Machine Translation

-- general concepts --

The concept of machine translation (MT)

- Translation = converting text or speech in one language (*a source language*) to a text in another language (*target language*)
- Usually performed by well-trained human translators

- Can be performed by computers, too (“machine translation”)

What is necessary for translation?

- Translation is a complex process
+ requires deep knowledge of both the source and target languages
+ And not only this:

    Also special skills which need hard work to be developed

- How to teach computers to do it?
+ Provide them with the rules of the two languages
+ Provide them with texts written in the two languages

Approaches

- Rule-based MT (RBMT)

+ Relies on complex set of rules for both source and target language
+ Requires very good knowledge of both languages and a lot of time
+ Does not require texts in any of the two languages

- Statistical MT (SMT, PBSMT), neural MT (NMT)

+ Requires parallel text written both in source and target language
  (larger texts => better translations)
+ Does not require any linguistic knowledge
+ Fast to develop (if an appropriate parallel text is available!)

A brief history

- 1950s:

  The very first experiments in machine translation: RBMT

- 1990s:

  emergence of SMT, suppressing RMBT systems

- 2015:

  emergence of NMT, suppressing SMT systems

- Now:

  NMT is the dominant approach

Rule-based MT (RBMT)

+ Does not require parallel texts
+ In principle, able to translate many different types of texts
+ Straightforward evaluation and improvement:
  every identified error can be corrected by an additional rule directed to this type of error

- Requires a complex set of language rules
- Always some exceptions => tough to include every possible rule
- Difficult to develop large dictionaries to cover all words and different meanings

Statistical MT (SMT)

+ Does not require hard work on different rules
+ Fast
  few days needed to develop a system for a new language pair,
  in comparison with few months for a RBMT system
+ Anyone can build an SMT system, not only language experts
+ More robust to errors or divergences in the source text

- a parallel text for the desired languages might not exist
  (or might be very small)
- More prone to morphological and syntactic errors
- More difficult to identify the cause of errors

Neural MT (NMT)

+ Better capable of learning than SMT because it relies on whole sentences instead of words and groups of words
+ Generates more natural (fluent) target language text
+ More possibilities to combine different knowledge sources

- Requires even more parallel texts
- Requires more time and computational resources
- Even more difficult to identify source/reason for errors

Who is using machine translation?

- MT is used every day, by different people

Researchers:

- developing complex computer programs for machine translation
- investigating data and methods for using different types of data
- developing methods for reliable evaluation of machine translation outputs
- some research groups are working with translators
  (our group in DCU works in all these directions)
- many work in big companies such as Google, Amazon, Microsoft, Facebook

Who is using machine translation?

Users translate millions of words every day

- “Translate this post” on social media platforms
- run different kinds of texts through publicly available online machine translation systems

A great contribution of machine translation -- it would not be possible without it!

- a professional translator is not really needed for news or social media article
- still, it might be very hard or impossible to find a person fluent in the source and in the target language and willing to translate such content for you
- even if yes, a person would not be able to translate it as fast as an MT system!

What about human translators?

- Machine translation is getting better and better, every year, maybe even every month
- But: it is important to be realistic!
- Recent improvements in the MT quality led to increasing hype in media and big companies

  > “Human parity achieved”, “MT systems translate better than human translators”, etc.

  ! MT is really only close to the human quality of translation in a very few limited scenarios

What about human translators?

- MT should not be used without revision by a human translator
- Users should not trust MT without any doubts

The “hype” trend also leads to increased uncertainty among the translators

However, translators are using MT, too!

- Some translators like to use MT for their work
+ They say it helps and makes the process faster
+ It can increase their productivity

What about human translators?

- Unfortunately, many translators do not like MT
+ A number of machine translated sentences is still very bad
+ They would like to see an indicator about the quality to know what to expect
+ It creates a task called “post-editing” which translators do not like

Especially if they are requested to edit the MT output only minimally

Post-editing = correcting errors in MT output

Are human translators still needed?

In short: yes!

In long:

- Even humans make mistakes, even human translations need revisions
  ! Do not fully trust MT without any revision
- MT is definitely getting better and better

  ! The claims and the hype about “human parity” are overstated
  ! MT quality heavily depends on the involved languages, translation direction, as well as type of the text (e.g. news articles, or medical texts, or forum posts)
- Will human translators be replaced one day?
  ! One day -- maybe; but this day still hasn’t arrived
  ! Nowadays: combine MT and human translation

What about linguistic knowledge?

