

COLX-531: Neural Machine Translation

Muhammad Abdul-Mageed

muhmmad.mageed@ubc.ca

Deep Learning & NLP Lab

The University of British Columbia

Table of Contents

1 Motivation

2 History of SMT

3 Resources

Machine Translation

The screenshot shows the Google Translate interface with two side-by-side panels. Both panels have a header with language selection: English-Detected, French, English, Spanish, and a dropdown arrow. Below the header is a horizontal double-headed arrow icon. The left panel contains the English text:

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

The right panel shows the French translation:

C'était le meilleur des temps, c'était le pire des temps, c'était l'âge de la sagesse, c'était l'âge de la folie, c'était l'époque de la croyance, c'était l'époque de l'incrédulité, c'était la saison de la Lumière, c'était la saison des Ténèbres, c'était le printemps de l'espoir, c'était l'hiver du désespoir, nous avions tout devant nous, nous n'avions rien devant nous, nous allions tous directement au Ciel, nous allions tous directement dans l'autre sens - en Bref, la période était si proche de la période actuelle, que certaines de ses autorités les plus bruyantes ont insisté pour qu'elle soit reçue, pour le bien ou pour le mal, dans le degré de comparaison superlatif seulement.

Both panels include a speaker icon, a word count (611/5000), and a pencil icon. The right panel also includes a star icon and a share icon.

Figure: Google Translate | Charles Dickens, *A Tale of Two Cities*.

Why MT?

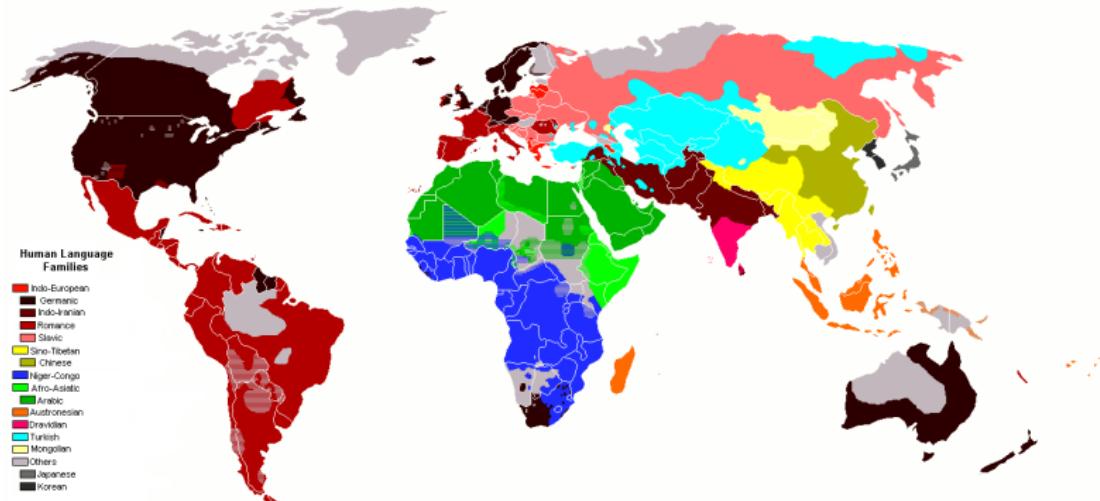


Figure: Human language families. Source: Wikipedia.

