociados.
His associates are not strong.	Sus asociados no son fuertes.
Torres has a company, too.	Torres también tiene una empresa.
His clients are angry.	Sus clientes están enfadados.
The associates are also angry.	Los asociados también están enfadados.
The company has three groups.	La empresa tiene tres grupos.
The groups are in Europe.	Los grupos están en Europa.
The modern groups sell strong pharmaceuticals.	Los grupos modernos venden medicinas fuertes.
The groups do not sell aspirin.	Los grupos no venden aspirina.

Learning to translate from parallel texts

- Try to translate the following sentence into Spanish using the given parallel text:

Clients do not sell pharmaceuticals in Europe.

- Does the parallel text on the previous slide provide enough information?

Learning to translate from parallel texts

- Try to translate the following sentence into Spanish using the given parallel text:

Clients do not sell pharmaceuticals in Europe.

- Does the given parallel text on the previous slide provide enough information to translate this sentence?

- **Yes! All necessary information can be found in the given parallel text and the translation is:**

Clientes no venden medicinas en Europa .

Learning to translate from parallel texts

- What about this sentence?

The pharmaceuticals are very strong.

Learning to translate from parallel texts

- What about this sentence?

The pharmaceuticals are very strong.

- **the word "very" does not appear in the parallel text, so the text does not provide the information about the corresponding Spanish word**

Las medicinas son ?? fuertes.

10.1.2 Another parallel text: English-Croatian

English	Croatian
A mouse.	Miš.
The mouse.	Miš.
A cat.	Mačka.
The cat.	Mačka.
A cat chases a mouse.	Mačka juri miša.
A cat chases a mouse.	Miša juri mačka.
A mouse chases a cat.	Miš juri mačku.
A mouse chases a cat.	Mačku juri miš.
The cat is with the mouse.	Mačka je sa mišem.

- Fewer rules for word order (syntax) in Croatian
- Richer morphology (many different word forms)
- No articles

Learning from parallel texts is now more difficult

- for translating the following English sentence into Croatian:

The mouse is with the cat.

Learning from parallel texts is now more difficult

- for translating the following English sentence into Croatian:

The mouse is with the cat.

- **All words are occurring in the given parallel text**
- **but not all possible Croatian word *forms* are there**

- **The translation is:**

Miš je sa mačkom.

- **and the form “mačkom” is not in the parallel text.**

Learning from parallel texts

- the more different words in parallel texts, the better translation
- the more different word forms in parallel text, the better translation
- the more different structures in parallel text, the better translation

therefore, parallel texts for learning have to be large

Why is machine translation hard?
– morphology and syntax –

- different syntax (sentence structure) in different languages
- different morphology (number of word forms, rules) in different languages
    - NMT systems handle these differences quite well, if they have learnt from large parallel texts
    - there are still some errors, though

Why is machine translation hard?
– availability and size of parallel texts –

- large parallel texts are not available in all languages
  - actually, only for a small set of language pairs

- also not for all types of texts
  - for example, there are many for news, movie subtitles, but almost none for social media texts

Why is machine translation hard?
– ambiguity –

a large number of words (even phrases or sentences) in a language are ambiguous:

they can mean different things depending on the context

I got it!

I received it?

I bought it?

I understood it?

Teangacha Iar-bhunscoile Éireann
Post Primary Languages Ireland

AILO

Languages Connect

Why is machine translation hard?
– ambiguity –

ambiguous sentences:

"Minister accused of having 8 wives in jail"

"A man saw a woman with a telescope"

for computers, even

"A man saw a dog with a telescope"

can be ambiguous

Why is machine translation hard?
– ambiguity –

- NMT translates much better than SMT
- mainly because of syntax and morphology

- ! ambiguity is still a challenge

Why is machine translation hard?
– evaluation –

- there is no single correct translation of a sentence
  => even evaluating machine translation is not trivial

  - human evaluation:
    requires time, effort and experienced evaluators

  - automatic evaluation:
    measures can calculate a score based on a “correct” translation (or a set of them)
    fast, but cannot provide all necessary information
    especially qualitative info (what is wrong, and why)

Why is machine translation hard? – evaluation –

- Users (and journalists) devise their own evaluations, almost always flawed
    - Expecting the system to translate idioms, or jokes - not a fair test
    - “Back and forth” translation, ie translate into a language and then translate the result back to English
        - not a fair test: if the back-translation is good there is no guarantee the target translation was good - it could be just word-for-word garbage in both directions
        - if the translation is bad you don’t know where it went wrong - on the outward trip or on the way back?
- a good way to assess the usefulness is to use it (!), for example to translate a web page from a language you don’t know, and see how much of the result is understandable/useful
- But translating INTO a language you don’t know is risky

Exercise 10.1.3 Learning from parallel texts (⅓)

Translate this sentence in Centauri below, into Arcturan:

farok crrrok hihok yorok clok kantok ok-yurp

How might *you* translate between two languages you know **nothing about**?!

Use the parallel text on the next slide!

Exercise 10.1.3 Learning from parallel texts (⅔)

Centauri	Arcturan
ok-voon ororok sprok .	at-voon bichat dat .
ok-drubel ok-voon anok plok sprok .	at-drubel at-voon pippat rrat dat .
erok sprok izok hihok ghirok .	totat dat arrat vat hilat .
ok-voon anok drok brok jok .	at-voon krat pippat sat lat .
wiwok farok izok stok .	totat jjat quat cat .
lalok sprok izok jok stok .	wat dat krat quat cat .
lalok farok ororok lalok sprok izok enemok .	wat jjat bichat wat dat vat eneat .
lalok brok anok plok nok .	iat lat pippat rrat nnat .
wiwok nok izok kantok ok-yurp .	totat nnat quat oloat at-yurp .
lalok mok nok yorok ghirok clok .	wat nnat gat mat bat hilat .
lalok nok crrrok hihok yorok zanzanok .	wat nnat arrat mat zanzanat .

