Introduction – Lecture 1

Bill Byrne

Lent 2022

Neural Machine Translation and Dialogue Systems – MLMI8

MPhil in Machine Learning and Machine Intelligence

UNIVERSITY OF CAMBRIDGE

Course Overview: Lectures and Assessment

Lectures (tbc):

- L1 – Course Overview and Introduction to Machine Translation
- L2 – Introduction to Dialogue Systems
- L3 – Transformers
- L4 – Weighted Finite State Transducers (a bit of still-useful history)
- L5 – Models and Metrics: Translation and Dialogue
- L6 – Search and Modeling Issues
- L7 – Ethical Concerns in Translation and Dialogue – Dr Marcus Tomalin

Assessment: 100% coursework*

- Submitted report based on two practical exercises:
  - Gender Bias in Machine Translation
  - GPT-2 Task-Oriented Dialogue State Tracking
  - Due 26 April 2022
- HPC accounts, invitation instructions to be provided shortly

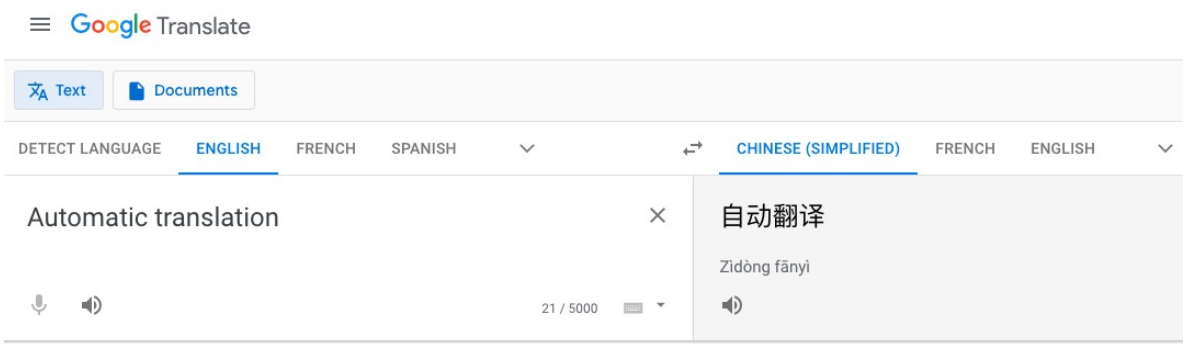
* Details will be posted on the module Moodle page

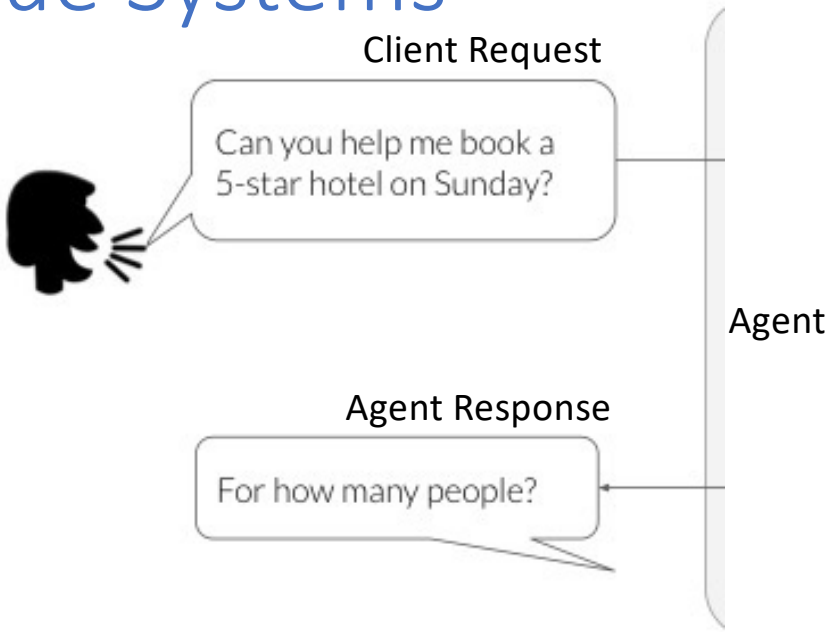
Today's Lecture

- Why study Dialogue Systems and Machine Translation in the same course?
  - i.e. why are we here?
- Overview of Machine Translation*
  - history, domains, where to find translations – old and new
- Why Translation is Difficult
  - morphology, domain, genre, word order, inherent variability, philosophical impossibility, ...

