Concepts and Applications in NLP

Machine Translation

Marion Di Marco

December 10, 2024

Introduction

- MT task: translate a text into English:

> Id-Dinja tagħmel parti mis-sistema solari, li fiċ-ċentru tagħha tinsab ix-xemx li għandha 99.86% mill-massa tas-sistema solari kollha. Il-kamp gravitazzjonali assoċjat mal-massa tax-xemx jiġbed il-bqija tal-kostiwenti l-oħra, inkluża id-Dinja, jorbitaw madwarha. Il-maġġor parti tal-oġġetti jinsabu fuq l-istess pjan, jorbitaw max-xemx fl-istess direzzjoni.

- One of the oldest problems in Artificial Intelligence

Outline

A Very Brief History of Machine Translation

- Machine translation was one of the first applications envisioned for computers
- Warren Weaver (1949):
  "I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text."
- IBM (1954): basic word-for-word translation system

A Very Brief History of Machine Translation

- 1970ies: Rule-based MT
  - parse source-sentences with a rule-based parser
    (finite-state based) morphological analysis
  - transfer source syntax structure → target-language representation
    hand-written rules
  - generate text from target-language representation
- 2000: Statistical Machine Translation (SMT)
  - relies on corpus statistics, no linguistic information
- 2016: Neural Machine Translation (NMT)
  - relies on corpus statistics, no linguistic information
  - based on deep learning techniques
  - sequence-to-sequence models, attention mechanisms, transformers
- Now: also Large Language Models

Machine Translation Approaches

- "Vauquois triangle"

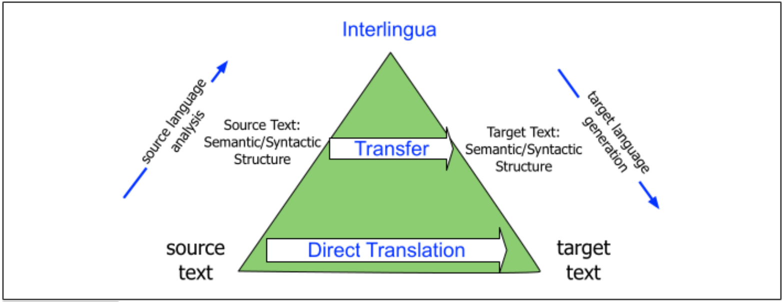
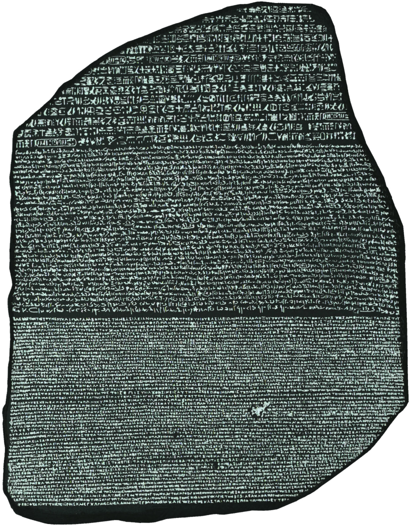


Figure 13.13 The Vauquois (1968) triangle.

- Direct translation: word-by-word, based on dictionaries
- Interlingua: language-independent representation scheme
- Depth of analysis ↔ amount of transfer knowledge

Figure from Jurafsky & Martin

Parallel Data: Rosetta Stone



- The Egyptian language was a mystery for centuries
- A stone with Egyptian text and its Greek translation was found (1799)
- We can *learn* how to translate Egyptian!

Figure from https://de.wikipedia.org/wiki/Stein_von_Rosette

Parallel Data

- Europarl:

Ich habe mich bei der gemeinsamen Entschließung zur Bonner Konferenz über den Klimawandel der Stimme enthalten.	I abstained on the joint resolution on the conference on climate change.
...	...
Nach mehr als fünf Jahre währender Vorbereitung haben wir nun heute endlich über den Vorschlag für eine Richtlinie des Rates in Bezug auf Konfitüren, Gelees, Marmeladen und and Maronenkrem abgestimmt.	After more than five years in the pipeline, we have finally voted today on the proposal for a Council directive relating to fruit jams, jellies, marmalades sweetened chestnut purée.
Jammy dodgers sind eine schöne britische Institution, und viele im Vereinigten Königreich hatten befürchtet, die Richtlinie würde zu einem Verbot dieses Gebäcks führen.	The jammy dodger is a fine British institution and many in the UK had feared that the directive would result in the outlawing of this biscuit.
...	...
Sie haben es ja schon gesagt, der Marktanteil europäischer Filme in den Kinos der Europäischen Union befindet sich mit nur 22,5 % auf einem historischen Tiefstand.	As you said, the market share of European films in the cinemas of the European Union is at an historic low point of only 22.5 %.

