NATURAL LANGUAGE PROCESSING

LESSON 11: MACHINE TRANSLATION

OUTLINE

- What is Machine Translation (MT)
- Difficulties
- Types of MT
  - Rule-Based Machine Translation (RBMT)
  - Statistical Machine Translation (SMT)
  - Neural Machine Translation (NMT)

WHAT IS MACHINE TRANSLATION?

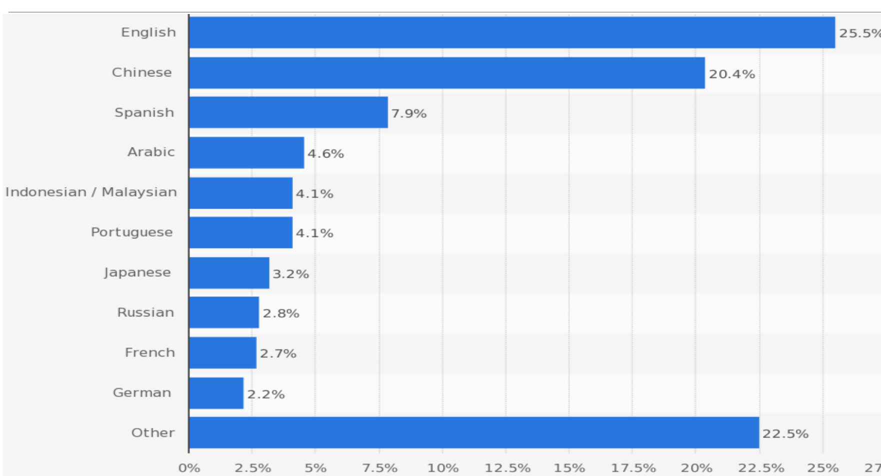
Machine translation (MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Turkish).

To process any translation, human or automated, the meaning of a text in the original (source) language must be fully restored in the target language.

Translation is not a mere word-for-word substitution. A translator must interpret and analyze all of the elements in the text and know all details about each word.

This requires extensive expertise in grammar, syntax (sentence structure), semantics (meanings), etc., in the source and target languages.

WHAT IS MACHINE TRANSLATION?



WHAT IS MACHINE TRANSLATION?

Human and machine translation each have their share of challenges.

For example, no two individual translators can produce identical translations of the same text in the same language pair, and it may take several rounds of revisions to meet customer satisfaction.

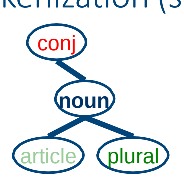
But the greater challenge lies in how machine translation can produce publishable quality translations.

MORPHOLOGICAL VARIATIONS

- Affixes (prefix/suffix)

write	→	written	Yaz-	→	Yazdı/mış
kill	→	killed	Öldür-	→	Öldürdü/müş
do	→	done	Yap-	→	Yaptı/mış

- Tokenization (segmentation + normalization)



And the cars	En
Ve Arabalar	Tr
Et les voitures	Fr

DIFFICULTIES

- The word order varies between languages
  - English : SVO
    Microsoft bought Yahoo
  - Turkish : SOV
    Microsoft Yahoo bought
  - Arabic, VSO :
    - bought Microsoft Yahoo
  - English / Turkish : Adj N
    A new car
    Yeni araba
  - Arabic, N Adj :
    - Car new

DIFFICULTIES

- Lexical Ambiguity
  - Bank -> Banka (financial)
  - Bank -> Sahil (coast)

- Syntactic Ambiguity – structural non-lexical
  - Thomas saw a man with a telescope
    - (3a) Thomas [SV saw [SN a man] [SP with a telescope]]
    - (3b) Thomas [SV saw [SN a man [SP with a telescope]]]
  - Who is using the telescope?

DIFFICULTIES

- Semantic Ambiguity: A form is semantically ambiguous if it can be mapped to at least two distinct meanings.
  - John and Mary are married." (To each other? or separately?)
    John ve Mary evli kişilerdir / John, Mary ile evlidir.

- Cultural aspects, e.g., calendars and dates
  - English : 09/10/2017 10th September 2017
  - Turkish : 09/10/2017 9 Ekim 2017

DIFFICULTIES

- Resolution of references

➢ O, onu seviyor.
  - **Bing**: She loves her.
  - **Google**: He loves it.