* D. Bellos. `Is That a Fish in Your Ear?: Translation and the Meaning of Everything`, 2012

Machine Translation and Dialogue Systems





Machine Translation (MT) is the full or partial automation of translation as carried out by humans

Dialogue Systems (DS) are computer systems that enable human-computer interaction using natural language

Both are sequence-to-sequence problems: map strings from a source language to a target language

- MT generates a translation of a source sentence
- DS generates a response to a client request

Different problem domains, but can be approached using the same modelling tools and tricks [1]

[1] Vaswani et al. Attention is All you Need, NIPS'17

Machine Translation

Machine Translation is the full or partial automation of translation as carried out by humans.

This is a difficult objective to define in absolute terms, since human translation often involves multiple overlapping tasks:

- Translation
- Interpretation
- Transliteration
- Summarization
- Glossing
  - Syntactic and Semantic Analysis
  - Commentary

All of these tasks require human intelligence, sensitivity, and understanding clearly beyond the state-of-the-art in NLP and artificial intelligence.

However translated texts are readily available in a variety of domains

- Religious texts
- Governmental proceedings
- Business documents
- Literary texts
- News sources
- Consumer
- Military and humanitarian resources

These provide ready sources of data for system building, evaluation, and deployment

Translation vs Interpretation

Translation: the interpreting of the meaning of a text and the subsequent production of an equivalent text, called a ”translation,” that communicates the same message in another language

Interpretation: the intellectual activity that consists of facilitating oral or sign-language communication, either simultaneously or consecutively, between two or among three or more speakers who are not speaking, or signing, the same language

The Eadwine Psalter
Canterbury, Christ Church, mid-twelfth century

…. Of all surviving twelfth-century English manuscripts, this is the most complex in design.

Three Latin versions of the Psalms laid in parallel columns are integrated with a Latin commentary, Old English and Anglo-Norman translations written between the lines and in the margins. Each Psalm opens with a magnificent drawing inspired by its text. 'By the waters of Babylon' illustrates Psalm 136 shown here.

The page layout remains lucidly clear, despite its complexity and the large team involved in the project – at least ten scribes and six artists. …

The Cambridge Illuminations
The Fitzwilliam Museum

The Origins of Modern Machine Translation

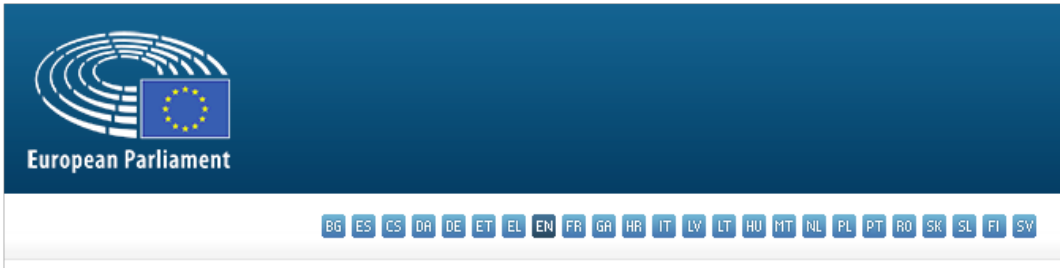
Introduction of extempore simultaneous interpretation at the Nuremberg Trials

… holding a trial that was "fair" and "expeditious" meant speeding up translations of the four languages of the nations involved: English, German, Russian and French. The solution was thought up by Col. Leon Dostert. …