- For many language: large parallel corpora available
- Europarl, CommonCrawl, NewsCommentary, WikiTitles, United Nations Parallel Corpus, Open Subtitles, ParaCrawl, ...

Parallel Data for MT Training

- Machine translation models are trained on parallel data
- Standard training data: aligned pairs of parallel sentences
- Simplification: translate each sentence independently
  → we only consider individual sentences
- Phrase-Based Statistical Machine Translation:
  word alignment in within parallel sentence pairs
- But: it is not always that easy ...

Parallel Data: Sentence Alignment

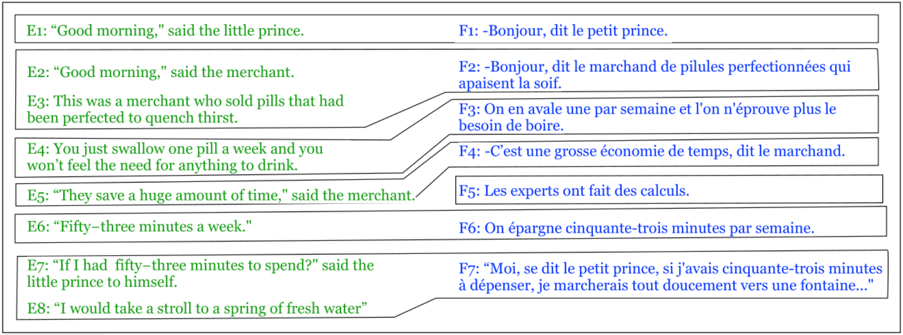


Figure 13.4 A sample alignment between sentences in English and French, with sentences extracted from Antoine de Saint-Exupery's *Le Petit Prince* and a hypothetical translation. Sentence alignment takes sentences $e_1, \ldots, e_n$, and $f_1, \ldots, f_n$ and finds minimal sets of sentences that are translations of each other, including single sentence mappings like (e_1,f_1), (e_4,f_3), (e_5,f_4), (e_6,f_6) as well as 2-1 alignments (e_2/e_3,f_2), (e_7/e_8,f_7), and null alignments (f_5).

Figure from Jurafsky & Martin

Outline

Language Divergences and Typology

- There are about 7000 languages
- Some aspects about language seem to be universal
  - words for referring to people, for talking about eating and drinking
  - every language seems to have nouns and verbs
- Languages can differ in many ways
- Idiosyncratic differences
  to be dealt with one by one, e.g. lexical differences
- Systematic differences
  can be modeled in a general way, e.g. adjective before or after the noun
- More information: WALS, the World Atlas of Language Structures

Word Order Typology

- Word order of verbs, subjects, and objects in declarative clauses
  - SVO: subj-verb-obj (e.g. German, French, English, and Mandarin)
  - SOV: subj-obj-verb (e.g. Hindi and Japanese)
  - VSO: verb-subj-obj (e.g. Irish and Arabic)
- Languages sharing the same word order often have other similarities
  - VO languages often have prepositions
  - OV languages often have postpositions

English:	*He wrote a letter to a friend*			
Japanese:	*tomodachi*	*ni*	*tegami-o*	*kaita*
	friend	to	letter	wrote
Arabic:	*katabt*	*risāla*	*li*	*ṡadq*
	wrote	letter	to	friend

Word Order Typology

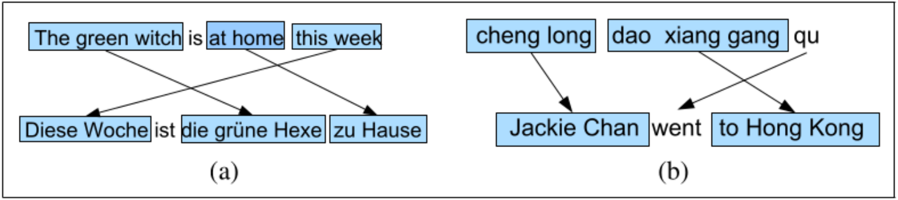


Figure 13.2 Examples of other word order differences: (a) In German, adverbs occur in initial position that in English are more natural later, and tensed verbs occur in second position. (b) In Mandarin, preposition phrases expressing goals often occur pre-verbally, unlike in English.