Why MT? *Contd.*

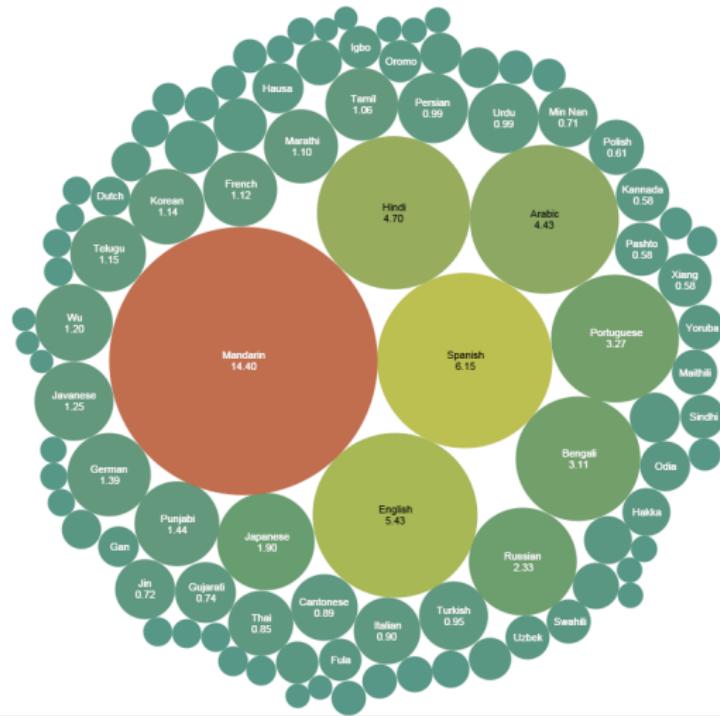


Figure: Languages by native speakers. Source: Wikipedia, based on Ethnologue.

Why MT? Contd.

Top Ten Languages Used in the Web - April 30, 2019 (Number of Internet Users by Language)					
TOP TEN LANGUAGES IN THE INTERNET	World Population for this Language (2019 Estimate)	Internet Users by Language	Internet Penetration (% Population)	Internet Users Growth (2000 - 2019)	Internet Users % of World (Participation)
English	1,485,300,217	1,105,919,154	74.5 %	685.7 %	25.2 %
Chinese	1,457,821,239	863,230,794	59.2 %	2,572.3 %	19.3 %
Spanish	520,777,464	344,448,932	66.1 %	1,425.8 %	7.9 %
Arabic	444,016,517	226,595,470	51.0 %	8,917.3 %	5.2 %
Portuguese	289,923,583	171,583,004	59.2 %	2,164.8 %	3.9 %
Indonesian / Malaysian	302,430,273	169,685,798	56.1 %	2,861.4 %	3.9 %
French	422,308,112	144,695,288	34.3 %	1,106.0 %	3.3 %
Japanese	126,854,745	118,626,672	93.5 %	152.0 %	2.7 %
Russian	143,895,551	109,552,842	76.1 %	3,434.0 %	2.5 %
German	97,025,201	92,304,792	95.1 %	235.4 %	2.1 %
TOP 10 LANGUAGES	5,193,327,701	3,346,642,747	64.4 %	1,123.0 %	76.3 %
Rest of the Languages	2,522,895,508	1,039,842,794	41.2 %	1,090.4 %	23.7 %
WORLD TOTAL	7,716,223,209	4,386,485,541	56.8 %	1,115.1 %	100.0 %

NOTES: (1) Top Ten Languages Internet Stats were updated in April 30, 2019. (2) Internet Penetration is the ratio between the sum of Internet users speaking a language and the total population estimate that speaks that specific language. (3) The most recent Internet usage information comes from data published by [Nielsen Online](#), [International Telecommunications Union](#), [GfK](#), and other reliable sources. (4) Population estimates are based mainly on figures from the [United Nations Population Division](#) and local official sources. (5) For definitions, methodology and navigation help, please see the [Site Surfing Guide](#). (6) These statistics may be cited, stating the source and establishing an active link back to [Internet World Stats](#). Copyright © 2019, Miniwatts Marketing Group. All rights reserved worldwide.

Figure: Source: Internet World Stats.

Why MT? Contd.

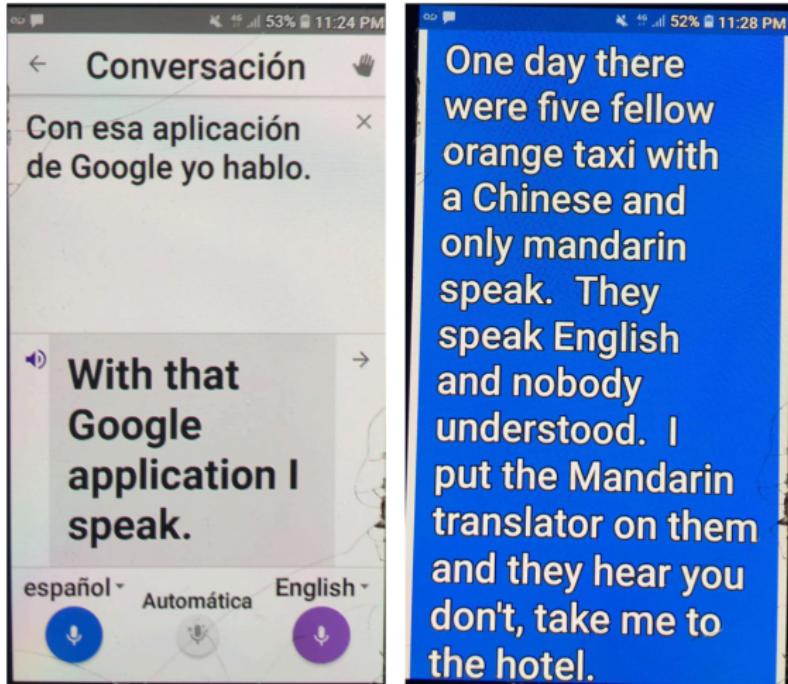


Figure: Taxi driver story in Costa Rica, Feb. 2020.