- Nowadays, the dominant approach is corpus-based

  you can build a MT system for a language pair, of which you have absolutely no knowledge of (neither source nor target language).

- linguistic knowledge can help if you can incorporate it appropriately into a corpus-based approach
- It also helps you to understand the behaviour of a system, and its strengths and weaknesses

The importance of parallel texts

- Parallel texts are crurcial for statistical and neural machine translation systems
+ a lot of texts have already been translated by humans
+ such texts are usually called parallel texts (also called bilingual or multilingual parallel texts)
+ In a parallel text, the same content is available in two (or more) different languages

A small example of a parallel French-English text

J’aime le garçon.	I love the boy.
J’aime le chien.	I love the dog.
Ils aiment le chien.	They love the dog.
Ils parlent à la fille.	They talk to the girl.
Ils parlent au chien.	They talk to the dog.
Je parle à la mère.	I talk to the mother.

Parallel texts: challenges

- Parallel texts are crucial for SMT and NMT
- However, they do not solve all the difficulties of translation

  For morphologically rich language (a language with a large number of different forms of the same word), there might be some unseen forms of the word

Parallel texts: example of unseen word form

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.

There are different forms for words “cat” and “mouse”, also different structures of the sentence (a language with a flexible word order)

Parallel texts: example of unseen word form

Now, is the knowledge extracted from the parallel text enough to translate the following sentence into Croatian?

The mouse is with the cat.

+ Every English word has been seen in the parallel text…but still!
+ We have seen the phrase “with the mouse”, but not “with the cat”!
+ This phrase requires a form of the word “cat” which has not been seen in the parallel text.

Parallel texts: availability

Another challenge about parallel texts is their availability

- what language do we want to translate from, and into which language do we wish to translate?
- do we want a system for news articles, conversations, pharmaceutical texts or user reviews?
- or would we prefer that our system is more general; not specified for any particular type of text (called “domain”)?

! Some languages and domains are more equal than others

Getting a right parallel text is generally not trivial and it represents one of the challenges for MT

The importance of domain

Let us take a look at a parallel English-Spanish text:

English There are, on average, 300 days per year of sunshine in Argentina.
Spanish Hay, de media, 300 días al año soleados en Argentina.
English Average daily peak insolation varies from 1150 W/m2 in June to 280 W/m2 in December.
Spanish La insolación álgida diaria media varía de 1.150 W/m2 en junio a 280 W/m2 en diciembre.
English The driest months tend to be January and February, and the wettest, May, June, and November.
Spanish Los meses más secos tienden a ser enero y febrero, y los más lluviosos, mayo, junio y noviembre.

The importance of domain

Now, will the given parallel text be useful for learning how to translate some of the following sentences?

1. The valleys in Argentina have 300 days of sunshine.
2. Argentina: there is an excellent collection of graffiti from around Buenos Aires.
3. Almost all of Europe and the better part of Asia sit atop the Eurasian Plate.
4. Oh boy, Argentina is so cool, and Buenos Aires ROCKS!
5. Abilify is a medicine containing the active substance aripiprazole.

The importance of domain

1. About the climate in Argentina, same as the parallel text
   It will be possible to learn now to translate the sentence
2. Also about Argentina but different topic
   hard to learn useful stuff from the given parallel text
3. About geography, but something completely different
   hard to learn anything
4. About Argentina, but different topic and also informal language
   hard to learn anything useful
5. Medical text - something completely different
   Hard to learn anything

Format of parallel texts

Before using a parallel text for SMT or NMT, it has to be ensured that the text is:

- aligned on the sentence level
- tokenised

And for NMT, in addition:

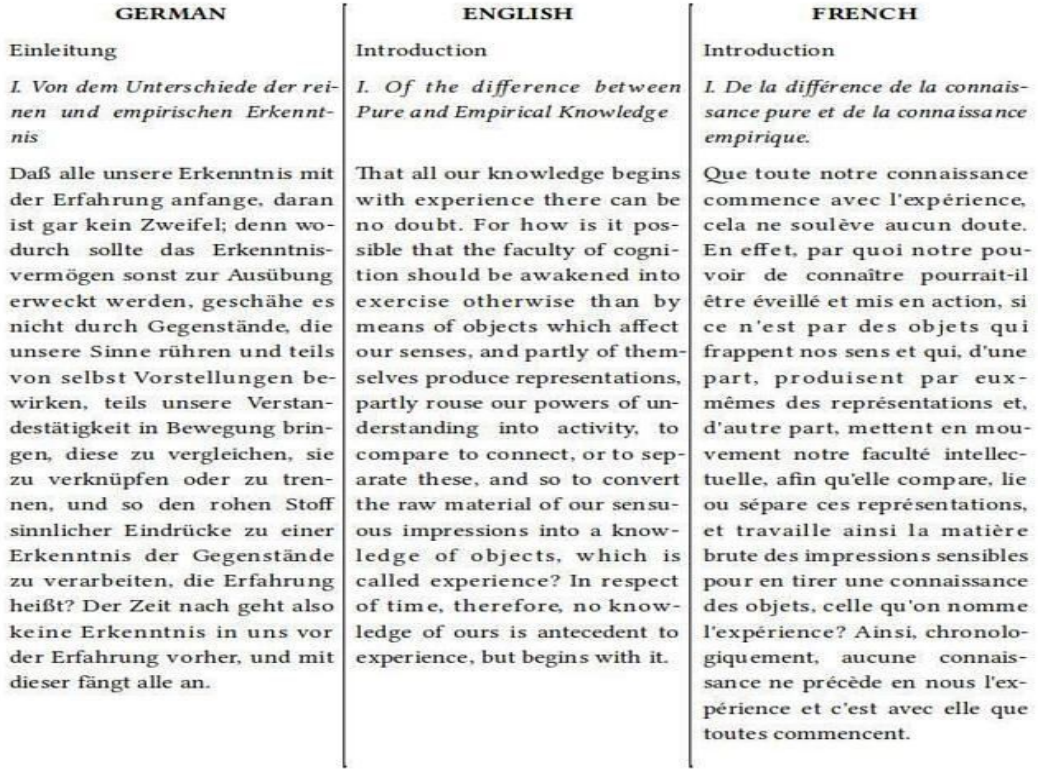
- The words should be split into so-called sub-word units

A text without a sentence alignment

- A trilingual parallel corpus (German, English and French)

- However, it cannot be used directly for MT

+ It has first to be aligned on the sentence level



GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
I. Von dem Unterschiede der reinen und empirischen Erkenntnis	*I. Of the difference between Pure and Empirical Knowledge*	*I. De la différence de la connaissance pure et de la connaissance empirique.*
Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.	Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.

The same text, now sentence-aligned

- Each sentence in one language has to be paired with the corresponding sentence in another language
- There are tools for automatic sentence alignment
- Note: “segment” is better term then “sentence”

\+ “Introduction” is not really a sentence

GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
Von dem Unterschiede der reinen und empirischen Erkenntnis	Of the differences between pure and empirical knowledge	De la différence de la connaissance pure et la connaissance empirique
Dass alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel;	That all our knowledge begins with the experience there can be no doubt.	Que tout notre connaissance commence avec l’expérience, cela ne soulève aucun doute.
Denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenständen zu verarbeiten, die Erfahrung heisst?	For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience?	En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n’est par des objets qui frappent nos sens et qui, d’une part, produisent par eux-mêmes des représentations et d’autre part, mettent en mouvement notre faculté intellectuelle, afin qu’elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu’on nomme l’expérience?
Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	In respect of time, therefore, no knowledge of ours in antecedent to experience, but begins with it.	Ainsi, chronologiquement, aucune connaissance ne précède en nous l’expérience et c’est avec elle que toutes commencement.

Tokenisation: separating punctuation marks

English	**French**
Sentence aligned, raw:	
I talk to a dog.	Je parle au chien.
The dog plays.	Le chien joue.
Tokenised:	
I talk to a dog .	Je parle au chien .
The dog plays .	Le chien joue .

Without tokenisation:

+ a model would learn that “chien.” corresponds to “dog.” and “chien” to “dog” as two completely different word pairs
+ “dog?” would be treated as a completely new, unseen word

Sub-word units (for NMT)

For NMT, an additional preprocessing method is necessary

- separating the words into so-called “sub-word” units

The do@@ g play@@ s . Le chi@@ en jo@@ ue .

- not linguistically motivated, but based on the most frequent character sequences
- Necessary in order to reduce the vocabulary as much as possible

Sub-word units (for NMT)

Why are necessary?