➢ Julie, Paul'un artık ona bakmamasını istiyor.
  - **Bing** : Julie wants Paul to stop looking at her.
  - **Google**: Julie wants Paul don't look at him anymore.

➢ Arabanın Kapısı (Senin/Onun)

➢ In general, translation is more difficult when the target language is morphologically richer than the source.

TYPES OF MT

A few different types of Machine Translation are available in the literature, the most widely use being

- Rule-Based Machine Translation (RBMT),
- Statistical Machine Translation (SMT),
- Google Neural Machine Translation (GNMT).

RULE-BASED MT (RBMT)

Historically, the first approach used to translate texts was based on linguistic rules.

The set of rules defines the possibilities of associating words according to their lexical categories and makes it possible to model the structure of a given sentence.

The software uses these complex rule sets and then transfers the grammatical structure of the source language into the target language.

This requires a lot of work on the part of linguists to define vocabulary and grammar.

RULE-BASED MT (RBMT)

The following example can illustrate the general frame of RBMT:

A girl eats an apple.

Source Language = English;
Demanded Target Language = Turkish

RULE-BASED MT (RBMT)

Minimally, to get a Turkish translation of this English sentence one needs:

1. A dictionary that will map each English word to an appropriate Turkish word.
2. Rules representing regular English sentence structure.
3. Rules representing regular Turkish sentence structure.

Finally, we need that rules can relate these two structures together.

RULE-BASED MT (RBMT)

Accordingly, we can state the following stages of translation:

1st: getting basic part-of-speech information of each source word:

a > determiner
girl > noun
eats > verb
an > determiner
apple > noun

2nd: getting syntactic information about the verb "to eat":

NP-eat-NP

here: eat – Present Simple, 3rd Person Singular, Active Voice

RULE-BASED MT (RBMT)

3rd: parsing the source sentence:

(NP a girl) = the subject of eat
(NP an apple) = the object of eat

4th: translate English words into Turkish

a (category > determiner) => bir (category > determiner)
girl (category > noun) => kız (category > noun)
eat (category > verb) => yemek (category > verb)
an (category > determiner) => bir (category > determiner)
apple (category > noun) => elma (category > noun)

RULE-BASED MT (RBMT)

5th: Mapping dictionary entries into appropriate inflected forms

A girl eats an apple (SVO) => Bir kız yiyor bir elma ~~(SVO)~~

Final generation:

A girl eats an apple (SVO) => Bir kız bir elma yiyor (SOV)

STATISTICAL MT (SMT)

- Statistical machine translation modeling is based on the mathematical theory of probabilistic distribution and estimation developed in 1990 with IBM's researchers .
- The initial hypothesis is that any sentence of one language is a possible translation of a sentence into another language.
- If we translate from a source language ***s*** to a target language ***t***, the goal is to find the target sentence ***t*** most appropriate to translate the source sentence ***s***.

STATISTICAL MT (SMT)

- For each pair of possible sentences (s, t), we assign a probability P(t|s) that can be interpreted as the probability that *t is the translation of s*.
- In statistical machine translation, probabilistic models are used to find the best possible translation *t** of a given source sentence *s*, among all possible *t* translations in the target language.
- This involves applying statistical learning methods to train the system with millions of words, including monolingual texts in the target language and aligned texts composed of translation examples between the two languages.

STATISTICAL MT (SMT)

The parameters of the statistical models are estimated from the analysis of a large amount of monolingual or bilingual learning data: called corpus. The corpora make possible to extract a set of useful information for statistical processing.

In the case of statistical machine translation, we need texts composed of translation examples between the two languages, or more precisely a set of sentences translated into source and target languages and aligned in pairs.

It's also called bitext which represents a parallel bilingual corpus (a text in a source language and its translation) where the translation links between sentences are explicit.

STATISTICAL MT (SMT)

A bitext is obtained from a bilingual corpus by aligning the corpus with the sentences.

The bitext is used for the training, development and evaluation of the statistical machine translation system.

The aim of the learning corpus is to train and construct the model using statistical learning methods.

