CMPE 409 Machine Translation

Murat ORHUN

March 9, 2022

İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

1. Introduction
2. Applications
3. Challenges
4. History
5. Available resources
6. MT paradigms
7. References

İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Introduction

Machine translation, sometimes referred to by the abbreviation MT, is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.

İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Introduction

On a basic level, MT performs mechanical substitution of words in one language for words in another, but that alone rarely produces a good translation because recognition of whole phrases and their closest counterparts in the target language is needed.

Introduction

Not all words in one language have equivalent words in another language, and many words have more than one meaning.

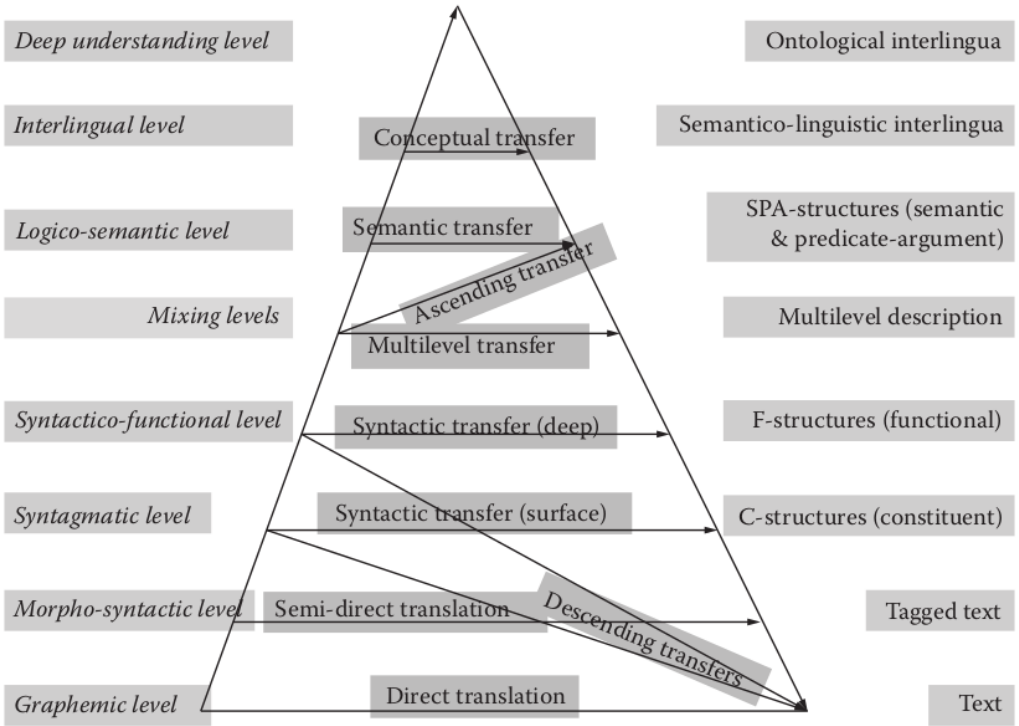
Introduction

Solving this problem with corpus statistical and neural techniques is a rapidly-growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies.

MT Approaches: Vauquois Triangle

- MT approaches have been grouped into a number of categories in the famous **Vauquois triangle**, also called the **Vauquois pyramid** (Vauquois, 1968, 1988).
- Prof. Bernard Vauquois was a translation theorist. Originally trained as a physicist, he got interested in automatic translation when the problem of translation between English and Russian assumed importance during the Cold War days.

MT Approaches: Vauquois Triangle



BUL BİLGİ UNIVERSITY
Computer Engineering Department

MT Approaches: Vauquois Triangle

- The left side of the triangle is the **ascending** side and the right side is the **descending** side.
- The left corner mentions the **source** language and the right corner the **target** language.
- When we ascend up the left-hand side, we perform analysis of various kinds on the source input sentence.