… This was 1945, so digital recordings and tapes weren't around. But Dostert pressed on and consulted with IBM to develop a system of microphones and headsets to transmit the cacophony of languages. He hired interpreters and practiced this new type of interpreting with them. …

https://www.pri.org/stories/2014-09-29/how-do-all-those-leaders-un-communicate-all-those-languages

European Parliamentary Proceedings



TITLE VII : SESSIONS

CHAPTER 3 : GENERAL RULES FOR THE CONDUCT OF SITTINGS

Rule 167 : Languages

1. All documents of Parliament shall be drawn up in the official languages.

2. All Members shall have the right to speak in Parliament in the official language of their choice. Speeches delivered in one of the official languages shall be simultaneously interpreted into the other official languages and into any other language that the Bureau may consider to be necessary.

3. Interpretation shall be provided in committee and delegation meetings from and into the official languages that are used and requested by the members and substitutes of that committee or delegation.

TITRE VII : SESSIONS

CHAPITRE 3 : RÈGLES GÉNÉRALES POUR LA TENUE DES SÉANCES

Article 167 : Régime linguistique

1. Tous les documents du Parlement sont rédigés dans les langues officielles.

2. Tous les députés ont le droit, au Parlement, de s'exprimer dans la langue officielle de leur choix. Les interventions dans une des langues officielles sont interprétées simultanément dans chacune des autres langues officielles et dans toute autre langue que le Bureau estime nécessaire.

3. L'interprétation est assurée, au cours des réunions de commission et de délégation, à partir des langues officielles utilisées et exigées par les membres titulaires et les membres suppléants de la commission ou de la délégation concernée, et vers ces langues.

https://www.europarl.europa.eu/doceo/document/RULES-9-2019-07-02-RULE-167_EN.html
https://www.europarl.europa.eu/doceo/document/RULES-9-2019-07-02-RULE-167_FR.html

Introduction

Since November 2007 the European Commission's Directorate-General for Translation has made its multilingual Translation Memory for the Acquis Communautaire, DGT-TM, publicly accessible in order to foster the European Commission's general effort to support multilingualism, language diversity and the re-use of Commission information. …

DGT's Translation Memory

This extraction of aligned sentences can be used to produce a parallel multilingual corpus of the European Union's legislative documents (Acquis Communautaire) in 24 EU languages. …

OPUS ... the open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving ...
Contributions are very welcome! Please contact <jorg.tiedemann@helsinki.fi >

Search & download resources: -- select -- -- select -- all show all versions

Latest News

- 2020-06-30: New: OPUS-100 corpus
- 2020-05-22: New: ELRC public
- 2019-10-16: New: MultiParaCrawl
- 2019-10-14: New: Infopankki v1
- 2019-09-28: Update: ParaCrawl v5
- 2019-08-28: JW300 corpus added
- 2019-08-14: Various new and updated corpora
- 2018-10-06: New corpus: memat (Xhosa/English)
- 2018-02-15: New corpora: ParaCrawl, XhosaNavy
- 2017-11-06: New version: OpenSubtitles2018
- 2017-11-01: New URL: http://opus.nlpl.eu
- 2016-01-08: New version: OpenSubtitles2016

Search & Browse

- OPUS multilingual search interface
- Europarl v7 search interface
- Europarl v3 search interface
- OpenSubtitles search interface
- OpenSubtitles 2018 search interface
- EUconst search interface
- Word Alignment Database (old DB)

Tools & Info

- OPUS Wiki
- OPUS API
- OPUS interface (@github)
- OPUS translator (@github)
- OPUS tools (Python package)
- OPUS tools (Perl package)
- Uplug (@bitbucket)

Sub-corpora (downloads & infos):