The development corpus will be used to adjust and improve the models learned while the test corpus allows to check and test the quality of the learned model.

STATISTICAL MT (SMT)

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	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	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

STATISTICAL MT (SMT)

Statistical translation is defined as the search for the target sentence with the highest probability of being the translation of a source sentence.

By applying Bayes' theorem on the pair of sentences (s, t), where the sentence t in the target language is the translation of the sentence s into the source language, we obtain for each pair a probability P(t|s) that the machine produces the word t as a translation of the sentence s:

$$P(t|s) = \frac{P(S|t)P(t)}{P(s)}$$

Since we calculate the *arg max*$_t$ and s is independent of t, using only the product Pr (s | t) Pr (s), we arrive at the fundamental equation :

$$argmax\ P(t|s) = argmax(P(s|t)P(t))$$

STATISTICAL MT (SMT)

$$argmax\ P(t|s) = argmax(P(s|t)P(t))$$

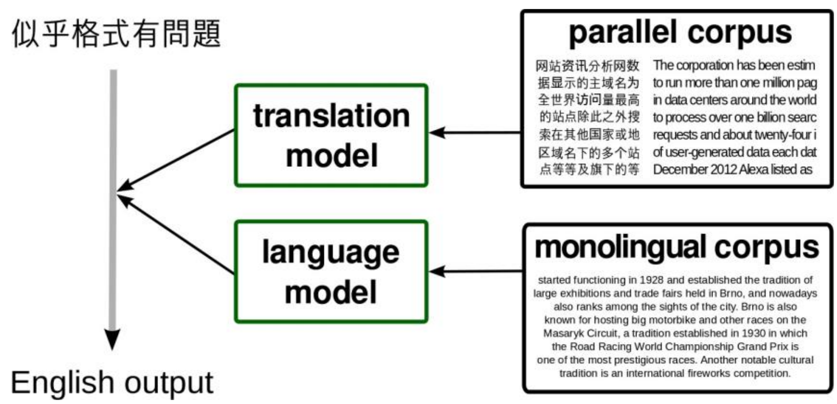
In this formula, P(t) is called the target language model and P(s|t) is called the translation model.

Both models are learned empirically from corpora.

P(t) is the probability of the sentence t in the target language and P(s|t) has the function of verifying that the source sentence is a translation of the target sentence t.

The sentence t* used for the translation of the sentence s will be the sentence which maximizes the product of the two probabilistic models of equation.

STATISTICAL MT (SMT)



SMT EXAMPLE

I love the boy	Ben Oğlanı seviyorum
I love the dog	Ben Köpeği seviyorum
They love the dog	Onlar köpekleri seviyorlar
they talk to the girl	Onlar kızla konuşuyorlar
they talk to the dog	Onlar köpekle konuşurlar
I talk to the mother	Ben Anneyle konuşurum

Aligned Data

SMT EXAMPLE

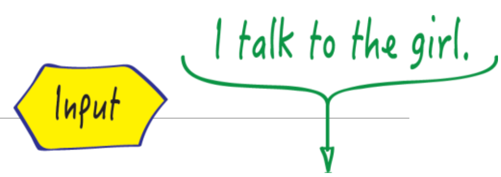
I love the boy
Ben Oğlanı seviyorum
I love the dog
Ben Köpeği seviyorum
they love the dog
Onlar Köpeği seviyorlar
they talk to the girl
Onlar kızla konuşuyorlar
they talk to the dog
Onlar köpekle konuşyorlar
I talk to the mother
Ben Anneyle konuşuyorum

Aligned Data

I	ben	3
They	onlar	3
Love	seviyorum	2
	seviyorlar	1
Talk	konuşyorlar	2
	konuşuyorum	1
The boy	Oğlanı	1
The dog	Köpeği	2
To the girl	kızla	1
To the dog	köpekle	1
To the mother	anneyle	1

Statistics

SMT EXAMPLE



I love the boy
Ben Oğlanı seviyorum
I love the dog
Ben Köpeği seviyorum
they love the dog
Onlar Köpeği seviyorlar
they talk to the girl
Onlar kızla konuşuyorlar
they talk to the dog
Onlar köpekle konuşyorlar
I talk to the mother
Ben Anneyle konuşuyorum

Aligned Data

I	ben	3
They	onlar	3
Love	seviyorum	2
	seviyorlar	1
Talk	konuşyorlar	2
	konuşuyorum	1
The boy	Oğlanı	1
The dog	Köpeği	2
To the girl	kızla	1
To the dog	köpekle	1
To the mother	anneyle	1

Statistics

SMT EXAMPLE

Input

I talk to the girl.