Translation is Hard! (But Fun!)



ISTOCK

Figure: "9 little translation mistakes caused big problems" on mentalfloss.com.
Quotes from "Found in Translation" book by Nataly Kelly & Jost Zetzsche.

Translation is Hard! (But Fun!) *Contd.*

2. Your lusts for the future

When President Carter traveled to Poland in 1977, the State Department hired a Russian interpreter who knew Polish, but was not used to interpreting professionally in that language. Through the interpreter, Carter ended up saying things in Polish like "when I abandoned the United States" (for "when I left the United States") and "your lusts for the future" (for "your desires for the future"), mistakes that the media in both countries very much enjoyed.

Figure: Source: mentalfloss.com.

Translation is Hard! (But Fun!) Contd.

The image shows a screenshot of the Google Translate interface. At the top, there's a grey header bar with the word "Translate" in red on the left, a double-headed arrow icon, and language selection buttons for English, Arabic, French, and a dropdown menu. To the right of these is a blue "Translate" button. Below the header, there are four pairs of text boxes, each consisting of an input box on the left and an output box on the right. The first pair contains the Arabic sentence "يا له من يوم جميل في فانکوفر!" followed by its English translation "What a lovely day in Vancouver!". The second pair contains " صباحنا قشطة!" followed by "Morning cream!". The third pair contains "الولد ده لسا ضارب كشري." followed by "The boy is a lion.". The fourth pair contains "الولد ده لسا ضارب كشري.." followed by "The boy is a loser.". The entire interface is set against a light grey background.

يا له من يوم جميل في فانکوفر!	What a lovely day in Vancouver!
صباحنا قشطة!	Morning cream!
الولد ده لسا ضارب كشري.	The boy is a lion.
الولد ده لسا ضارب كشري..	The boy is a loser.

Figure: Google Arabic/Egyptian Arabic-English Translation samples, Oct. 24, 2018

Facebook's AI Just Set A New Record In Translation And Why It Matters



Hieroglyphics are seen on the sarcophagus once belonging to the Judge and prime minister Gemeneferherbek of Sals at the Egyptian Museum in Turin, Italy, Tuesday, February 21, 2006.
Photographer: Adam Berry/Bloomberg News

Figure: Source: Forbes. (More than 10 BLEU points improvement!)

History of MT: Origins



Muhammad Abdul-Mageed

5 mins · 🔍

...

Machine Translation has a very interesting history.

"The origins of machine translation can be traced back to the work of Al-Kindi [left], a 9th-century Arabic cryptographer who developed techniques for systemic language translation, including cryptanalysis, frequency analysis, and probability and statistics, which are used in modern machine translation. The idea of machine translation later appeared in the 17th century. In 1629, René Descartes [right] proposed a universal language, with equivalent ideas in different tongues sharing one symbol." Read more: https://en.wikipedia.org/wiki/History_of_machine_translation"



History of MT: Warren Weaver

Warren Weaver's "Translation" memorandum (1949)

- Goal: Going beyond word-for-word translation. (4 proposals)
 - ① Use **context** to solve word sense **ambiguity**
 - ② Translation could be addressed as a *problem of formal logic*, **deducing "conclusions" in the target language from "premises" in the source language** (**seq2seq?**)
 - ③ **Cryptographic methods applicable to translation**
 - ④ There may be linguistic universals underlying all human languages (**cross-lingual methods?**)

History of MT: The Georgetown Experiment

Georgetown Experiment (Jan. 7, 1954)

- Leon Dostert, a French-born American scholar of languages with impact on MT
- Paul Garvin had a knowledge of Russian and carried out the linguistics part of the demonstration
- System had only **six grammar rules** and **250 lexical items in its vocabulary**
- A **toy system**, but was covered by journalists and received well by the public
- Encouraged **governmental funding** in **computational linguistics**
- Read more on [Wikipedia](#)...