- ada83 - Ada 83 manuals
- Bianet - Translated Turkish articles
- Bible (uedin) - Collection of Bible translations
- Books - A collection of translated literature
- CAPES - Thesis and dissertation abstracts
- DGT - A collection of EU Translation Memories provided by the JRC
- DOGC - Documents from the Catalan Goverment
- ECB - European Central Bank corpus
- EhuHac - Hizkuntzen Arteko Corpusa
- EiTB-ParCC - Parallel Corpus of Comparable News
- Elhuyar corpus
- ELRC public corpora
- EMEA - European Medicines Agency documents
- MultiParaCrawl - Non-English Bitexts from ParaCrawl
- MultiUN - Translated UN documents
- News Commentary (v11, v9.1,v9)
- OfisPublik - Breton - French parallel texts
- OO - the OpenOffice.org corpus (v2)
- OpenSubtitles - translated subtitles (v1, v2011, v2012, v2013, v2016)
- OPUS-100 corpus
- ParaCrawl corpus
- ParCor - A Parallel Pronoun-Coreference Corpus
- PHP - the PHP manual corpus
- QED - The QCRI Educational Domain Corpus
- Regeringsförklaringen - a tiny example corpus
- The sardware corpus
- SciELO - Artciles from SciELO
- SETIMES - A parallel corpus of the Balkan

Product Localisation

Mac iPad iPhone Watch TV Music Support

Country / Region
United States

AirPods Pro Service Program for Sound Issues

Apple has determined that a small percentage of AirPods Pro may experience sound issues. Affected units were manufactured before October 2020.

An affected AirPods Pro may exhibit one or more of the following behaviors:

- Crackling or static sounds that increase in loud environments, with exercise or while talking on the phone
- Active Noise Cancellation not working as expected, such as a loss of bass sound, or an increase in background sounds, such as street or airplane noise

Apple or an Apple Authorized Service Provider will service the affected AirPods Pro (left, right or both), free of charge.

Note: No other AirPods models are part of this program

أماكن الشراء الدعم Music TV Watch iPhone iPad Mac

البلد / المنطقة
الإمارات العربية المتحدة

برنامج خدمة AirPods Pro لمشاكل الصوت

تبيّن لشركة Apple أن نسبة صغيرة من AirPods Pro قد تكون بها مشاكل في الصوت. يعود تاريخ تصنيع الوحدات المتضررة إلى ما قبل أكتوبر 2020.

قد تظهر حالة واحدة أو أكثر من الحالات التالية على AirPods Pro المتضررة:

- أصوات طقطقة أو أصوات ساكنة ترتفع في البيئات عالية الصوت أو عند التمرين أو أثناء التحدث في الهاتف
- "إلغاء الضجيج النشط" لا يعمل كما هو متوقّع، مثل فقدان صوت الجهير أو ارتفاع مستوى الأصوات في الخلفية، مثل ضوضاء الشارع أو الطائرة

ستقوم Apple أو مقدم خدمة معتمد من Apple بخدمة AirPods Pro (اليسرى أو اليمنى أو كليهما) المتضررة مجانًا.

ملاحظة: لا يُعد أي من طرازات AirPods الأخرى مؤهلاً للاستفادة من هذا البرنامج

UNIVERSITY OF CAMBRIDGE

Social Media

Rapid adoption of social media is creating new text resources

- Twitter, Facebook, chatrooms, ...
- Short, noisy messages that are essentially untranslatable without context
- Qualitatively different from earlier collections – complete lack of editorial control