I love the boy
Ben Oğlanı seviyorum
I love the dog
Ben Köpeği seviyorum
they love the dog
Onlar Köpeği seviyorlar
they talk to the girl
Onlar kızla konuşuyorlar
they talk to the dog
Onlar köpekle konuşyorlar
I talk to the mother
Ben Anneyle konuşuyorum

Aligned Data

I	ben	3
They	onlar	3
Love	seviyorum	2
	seviyorlar	1
Talk	konuşyorlar	2
	konuşuyorum	1
The boy	Oğlanı	1
The dog	Köpeği	2
To the girl	kızla	1
To the dog	köpekle	1
To the mother	anneyle	1

Statistics

I	ben	3
Talk	konusuyorlar	2
	konuşuyorum	1
To the girl	kızla	1

How to choose?

SMT EXAMPLE

I	ben	3
Talk	konuşuyorlar	2
	konuşuyorum	1
To the girl	kızla	1

How to choose?

\+ Language Model

The Language Model:
- What is good in target language?
- Which words can follow which words
- and which can't? The "grammar"!
- Learnt from the data …
  - 1) Ben konuşuyorlar kızla ??
  - 2) Ben konuşuyorum kızla ??

  Ben ----- yorum
- Ben konuşuyorum kızla >> Ben konuşuyorlar kızla

Output => Ben kızla konuşuyorum (SOV)

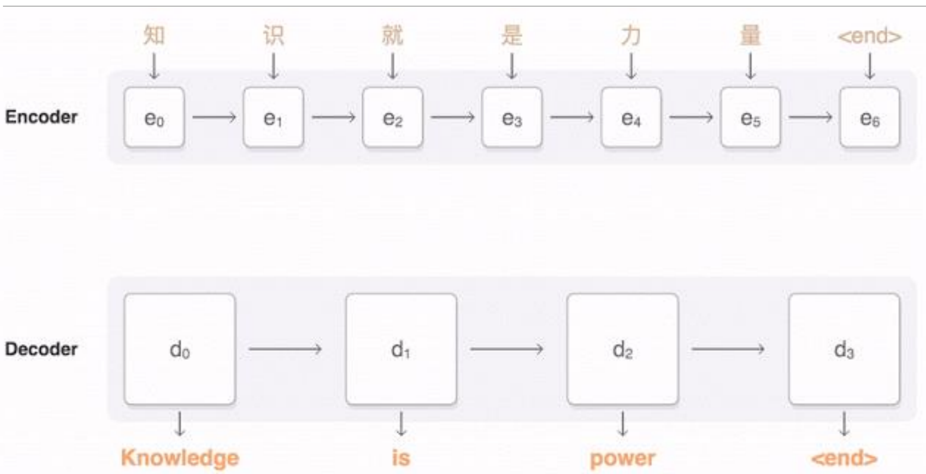
NEURAL MT SYSTEM (NMT)

In September 2016, Google announce the new method called Google Neural Machine Translation.

In contrast of traditionally used phrase-based machine translation (PBMT) system, which breaks an input sentence into individual words and phrases to be translated largely independently, the new Neural Machine Translation (NMT) system works on the entire input sentence as a single unit for translation.

To make this possible, the translator is first trained by showing it millions of examples of translations for every language pair.