Exercise 10.1.3 solution (3/3)

Centauri sentence:
farok crrrok hihok yorok clok kantok ok-yurp

Arcturan words:

in the best order (according to the Arcturan part of the parallel text):

{jjat, arrat, mat, bat, oloat, at-yurp}

Task 10.2
Ambiguity

Exercise 10.2 Ambiguity

- Ambiguity AILO puzzle Running on MT puzzle (and Solution)

- One of the most common errors in the modern Neural Machine Translation systems is word sense selection: the source language text may contain words which have multiple meanings and the MT system has chosen the wrong one.

- In the puzzle, the effect of this has been simulated: we have taken an ordinary English text and replaced a number of individual words with alternative words which share a meaning with the original word, but which are not correct in this context. For example, in the first line, we have "angry-legged" instead of "cross-legged".

- Your task is to find the incorrect words and their correct replacements

Task 10.3 Gender Bias in Language

Exercise 10.3 Gender Bias in Language (1/3)

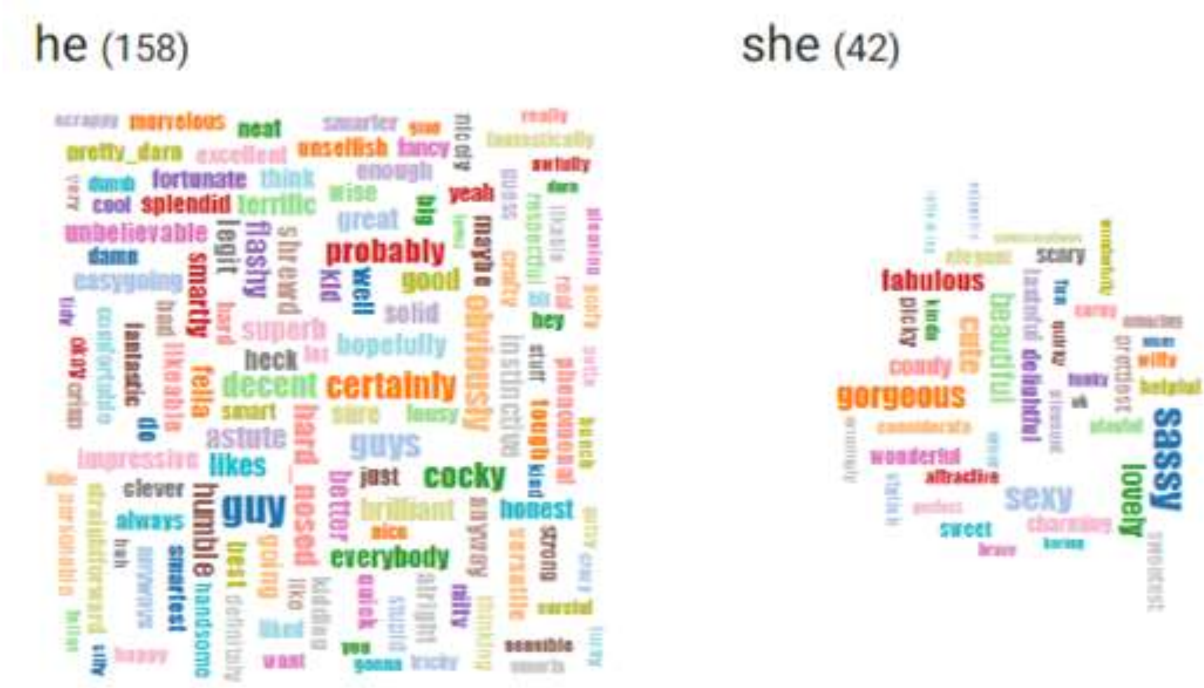
- When you hear about a ‘nurse’, it may be that a woman pops into your mind. But we all know there are male nurses also.
- Similarly, there are many female doctors.

- These stereotypes can also be found in Machine Translation data which lead to biased translations, and researchers are working to change this.

Exercise Gender Bias in Language (2/3)

- Have a look at http://wordbias.umiacs.umd.edu/ and see if there are words that are commonly associated to a particular gender which you hadn't considered before.

- Example searches: nurse, doctor, pilot, caring, smart, happy



Exercise Gender Bias in Language (3/3)

- Have a look into these sentences of real MT data https://github.com/gabrielStanovsky/mt_gender/blob/master/data/aggregates/en.txt
- Try to translate a few of the test MT sentences using Google Translate into a language with grammatical gender you know (e.g. French? Spanish? German?)
- Which gender is meant? What issues can you find?

Extra resource

Task 10.4
Translating Images

Extra Resource
Exercise 10.4 Translating Images

- Think of an object that is shiny/glittery, gold colour, has a loop one end, reflects the light, is made of a hard material.
- Write down what you think it is.

Exercise 10.4 Translating Images

- Watch the video

- After the video, take out your phones and visit the link below:

https://thing-translator.appspot.com/

Example:
Take a picture of a cup. Is the image description a cup?

What is the translation in a language you know? E.g. French

Your translation will be wrong depending on how much data the app has.

Exercise 10.4 It's not so easy to know what is meant

Exercise 10.4 It’s not so easy to know what is meant

The image description may not match the image.

If this wrong, the next stage, where you are translating it into another language, will be really wrong.

Thank you