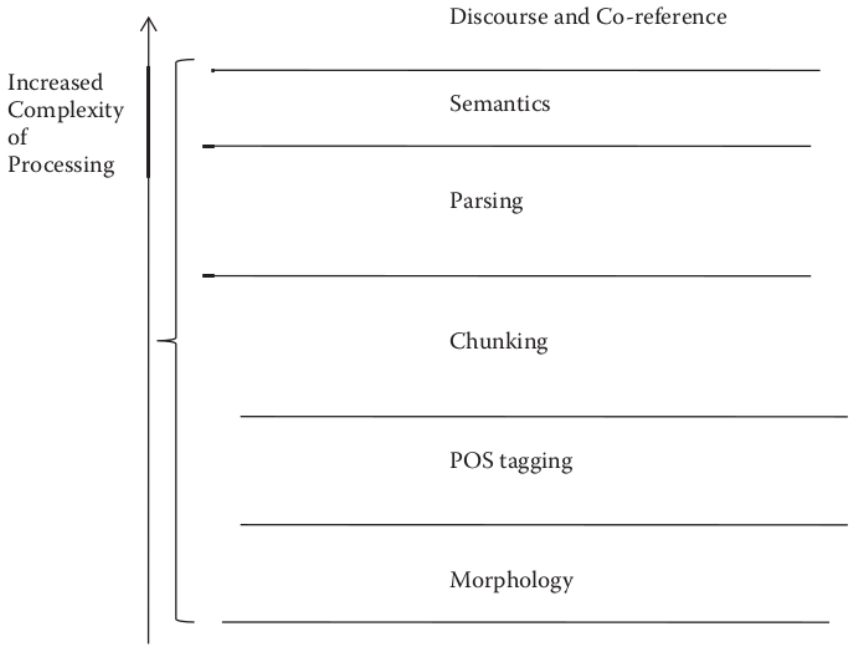
Input sentence could involve

1. Morphology analysis
2. Part of speech (POS) tagging
3. Noun and verb group identification (also called shallow parsing or chunking)
4. Parsing, followed by semantics extraction
5. Discourse resolution in the form of co-references
6. Pragmatics

MT Approaches: Vauquois Triangle

- Ascending the left-hand side of the Vauquois triangle until the apex amounts to traversing the NLP layers.
- After the analysis, the representation of the input sentence is taken through the stage of transfer.
- This means the representation is brought **on the side** of the target sentence

NLP Layers



İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Vauquois Triangle Elements

- **Ascending** transfers and **descending** transfers.
- It is important to remember that in the Vauquois triangle, the higher one goes toward the apex, the higher is the information richness of the representation.
- Example:

Example

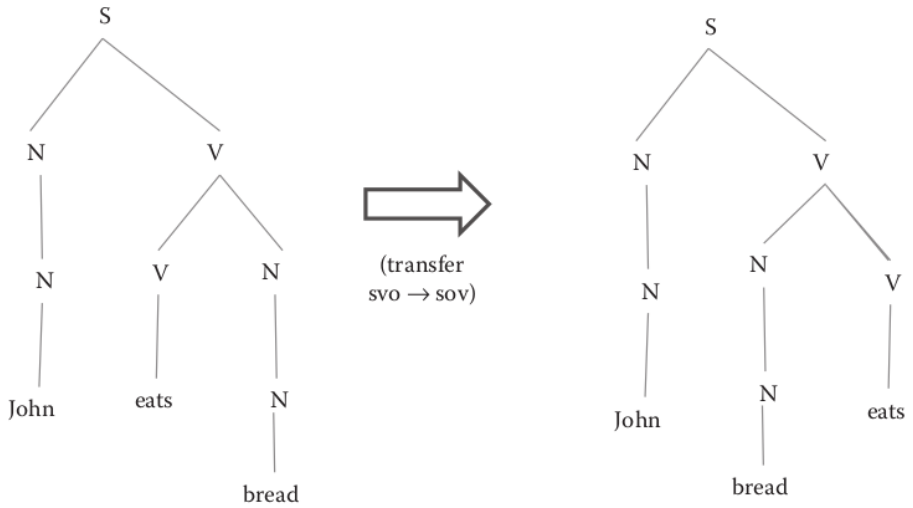
Graphemic level: The government levied new taxes.