History of MT: The ALPAC Report

ALPAC Report (1966)

- **The US government:** Concerned about lack of progress despite significant funding
- **ALPAC**, the Automatic Language Processing Advisory Committee, 7 scientists convened by US government
- **ALPAC concluded:**
 - MT *more expensive, less accurate, and slower than human translation*
 - MT *not likely to reach human quality in the near future*
- Research “almost completely abandoned” for over a decade in the **US**, and to a lesser extent in the **Soviet Union** and **United Kingdom**
- Research continued in **Canada, France, and Germany**
- Read more on [Wikipedia](#)...

History of MT: The 1980s | Example-Based MT



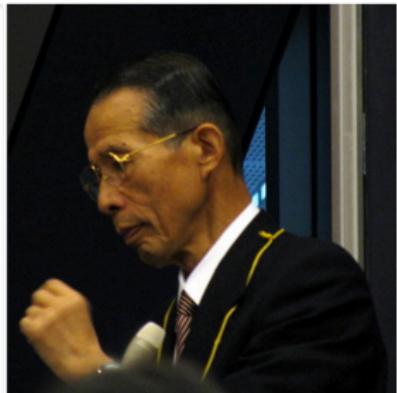
Muhammad Abdul-Mageed

11 mins · ▾

•••

During the 1980s there was a lot of activity in MT in Japan. Methodologically, translation was done through some variety of intermediary linguistic representation involving morphological, syntactic, and semantic analysis

Kyoto Nago, who received his PhD in Engineering from Kyoto University (1961) and later became a President (1997–2003), and his group popularized example-based MT. The approach ‘neglects’ syntactic and semantic rules and relies instead on the manipulation of large text corpora. Read more:
https://en.wikipedia.org/wiki/History_of_machine_translation



The Early 1990s | Industry

The 1980s: A lot of activity in Japan

- Research began into speech translation with the development of the **German Verbmobil project**
- Development of the **Forward Area Language Converter (FALCon) system** by the Army Research Laboratory, fielded 1997 to aid soldiers in Bosnia
- Transition away from **large mainframe computers** toward **personal computers and workstations**
- **AltaVista's Babel Fish** (using Systran technology) and **Google Language Tools** (also initially using Systran technology exclusively)
- Read more on [Wikipedia](#)...

The 2000s | Industry

The 2000s: Industry

- Focus statistical MT and example-based MT
- **Speech translation** expanding to more domains (e.g., parliamentary and news)
- **Multi-lingual MT** (e.g., web content, including audio and video)
(See [Wikipedia](#))

NMT

- Sparked by work on **word embeddings** (e.g., Bengio et al., 2003; Collobert & Weston, 2008; Mikolov et al., 2013)
- **Sequence-to-sequence models** (Sutskever, 2014)
- **Attention** (Graves, 2013; Bahdanau et al., 2014; Vaswani et al., 2017)
- Use of **monolingual data**, including **backtranslation**
- Use of **sub-word units** (byte-pair encoding) and **large vocabularies**
- **One-shot** and **low-resource MT**

Word, Phrase, & Eval Tools

- **GIZA++**: Implementation of IBM word-based models; used for word alignment ([link](#))
- **SRILM**: A tool for language modeling ([link](#))
- **Moses**: a decoding tool offering phrase-based and tree-based decoding ([link](#))
- **Pharaoh**: Predecessor of Moses, a beam search decoder for phrase-based SMT ([link](#))
- **BLEU**; **METEOR**: Evaluation metrics implemented in various tools (e.g., Pytorch, NLTK)

Various Datasets

- **LDS data:** Arabic-English, Chinese-English, French-English, etc.
([link](#))
- **European:** European Parliament Proceedings Parallel Corpus ([link](#))
- **OPUS:** An open source parallel corpus; a growing collection of translated texts from the web ([link](#))

Eval Campaigns

Eval Campaigns

- **NIST**: National Institute of Standards and Technology; focus on **Arabic-English** and **Chinese-English**; Since 2002, has coordinated evaluations of text-to-text MT technology through OpenMT series ([link](#))
- **IWSLT**: The International Workshop on Spoken Language Translation; focus on **speech translation** ([2019](#); [2020](#); [2021](#))
- **WMT**: Workshop on SMT; focus on **European languages** ([2020](#); [2021](#))
- Note: **Focus shifts**, depending on funding, interest, and progress in the field