Figure from Jurafsky & Martin

Lexical Divergences

- Word sense disambiguation
  - EN *bass* → fish (ES: *lubina*) or musical instrument (ES: *bajo*)

- Word senses depending on context
  - EN *wall* → DE *Wand* (walls inside a building)
    DE *Mauer* (walls outside a building)
  - EN *brother* → distinct words for older/younger brother in many other languages

- Lexical gaps
  - DE *Schadenfreude* → pleasure in someone else's misfortune
  - IS *gluggaveður* '*window-weather*' → weather that is best enjoyed from indoors, looking through the window

Lexical Divergences

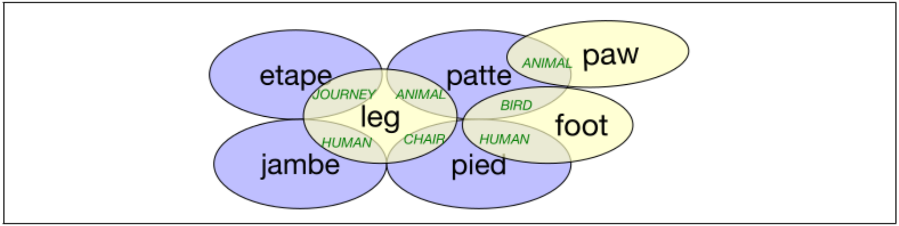


Figure 13.3 The complex overlap between English *leg*, *foot*, etc., and various French translations as discussed by Hutchins and Somers (1992).

Figure from Jurafsky & Martin

Lexical Divergences

- Differences in in how the conceptual properties of an event are mapped onto specific words
- Marking of *direction of motion* and *manner of motion* on the verb vs. a 'satellite'
  - EN: *the bottle floated out*.
  - ES: *la botella salió flotando.*
the bottle exited floating.
  - DE: *Pierre durchschwimmt den Fluß.*
Pierre through-swims the river.
  - FR: *Pierre traverse la rivière en nageant.*
Pierre crosses the river swimming.
- "Chassé-croisé"

Grammatical Constraints

- Explicit marking of number
- Explicit marking of grammatical gender on nouns and adjectives
- Marking of grammatical gender → grammatical gender on pronouns
  - DE: *Die Katze spielt mit der Maus.* **Sie** *mag das nicht.*
    The cat$_{she}$ plays with the mouse$_{she}$. **She** doesn't like this.
  - FR: *Le chat joue avec la souris.* **Il/Elle** *n'aime pas cela.*
    The cat$_{he}$ plays with the mouse$_{she}$. **He/She** doesn't like this.
- Level of politeness, e.g. Japanese

Example from: http://static.lingenio.de/Publikationen/Eberle_Integration_JLCL09.pdf

Morphological Typology

- Two dimensions of morphological variations
- Morphemes per word
  - isolating languages: one word – one morpheme
  - (poly)synthetic languages: one word may have (very) many morphemes
- Are morphemes segmentable?
  - agglutinative: morphemes have relatively clear boundaries
  - fusional: a single affix may conflate multiple morphemes
- Translating between morphologically rich languages:
  need to deal with structure below word level
- Subword tokenization in NMT: for example BPE (not ideal!)

Referential Density

- Some information is not always explicit, for example pronouns
  - some languages require a pronoun when talking about a referent
  - in some languages, pronouns can sometimes be omitted

[El jefe]$_i$ dio con un libro. $\emptyset_i$ Mostró su hallazgo a un descifrador ambulante.
[The boss] came upon a book. [He] showed his find to a wandering decoder.

- Pro-drop languages can omit pronouns, with varying degrees

- Referentially dense $\leftrightarrow$ referentially sparse

- Translating from languages with extensive pro-drop:
  (i) identify the zero-pronoun and (ii) fill it correctly

Translational Divergences: Example

- Between different languages: collection of translational diverences

大会/General Assembly 在/on 1982年/1982 12月/December 10日/10 通过了/adopted 第37号/37th 决议/resolution ，核准了/approved 第二次/second 探索/exploration 及/and 和平peaceful 利用/using 外层空间/outer space 会议/conference 的/of 各项/various 建议/suggestions 。

On 10 December 1982 , the General Assembly adopted resolution 37 in which it endorsed the recommendations of the Second United Nations Conference on the Exploration and Peaceful Uses of Outer Space .

- Sentence from the United Nations
  - word order: date, noun phrase *peaceful using outer space conference of various suggestions*
  - definite article *the* vs. none in Chinese
  - plural –*s* vs. modifier *various*
  - ...