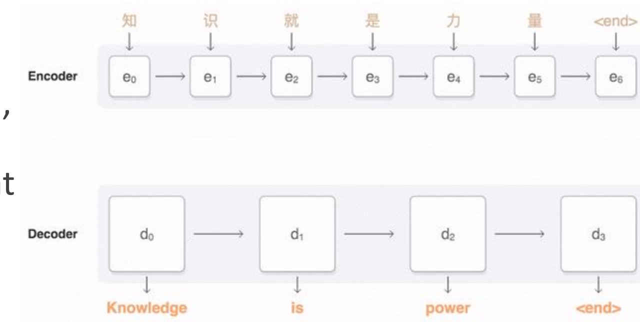
NEURAL MT SYSTEM (NMT)



NEURAL MT SYSTEM (NMT)

First, the network encodes the Chinese words as a list of vectors, where each vector represents the meaning of all words read ("Encoder").

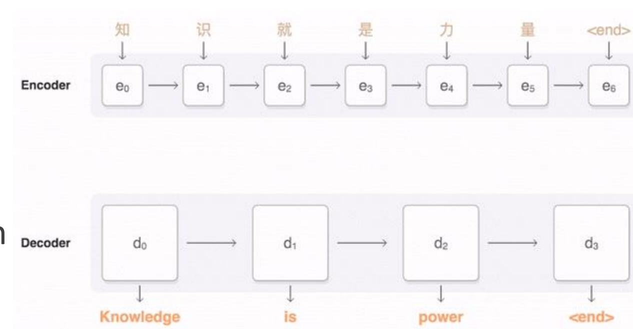
Once the entire sentence is read, the decoder begins, generating the English sentence one word at a time ("Decoder").



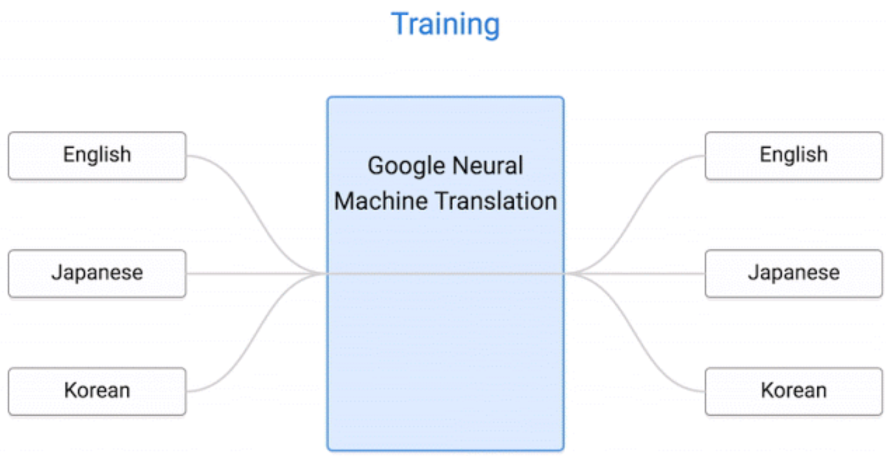
NEURAL MT SYSTEM (NMT)

To generate the translated word at each step, the decoder pays attention to a weighted distribution over the encoded Chinese vectors most relevant to generate the English word

("Attention"; the blue link transparency represents how much the decoder pays attention to an encoded word).

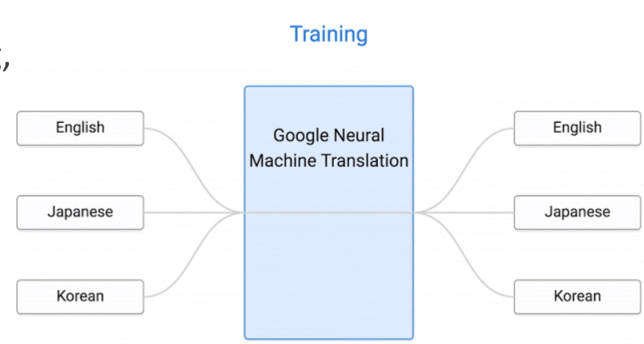


NEURAL MT SYSTEM (NMT)



NEURAL MT SYSTEM (NMT)

NMT is able to produce reasonable translations for language pairs that the system has never seen in training. As shown by the animation, during training, the framework is trained by showing it many examples of translations between English-Japanese and English-Korean pairs.



NEURAL MT SYSTEM (NMT)

Yet, the system is able to generate reasonable translations for the Japanese-Korean pair as well. This is possible because all the language pairs use the same neural network.