Morphosyntactic level: The/DT government/NN levied/VBD new/JJ taxes/NNS./.

Syntagmatic level:

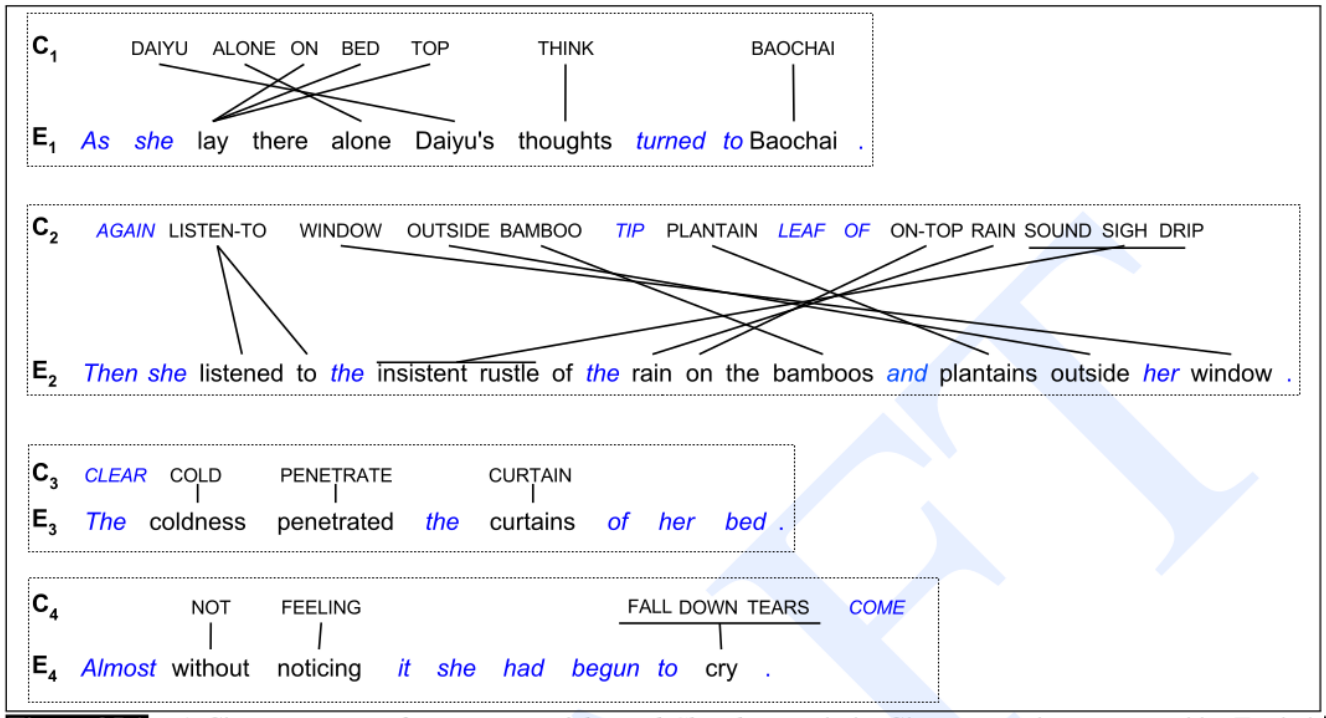
```
(S
    (NP (DT The) (NN Government))
    (VP (VBD levied)
      (NP (JJ new) (NNS taxes)))
    (..)))
```