Figure from Jurafsky & Martin

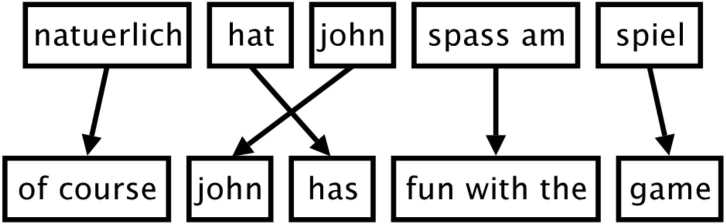
Outline

Phrase-Based Translation: Motivation

- Phrase-Based Models translate *phrases* as atomic units
- Advantages
  - many-to-many translation can handle non-compositional phrases
  - use of local context in translation
  - more data $\rightarrow$ learn longer phrases
- Phrases are extracted from word-aligned parallel data
- Decoder takes phrases and target-side language model and searches over translations
- Phrase-based translation was state-of-the-art for a long time before NMT
- Moses system: `http://www2.statmt.org/moses/`

Phrase-Based Translation: Idea

- Parallel sentence pairs with word alignemnt



- “Foreign” input (= source language) is segmented into phrases
- Each phrase is translated into English (= target language)
- Phrases are reordered

Figure from https://www2.statmt.org/book/slides/05-phrase-based-models.pdf

Phrase-Translation Table

- Main knowledge source: phrase translation probabilities

English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

- Phrase translations for *den Vorschlag* learned from the Europarl corpus
  - lexical variation *proposal* vs. *suggestions*
  - morphological variation *proposal* vs. *proposals*
  - included function words (*the*, *a*, ...)
  - noise (*it*)

Figure from https://www2.statmt.org/book/slides/05-phrase-based-models.pdf

Linguistic Phrases?

- The model is not limited to linguistic phrases
  such as noun phrases, prepositional phrases, ...
- Some non-linguistic phrase pair:
  - spass am → fun with
- Context information:
  Prior nouns often help with translation of preposition
- Experiments show that limitation to linguistic phrases hurts quality

Probabilistic Model

- Source language $\mathbf{f}$ (= foreign)
  target language $\mathbf{e}$ (= English)

- Bayes rule:

$$\begin{aligned} \mathbf{e}_{best} &= argmax_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) \\ &= argmax_{\mathbf{e}} p(\mathbf{f}|\mathbf{e}) p_{LM}(\mathbf{e}) \end{aligned}$$

  - translation model $p(\mathbf{e}|\mathbf{f})$
  - language model $p_{LM}(\mathbf{e})$

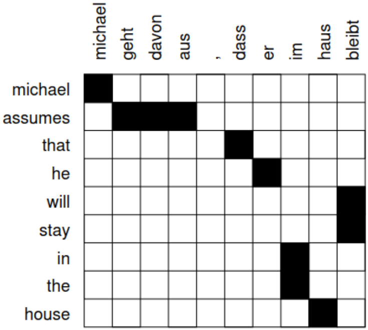
- **Translation model** → reproduce source-side content

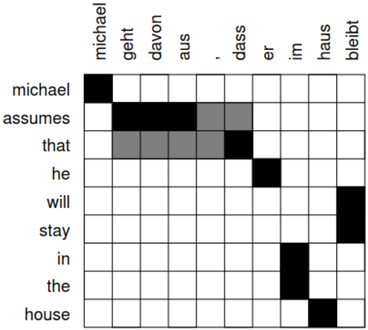
- **Language model** → make the output fluent English

- (Also: reordering model)

Phrase-Translation Table

- Learn phrase translations from parallel data
  - word alignment
  - extract phrase pairs
  - score phrase pairs ($\rightarrow$ translation probabilities)



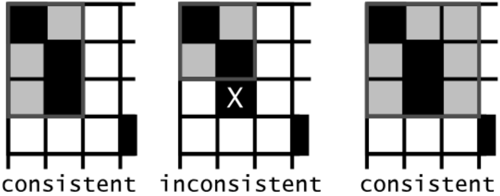


extract phrase pairs

consistent with word alignment

Figures from https://www2.statmt.org/book/slides/05-phrase-based-models.pdf

Extracting Phrase Pairs



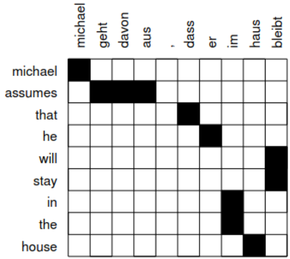
ok	**violated** one alignment point outside	**ok** unaligned word is fine

All words of the phrase pair have to align to each other.