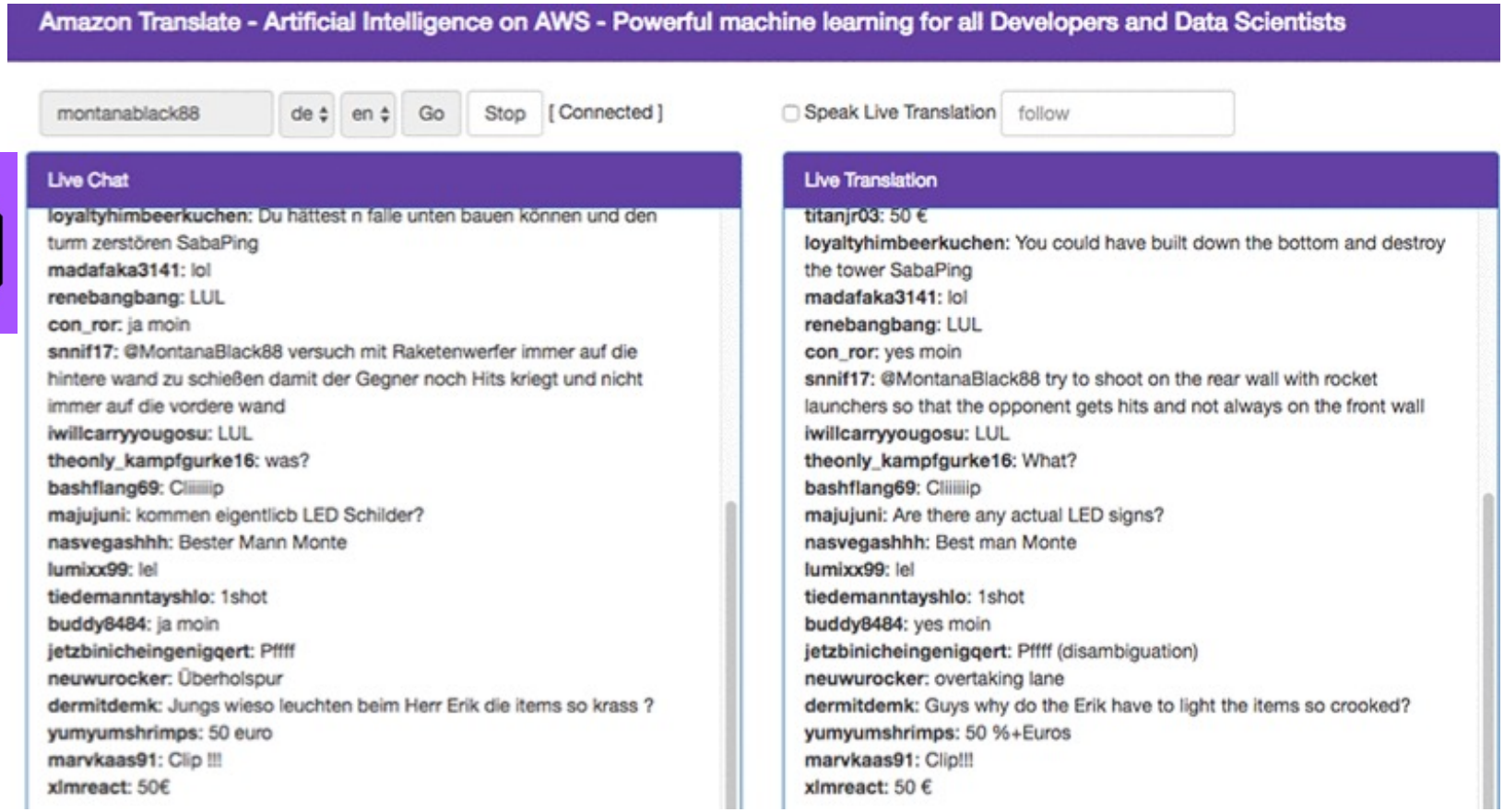
Spanish

- ahhhhhhhhhhhhhhhhhhhhhh siiiiill omg omg omg por diooooooos no lo puedes creer!!! espaa es tan pero tan chipadrevere!!!! y algo mas... me quedo hasta el savadoooo :D :D :D nena cuidate oiste... ojala la estes pasando super bien ok??? ha
- su musica es muy hermosalos adoro
- aww qe lindo0 esteban io tmb te qiero baby :)
- ellos practicaron duro aller
- me duele saber ke nunca te importe
- muy difisil
- hola,com estas?

English

- iam off tomarrow
- Eh! Waiter servame a beer more
- HI AS THESE? My name is EDUARDO I'm FROM MEXICO AND I wanted THANK YOU FOR AGREGARME..HAVE MSN? AH if it did not understand WAS THE FAULT OF THE TRANSLATOR. 34 AND I HOPE KNOW YOU SOON.KISSES
- im trying to write in korean
- he's def a cutie!!