Example Cont. svo to sov

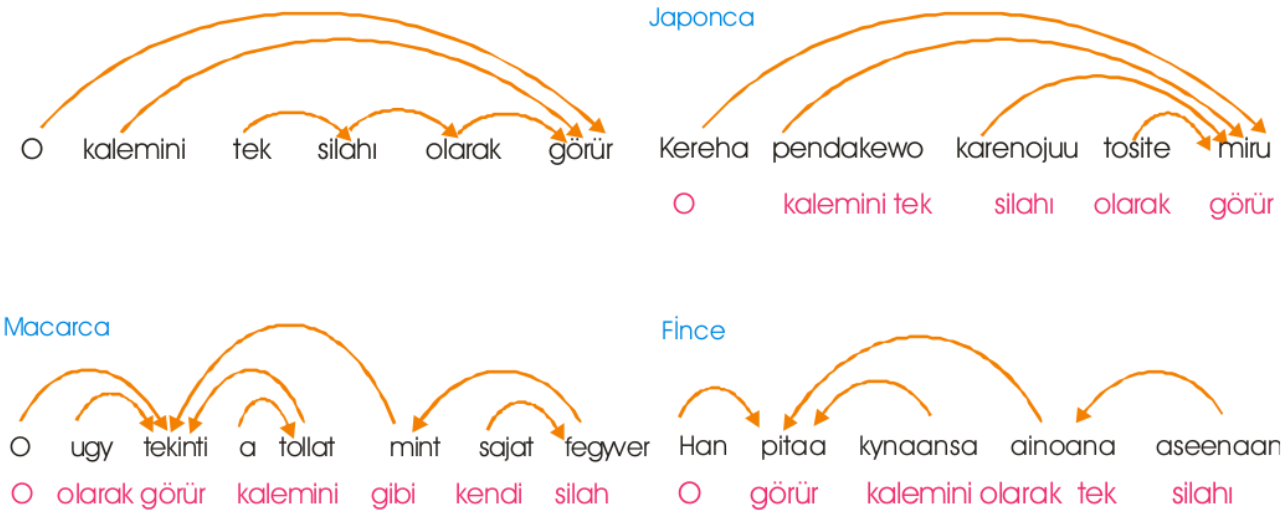


İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Chinese- English translation

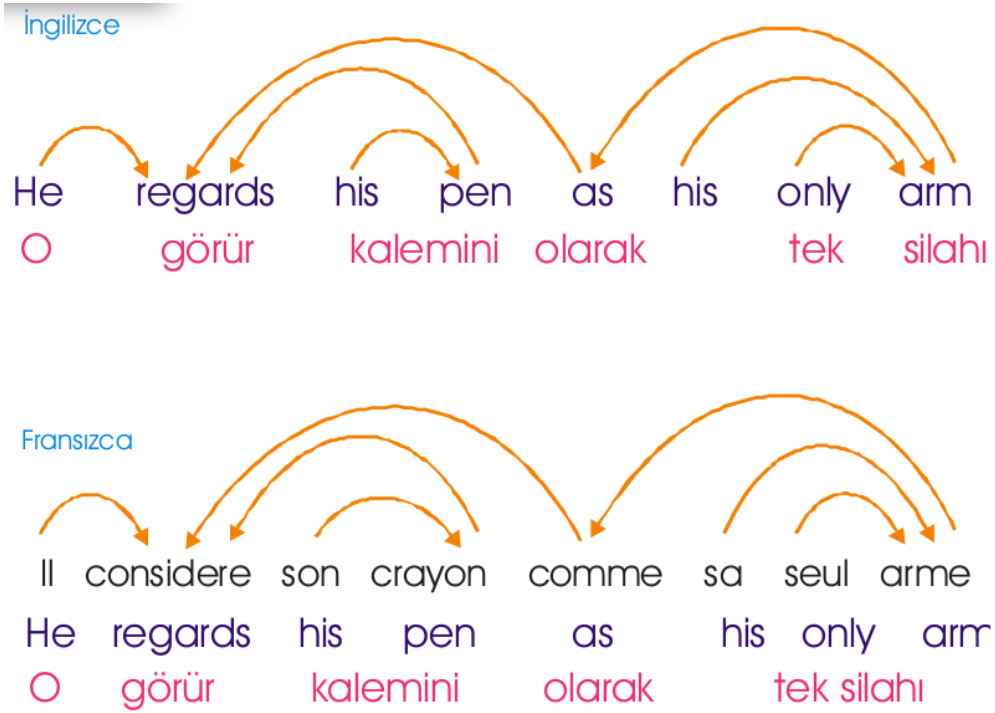


Turkish- Japanese- Hungarian- Finnish translation



İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Turkish- English-French translation