Figure from https://www2.statmt.org/book/slides/05-phrase-based-models.pdf

Larger Phrase Pairs

Larger Phrase Pairs



michael assumes — michael geht davon aus / michael geht davon aus ,
assumes that — geht davon aus , dass ; assumes that he — geht davon aus , dass er
that he — dass er / , dass er ; in the house — im haus
michael assumes that — michael geht davon aus , dass
michael assumes that he — michael geht davon aus , dass er
michael assumes that he will stay in the house — michael geht davon aus , dass er im haus bleibt
assumes that he will stay in the house — geht davon aus , dass er im haus bleibt
that he will stay in the house — dass er im haus bleibt ; dass er im haus bleibt ,
he will stay in the house — er im haus bleibt ; will stay in the house — im haus bleibt

Figure from https://www2.statmt.org/book/slides/05-phrase-based-models.pdf

Scoring Phrase-Translation Pairs

- Phrase pair extraction: collect all phrase pairs from the data
- Phrase pair scoring: assign probabilities to phrase translations
- Score by relative frequency:

$$\phi(f|e) = \frac{count(e,f)}{\sum_{f_i} count(e,f_i)}$$

Weighted Model

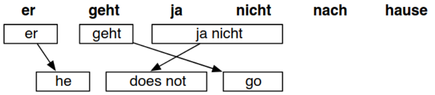
- The model consists of three sub-models
  - phrase translation model $\phi(f|e)$
  - reordering model d
  - language model $p_{LM}(e)$
- Add weights: $\lambda_\phi, \lambda_d, \lambda_{LM}$
- Such a weighted model is a log-linear model: $p(x) = exp \sum_{i=1}^{n} \lambda_i h_i(x)$
- More feature functions:
  - bidirectional translation probabilities $\phi(e|f)$ and $\phi(f|e)$
  - lexical weighting with word translation probabilities

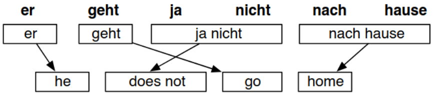
Phrase-Based Decoding

- Model
  - phrase-table: set of phrase pairs with translation probabilities $p(f|e)$
  - target-side n-gram language model:
  - reordering model

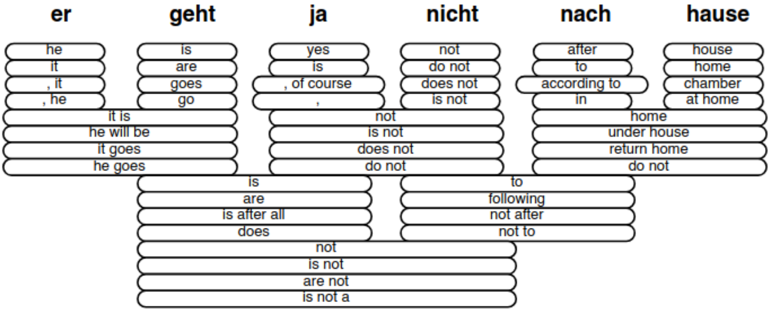
- For input **f**: find a sentence **e** produced by a series of phrase translations, including reordering

- Pick phrase in input, translate





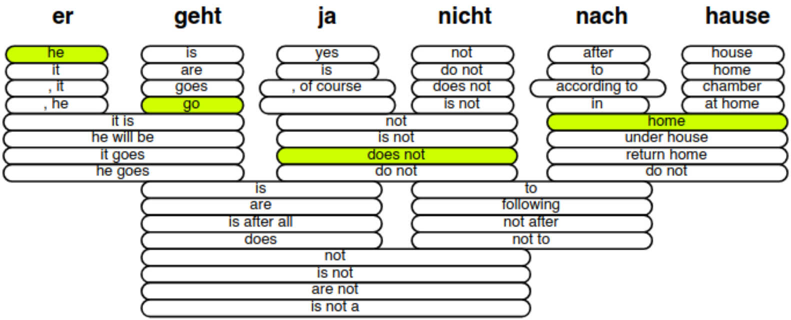
Translation Options



- Many translation options to choose from
  - Europarl phrase-table: 2727 matching phrase pairs for this sentence
  - pruning to the top 20 per phrase: 202 translation options remain