Gaming


Amazon Translate - Artificial Intelligence on AWS - Powerful machine learning for all Developers and Data Scientists
montanablack88
de
en
Go
Stop
[Connected]
Speak Live Translation
follow
Live Chat
loyaltyhimbeerkuchen: Du hättest n falle unten bauen können und den turm zerstören SabaPing
madafaka3141: lol
renebangbang: LUL
con_ror: ja moin
snnif17: @MontanaBlack88 versuch mit Raketenwerfer immer auf die hintere wand zu schießen damit der Gegner noch Hits kriegt und nicht immer auf die vordere wand
iwillcarryyougosu: LUL
theonly_kampfgurke16: was?
bashflang69: Cliiiiip
majujuni: kommen eigentlicb LED Schilder?
nasvegashhh: Bester Mann Monte
lumixx99: lel
tiedemanntayshlo: 1shot
buddy8484: ja moin
jetzbinicheingenigqert: Pffff
neuwurocker: Überholspur
dermitdemk: Jungs wieso leuchten beim Herr Erik die items so krass ?
yumyumshrimps: 50 euro
marvkaas91: Clip !!!
xlmreact: 50€
Live Translation
titanjr03: 50 €
loyaltyhimbeerkuchen: You could have built down the bottom and destroy the tower SabaPing
madafaka3141: lol
renebangbang: LUL
con_ror: yes moin
snnif17: @MontanaBlack88 try to shoot on the rear wall with rocket launchers so that the opponent gets hits and not always on the front wall
iwillcarryyougosu: LUL
theonly_kampfgurke16: What?
bashflang69: Cliiiiip
majujuni: Are there any actual LED signs?
nasvegashhh: Best man Monte
lumixx99: lel
tiedemanntayshlo: 1shot
buddy8484: yes moin
jetzbinicheingenigqert: Pffff (disambiguation)
neuwurocker: overtaking lane
dermitdemk: Guys why do the Erik have to light the items so crooked?
yumyumshrimps: 50 %+Euros
marvkaas91: Clip!!!
xlmreact: 50 €

Why is Machine Translation Difficult?

Variations within language pairs and translations domains challenge even bilingual humans

- Domain and Genre
- Word sense
- Morphology
- Word Order
- Idiomatic expressions
- Language specific issues
  - Chinese – word segmentation
  - Arabic – tokenization
- Translation vs Transliteration

Constructing models for use in statistical machine translation is computationally difficult

- Translation does not proceed in left-to-right fashion due to word reordering between language pairs
  - unlike Automatic Speech Recognition
- Large amounts of monolingual and bilingual text are needed for parameter estimation

Much of the material in this section was taken from Jurafsky and Martin, Chapter 21 (2nd edition)

Translation should respect Domain and Genre

Religious texts, legal documents, business correspondence, technical documents, news sources, etc. are sometimes referred to as sublanguages

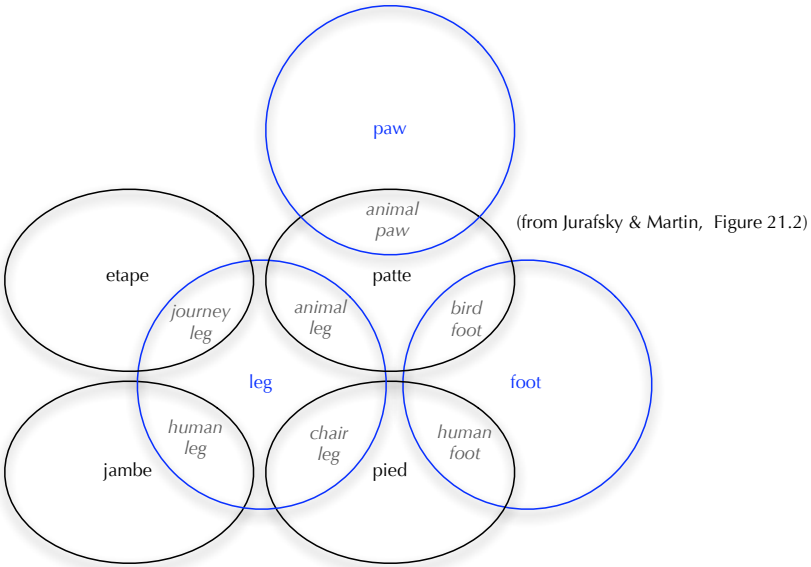
- Sublanguage domains differ across any of a range of dimensions
  - Specialized vocabularies
  - Stylistic differences, e.g. active vs. passive voice
  - Use of headlines, passage and chapter numbers, web addresses, proper names, ...
  - Variable sentence lengths, e.g. weather forecasts vs. news stories
- Care must be taken in using translation systems for multiple genres
  - Text from one genre may not be suitable for building systems in another genre