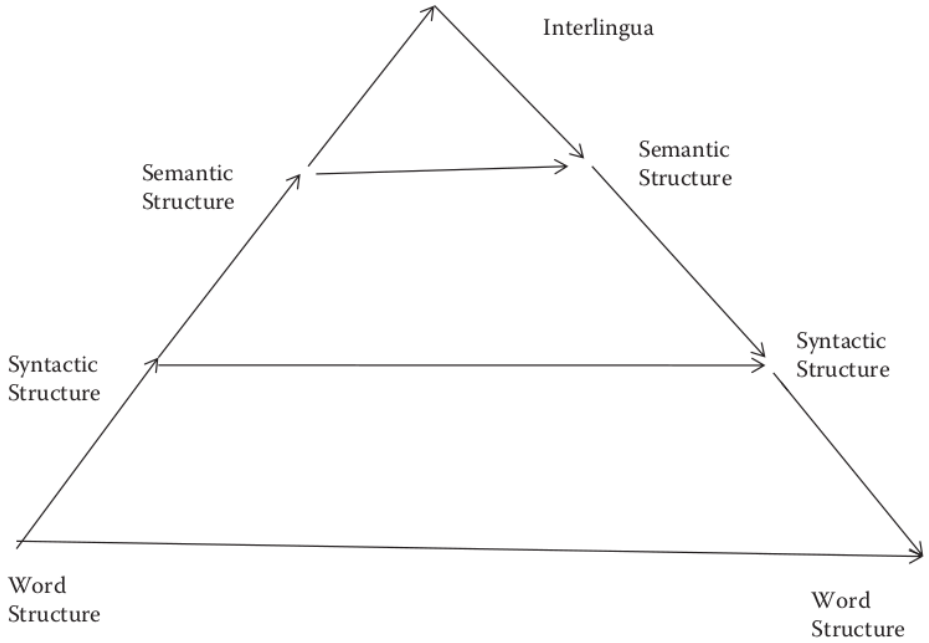


ISTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Vauquois Triangle Elements

- The consequence of the hypothesis is that the smaller the "distance" between the source language and the target language, the easier it is for the machine translation system to transfer between the two languages.
- Universality hypothesis: At the level of "deep meaning," all texts are the "same," whatever the language.
- The consequence of the hypothesis is that the smaller the "distance" between the source language and the target language, the easier it is for the machine translation system to transfer between the two languages.
- Similar or close languages

Simplified Vauquois Triangle



İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Applications

- Expecting Fully Automatic High Quality Machine Translation (FAHQMT)
- Limited domains
- General purpose

Météo

The METEO System is a machine translation system specifically designed for the translation of the weather forecasts issued daily by Environment Canada.The system was used from 1981 to 30 September 2001 by Environment Canada to translate forecasts issued in French in the province of Quebec into English and those issued in English in other Canadian provinces into French.

Météo

FPCN18 CWUL 312130	*FPCN78 CWUL 312130*
SUMMARY FORECAST FOR WESTERN QUEBEC ISSUED BY ENVIRONMENT CANADA	*RESUME DES PREVISIONS POUR L'OUEST DU QUEBEC EMISES PAR ENVIRONNEMENT CANADA*
MONTREAL AT 4.30 PM EST MONDAY 31 DECEMBER 2001 FOR TUESDAY 01 JANUARY 2002. VARIABLE CLOUDINESS WITH FLURRIES. HIGH NEAR MINUS 7.	*MONTREAL 16H30 HNE LE LUNDI 31 DECEMBRE 2001 POUR MARDI LE 01 JANVIER 2002. CIEL VARIABLE AVEC AVERSES DE NEIGE. MAX PRES DE MOINS 7.*
END/LT	*FIN/TR*

Figure 1. An example of an English weather report and its French translation.

SITY
Computer Engineering Department

Systran

- SYSTRAN, founded by Dr. Peter Toma in 1968,is one of the oldest machine translation companies
- SYSTRAN has done extensive work for the United States Department of Defense and the European Commission.