- Computations in NMT systems are very costly
- Complexity increases with the vocabulary size (number of distinct words)
  => The vocabulary size for NMT has to be restricted

- But: The more words in vocabulary, the better performance
  => Retaining most frequent words and removing the rest is not a good idea

- Sub-word units are a good balance
  => All words in the text can be used without exploding of the vocabulary size

Subword units: effects

+ Better handling of the internal structure of words (morphology)

  NMT systems able to generate word forms not seen in the training corpus

- NMT systems often generate words which do not exist at all

  Neither in the source language nor in the target language

  (“chi@@ en” and “do@@ g” might result in “chi@@ g”)

Publicly available parallel data: OPUS

OPUS parallel data collection
http://opus.nlpl.eu/

- A large number of language pairs
- Different types of texts (domains)
- The majority of the corpora is sentence aligned (column called “Moses”)
- A lot of corpora are noisy so a filtering step is necessary
- The data still has to be tokenised and converted into sub-word units

.

OPUS: availability of different language pairs

Millions of parallel segments	German	Spanish	French	Finnish	Croatian	Japanese	Tagalog	Swahili	Irish
English	234	290	367	61	55	6.2	1.8	1.4	6.2
German	/	64	69	34	20	5.2	1.4	1.1	1.3
Spanish	=	/	125	43	37	5.8	1.4	1.0	1.4
French	=	=	/	38	30	5.4	1.3	1.0	1.3
Finnish	=	=	=	/	24	4.7	1.3	1.0	1.2
Croatian	=	=	=	=	/	3.7	1.1	0.9	0.9
Japanese	=	=	=	=	=	/	1.4	1.1	0.4
Tagalog	=	=	=	=	=	=	/	1	0.26
Swahili	=	=	=	=	=	=	=	/	0.22

Availability of different language pairs

Already mentioned: some languages are more equal than others

- High-resourced languages: large amounts of data are available
- between 10 and 20 high-resourced languages.
- low -resourced languages: small amounts of available data
- Mid-resourced languages: something in between
- There are (still) no precisely defined limits for any of these categories
- + more- or less-investigated languages

The role of English

- the parallel data are generally very English-centric

+ parallel data involving English >> parallel data involving another language

  even for high-resourced language pairs

  (data for English-French or English-Spanish >> French-Spanish)

- English is involved in almost all MT experiments
  + either as the source or as the target language

- Language pairs without English are much less investigated

Availability of different domains

Some domains are more equal than others, too

- Even for the high-resourced languages:
  only a few dominating domains (for example news and subtitles)

- Other domains: small amounts of data or no data

Low-resourced scenarios

- the lack of parallel training corpora is common for most languages and domains
- An example of domain problem:
  user-generated content (social media posts, user product/movie reviews, etc.)
+ Huge amount of monolingual (mainly English) data, but almost no parallel texts

- the companies have access to much more data than publicly available corpora
+ however, the amount of this data for the majority of languages is still limited

=> low-resource settings represent an important challenge for modern NMT systems

Active research topic (many researches are working on it, trying diffferent methods)

Disadvantages of the sentence level

- Both SMT and NMT operate on the sentence level
- Contextual information from surrounding sentences is not used
+ This information can be very useful

Sentence 1: *This **chair** is way darker than it is in the picture.*

Sentence 2: ***It** is pretty.*

For many languages with grammatical gender (such as French and Spanish):

- It is important to know what “it” referes to in Sentence 2
- this information can be found in Sentence 1

Going beyond the sentence level

- Another active research topic
- Researchers are investigating different methods for including the information from surrounding sentences

  + such systems are called *context-aware NMT systems*

  + also *document-level NMT systems*

  although it is still not precisely defined what a “document” is

  basically, it refers to the general concept of going beyond a single sentence

Is this translation good?

- No matter where a machine translated text comes from (rules, statistics, or a neural network)
  + Its quality has to be checked

=> evaluation of machine translation

- far from a trivial task
- Another active research topic

Essential problem:

A language allows to express the same thing in many different ways.

Many ways to say the same thing in one language

1) He needs to get rid of a lot of junk.
2) He has to throw away tons of junk.
3) He has a lot of junk to throw away.
4) He needs to throw away a lot of stuff.
5) He needs to get rid of tons of junk.
6) He has tons of stuff to get rid of.

1) *The boy quick*ly ran across the finish line, seizing yet another victory.
2) The boy seized yet another victory when he quickly ran across the finish line.
3) The quick boy seized yet another victory when he ran across the finish line.

Many ways to translate the same thing in another language

There is no single one and only correct translation of a text

- MT systems are often evaluated manually:
+ evaluators assign scores to each sentence according to how good the translation of the sentence is
+ requires language knowledge and a lot of time

=> automatic evaluation metrics started to emerge

+ Based on similarity between MT translation and a human reference translation
+ Which brings us back to the main problem:
  + it is not **the** reference translation but **a** reference translation (just one of many)