Translation should also respect user preferences, when available, such as gender, formality, …

Word Sense and Context: Translation Must Depend on Word Sense

Ambiguous words should be translated based on their *sense* : e.g. bank, plant, leg, ...

- Should an instance of *leg* be translated as *etape , pied , patte ,* or *jambe* ?



Fortunately, word sense is often (easily) determined by *context*, even for idioms

you're just pulling my leg (English)

mi stai prendeno per il naso (Italian)

me est tomando el pelo (Spanish)

Morphology

Morphological variation across languages leads to modelling difficulties in translation

- Morphological analysis is typically applied prior to processing
- Surface variability leads to sparsity of individual tokens in text translations

Example of Arabic tokenization and morphological analysis

وينتج المفاعل البلوتونيوم اللازم لتصنيع القنبلة الذرية .

[and-produces] [the-reactor] [the-plutonium] [the-required] [to-build] [the-bomb] [the-atomic] . ⇐ *gloss*

wyntj AlmfAEl Alblwtwnywm AllAzm ltSnyE Alqnblp Al*ryp . ⇐ *Romanized text*

w+ yntj AlmfAEl Alblwtwnywm AllAzm l+ tSnyE Alqnblp Al*ryp . ⇐ *MADA Morphological Analyzer*

Sources of variation

- Polysynthetic languages – *Siberian*
  - A single word might correspond to a single English sentence
- Agglutinative languages – *Turkish*
  - Morphemes are segmentable
- Fusion languages – *Russian*
  - affixes and prefixes carry syntactic meaning

Movement – Variations in Word Order

Languages are often classified by the ordering of their subject, verb, and object clauses .

- S-V-O – English
- S-O-V – Japanese
- V-S-O – Arabic

For example, in translating from Arabic into Japanese, the verb might have to be 'moved' from a sentence initial position to a sentence final position.

Movement leads to computational difficulties in machine translation

- Ordering changes in translation are not deterministic
- The number of possible reorderings grows exponentially with sentence length
- Considering arbitrary reorderings in automatic translation can be NP-Complete*

* K. Knight, "Decoding Complexity in Word-Replacement Translation Models," Computational Linguistics, 25(4), 1999

Pronouns

- Japanese is an example of a **pro-drop** language
- *Kono kēki wa oishii. Dare ga yaita no?*
  - This cake TOPIC tasty. Who SUBJECT made?
  - This cake is tasty. Who made **it**?
- *Shiranai. Ki ni itta? know-NEGATIVE. liked?*
  - **I** don't know. Do **you** like it?