Special domain

- Medical report
- Title translation
- Close languages

Skype

Skype Translator

Whether you need to translate English to Spanish, English to French, or communicate in voice or text in dozens of languages, Skype can help you do it all in real time – and break down language barriers with your friends, family, clients and colleagues.

Our **voice translator** can currently translate conversations from 60 languages into 11 languages, including **English, Spanish, French, German, Chinese (Mandarin), Italian, Portuguese (Brazilian), Arabic, and Russian**.

Talking dictionary

Talking Electronic Dictionaries bring you prompt translations of words and expressions as well as advanced synthesis of speech. Translations are not only given in the form of text but are also pronounced. Foreign language students will enjoy these great audio aids. The extras include games, business and travel features, built-in language studying aids, professional add-on dictionaries, and other features.

http://www.ectaco.translation.net/

İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Applications

- Google Translate.
- Microsoft Translator.
- Yandex.
- IBM Watson Language Translator.
- Amazon Translate.
- Bing Translator.
- Cloud Translation API

Lexical Ambiguity

Example 1:
book the flight => reservar
read the **book** => libro
Example 2:
the box was in the **pen**
the **pen** was on the table
Example 3:
kill a man => matar
kill a process => acabar

Differing Word Orders

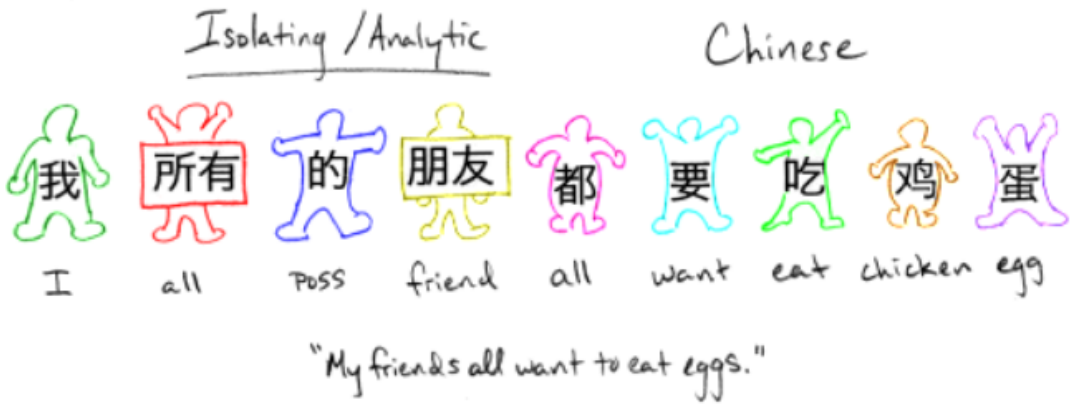
- English word order is *subject – verb – object*
- Japanese word order is *subject – object – verb*

English: IBM bought Lotus
Japanese: *IBM Lotus bought*

English: Sources said that IBM bought Lotus yesterday
Japanese: *Sources yesterday IBM Lotus bought that said*

RSITY
artment

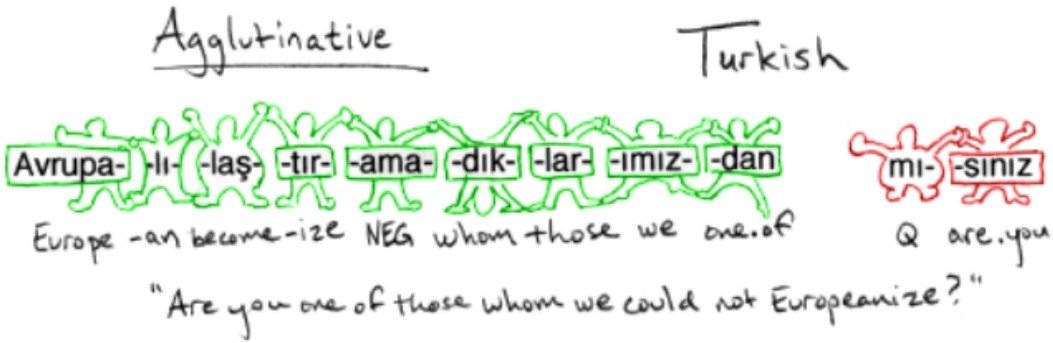
Topology



https://allthingslinguistic.com/post/50939757945/morphological-typology-illustrations-from

İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Topology



https://allthingslinguistic.com/post/50939757945/morphological-typology-illustrations-from

İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Topology

elma	erik
elma**lar**	erik**ler**
elma**cık**	erik**cik**
elma**cı**	erik**çi**
elma**cılık**	erik**çilik**
elma**lık**	erik**lik**
elma**m**	eriğ**im**
elma**n**	eriğ**in**
elma**sı**	eriğ**i**
elma**mız**	eriğ**imiz**
elma**nız**	eriğ**iniz**
elma**ları**	erik**leri**
elma**mın**	eriğ**imin**
elma**yı**	eriğ**i**
elma**ya**	eriğ**e**

İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

Doğal Dil İşleme Eşref Adalı

Other

- Syntactic Structure
- Syntactic Ambiguity
- Pronoun Resolution
- Tense
- Cultural problems

History

"I have a text in front of me which is written in Russian
pretend that it is really written in English and that it ha
strange symbols. All I need to do is strip off the code in
information contained in the text."
--Warren Weaver

Georgetown-IBM experiment (1954)

- „[...] human translations were subject to political bias and interference“
- Translation of 60 sentences from Russian into English
- Topic: organic chemistry
- System: six grammar rules and 250 words in the vocabulary

Georgetown-IBM experiment (1954)

- „[...] human translations were subject to political bias and interference“
- Translation of 60 sentences from Russian into English
- Topic: organic chemistry
- System: six grammar rules and 250 words in the vocabulary

Georgetown-IBM experiment (1954)

- Conclusions
  - The problem was solved
  - But semantic disambigution are impossible to be solved automatically

Russian (Romanized)	English translation
Mi pyeryedayem mislyi posryedstvom ryechyi.	We transmit thoughts by means of speech.
Vyelyichyina ugla opryedyelyayetsya otnoshyenyiyem dlyini dugi k radyiusu.	Magnitude of angle is determined by the relation of length of arc to radius.
Myezhdunarodnoye ponyimanyiye yavlyayetsya vazhnim faktorom v ryeshyenyiyi polyityichyeskix voprosov.	International understanding constitutes an important factor in decision of political questions.

https://en.wikipedia.org/wiki/Georgetown

ALPAC Report

- Automatic Language Processing Advisory Committee
- Study of reality of MT
- Conclusions:
  - post-editing not cheaper than full translation
  - Little Russian scientific literature worth to be translated
  - No shortage of human translators
  - No advantage in using machine translation
  - Better fund linguistic research for human translation
- Funding for MT stopped in the US as a consequence