- Some languages like Spanish can drop subject pronouns
  - In Spanish the verbal inflection often indicates which pronoun should be restored (but not always)
    - -o=I
    - -as = you
    - -a = he/she/it
    - -amos = we -an they

examples from Wikipedia

Different Tenses

- Spanish has two versions of the past tense: one for a definite time in the past, and one for an unknown time in the past
- When translating from **English to Spanish** we need to choose which version of the past tense to use

Transliteration

Transliteration is writing a word from one language using the *closest* corresponding orthography of a different language

- Closeness is often defined ‘phonetically’ (or syllabically)
- Transliteration attempts to describe the pronunciation of a foreign word

  “Bush” -> /bu shu/ “Clinton” -> /kk lin dun/ Obama -> /ou ba ma/
- A written form with that pronunciation is chosen as the transliteration

Transliterations are rarely unique and can be difficult to spot in text

- “Kosovo” can be transliterated into Chinese in several ways

  科索沃 /ke-suo-wo/, 科索佛 /ke-suo-fo/,
  科索夫 /ke-suo-fu/, 科索伏 /ke-suo-fu/, or
  柯索佛 /ke-suo-fo/.

- All of the above should be translated by into English as ”Kosovo”

UNIVERSITY OF CAMBRIDGE

Meng et al., *Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval*, ASRU 2001 https://ieeexplore.ieee.org/abstract/document/1034649

Codeswitching in translation

In linguistics, **code-switching** or language alternation occurs when a speaker alternates between two or more languages, or language varieties, in the context of a single conversation.

From the English language version of **Terminator 2**:

- *John Connor: No, no, no, no. You gotta listen to the way people talk. ...*
  *And if you want to shine them on, it's **"hasta la vista, baby"**.*
- *The Terminator: **Hasta la vista, baby**.*

In the European Spanish version of the film, "**Hasta la vista, baby**" was dubbed as "**Sayonara, Amigo**."

https://en.wikipedia.org/wiki/Code-switching

Variability of human translation

It would appear that a speech made at the weekend by Mr Fischler indicates a change of his position.

1. Por lo visto el sr. Fischler pronunci un discurso este fin de semana en el que pareca haber cambiado de actitud.
2. Parece que el Sr. Fischler, en su discurso del fin de semana, mostr un cambio de opinin.
3. Parece que el sr. Fischler ha cambiado de opinin, segn se desprende de su discurso del
4. pasado fin de semana.
5. Pareciera que el discurso realizado por el seor Fischler el fin de semana indica un cambio de opinin.
6. Parece que el discurso que hizo el sr. Fischler el fin de semana sugiere un cambio en su postura.

Discourse and Meaning

‘Truth values’ can be difficult to maintain at all possible levels of translation

- *This sentence has five words* ✓
- В этом предложении пять слов ✓
- Это предложение состоит из пяти слов ✗

Translation is Impossible - but possibly easier with more data

Schultz poses a quiz after W.V.O.Quine on “the indeterminacy of translation”

You are travelling in the country of Quine. A speaker of Quinean shows you this picture and says, `This is a Gavagai.`

What is a Gavagi?

A. Grass
B. Rabbit plus grass
C. Dinner






From Kathryn Schultz, *Being Wrong: Adventures in the Margin of Error*, 2010

Q. Why Statistics for Machine Translation?
A. Human Understanding of Translation Offers Little Guidance

An example: A discussion on the importance of preserving length in translation:

> *a translation must be ‘ quantitatively equivalent’ to the original … It is not the problem of counting the number of signs, of signifiers and signified, but rather of counting the number of words …*
>
> *There is an ideal law, even though inaccessible, not to translate ‘ word per word’ or ‘ word by word’, but to keep close as far as it is possible to the equivalent of a word through another word.”*

One interpretation of this: translation length should be based on words, not concepts

But inaccessible ideals laws aren’t much help in building computing systems …

… its easier to work from examples

J. Derrida 1998, quoted by Umberto Eco in *Mouse or Rat?* , 2003

Today's Lecture - recap

- Why study Dialogue Systems and Machine Translation in the same course?
  - i.e. why are we here?
- Overview of Machine Translation
  - History, Translation Domains
- Why Translation is Difficult
  - Morphology, domain, genre, word order, inherent variability, philosophical impossibility, ...

Next Lecture - preview

- Introduction to Dialogue Systems
  - types of dialogue systems, characteristics of dialogues, turn taking, ...
- Task-oriented dialogue systems
  - system architecture, dialogue acts
- Natural language understanding, dialogue state tracking, natural language generation