http://www.hutchinsweb.me.uk/MTNI-14-1996.pdf

MT Systems

- Météo
- Systran
- Logos
- METAL
- Trados

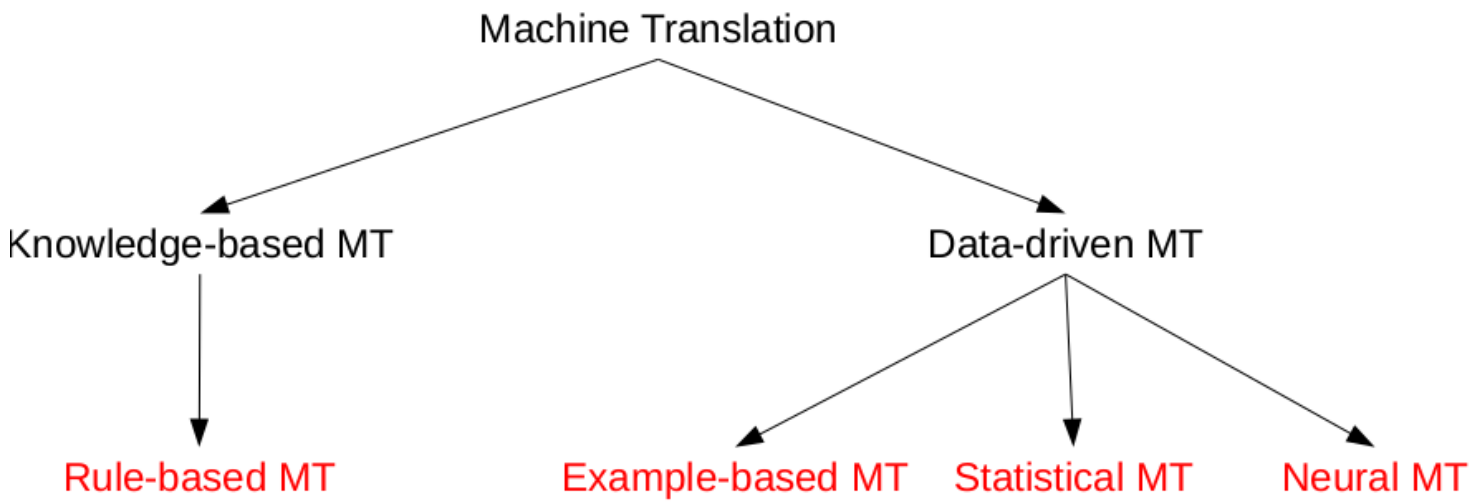
MT Systems

- Rule-Based Machine Translation (RBMT)
- Statistical Machine Translation (SMT)
- Hybrid Machine Translation (HMT)
- Neural Machine Translation (NMT)

MT Systems

- NLTK
- Corpora
- Evaluation

MT paradigms



Rule Based - MT

Rule-based machine translation (RBMT; "Classical Approach" of MT) is machine translation systems based on linguistic information about source and target languages basically retrieved from (unilingual, bilingual or multilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively

Rule Based - MT

- Direct Systems (Dictionary Based Machine Translation) map input to output with basic rules.
- Transfer RBMT Systems (Transfer Based Machine Translation) employ morphological and syntactical analysis.
- Interlingual RBMT Systems (Interlingua) use an abstract meaning

Example Based - MT

Example-based machine translation (EBMT) is a method of machine translation often characterized by its use of a bilingual corpus with parallel texts as its main knowledge base at run-time. It is essentially a translation by analogy and can be viewed as an implementation of a case-based reasoning approach to machine learning.

Example Based - MT

Example of bilingual corpus

English	Japanese
How much is that **red umbrella**?	Ano **akai kasa** wa ikura desu ka.
How much is that **small camera**?	Ano **chiisai kamera** wa ikura desu ka.

Statistical machine translation

Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation.

SMT-Benefits

- More efficient use of human and data resources
  - There are many parallel corpora in machine-readable format and even more monolingual data.
  - Generally, SMT systems are not tailored to any specific pair of languages.
  - Rule-based translation systems require the manual development of linguistic rules, which can be costly, and which often do not generalize to other languages.
- More fluent translations owing to use of a language model

SMT-Shortcomings

- Corpus creation can be costly.
- Specific errors are hard to predict and fix.
- Results may have superficial fluency that masks translation problems.[10]
- Statistical machine translation usually works less well for language pairs with significantly different word order.
- The benefits obtained for translation between Western European languages are not representative of results for other language pairs, owing to smaller training corpora and greater grammatical differences.

Neural machine translation

Neural machine translation (NMT) is an approach to machine translation that uses an artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model.

Recourse

- Jurafsky, D. and J. H. Martin. Speech and language processing*:* An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition, Upper Saddle River, NJ: Prentice-Hall, 2008.
- Machine Translation: an Introdu tory Guide , NCC Bla kwell, London, 1994, ISBN: 1855542-17x
- Bhattacharyya, P. Machine Translation, Indian Inst fo Tech Bombay. Mumbai India,2014

İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department