Slide from `https://www2.statmt.org/book/slides/06-decoding.pdf`

Translation Options



- The decoder does not know the right answer
  - pick the right translation option
  - arrange them in the right order

⇒ Search problem solved by heuristic beam search

Slide from `https://www2.statmt.org/book/slides/06-decoding.pdf`

Outline

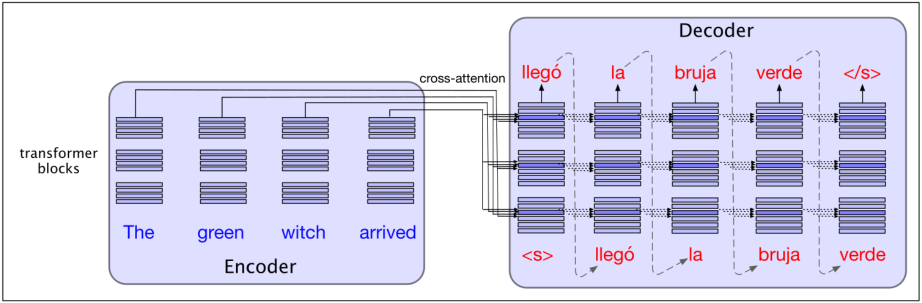
Neural Machine Translation

- Phrase-based MT was state-of-the-art until 2015/16
- Neural MT models can overcome some of the challenges of SMT
  - Training of one single end-to-end model vs. combination of several sub-models
  - Limited context size in SMT: n-gram LM and phrase length are a hard cut-off vs. attention in NMT that can focus on relevant context
  - NMT models can generalize better, SMT was more affected from rare words or phrases
- Encoder-decoder transformer

Neural Machine Translation

- Encoder-decoder models: very good at handling different types of translation divergencies
- Supervised machine learning: given a large set of parallel sentences, learn to map source sentences into target sentences
- Maximize the probability of target tokens $y_1, ..., y_m$ given a sequence of source tokens $x_1, ..., x_n$
- Encoder: takes the input words $x = [x_1, ..., x_n]$ and produces an output representation h
- Decoder: conditional language model that attends to encoder representation and generates target words
  At each timestep t: conditioning on source sentence and the previously generated target language words

Neural Machine Translation



- The encoder-decoder transformer architecture for machine translation
- Extra cross-attention layer: attend to all the encoder words

Figure from Jurafsky and Martin

Subword Segmentation

- Subword units are often based on WordPiece or BPE
  - handle unknown words
  - efficiency in training
- Frequency-based compression algorithms:
  - start with small vocabulary (character-level)
  - iteratively merge the most common tuples until desired vocabulary size is reached
  - Example:
    `t h e   c a t   s a t   o n   t h e   m a t`

    assuming "`t h`" is the most frequent tuple given an EN corpus:
    `th e   c a t   s a t   o n   th e   m a t`

  → keep frequent words intact, segment less frequent ones
- Example: playing → `play ##ing`

Subword Segmentation

- BPE: merges based on the most frequent set of tokens
- WordPiece: merges based on which one most increases the language model probability
- Unigram algorithm/SentencePiece:
  - start with a huge vocabulary: individual characters, frequent sequences of characters including space-separated words
  - estimating the probability of each token, tokenize the input data using various tokenizations, remove a percentage of tokens that don't occur in high-probability tokenization

Subword Segmentation

Original:	corrupted	Original:	Completely preposterous suggestions
BPE:	cor rupted	BPE:	Comple t ely prep ost erous suggest ions
Unigram:	corrupt ed	Unigram:	Complete ly pre post er ous suggestion s

- BPE tends to create lots of very small non-meaningful tokens
- BPE tends to merge very common tokens, like the suffix *ed*, onto their neighbor
- Unigram tends to produce tokens that are more semantically meaningful

Figure from Jurafsky and Martin

Subword Segmentation for Morphologically Rich Languages

- Frequency-based segmentation approaches are not optimal for morphologically rich languages
- Fail to fully capture the morphological complexities of words
- Cannot handle non-concatenative processes: $Apfel_{Sg} \rightarrow \ddot{A}pfel_{PL}$
- Previous research:
  evidence that linguistic guidance in segmentation can help
  for example, faster convergence, lower perplexity

  but: correlation with training data size
- What about multilingual models?

Outline

Evaluation

- MT output is evaluated along two dimensions
- **Adequacy**: how well the translation reproduces the content of the source sentence
- **Fluency**: how fluent the translation is in the target language (grammatical, clear, readable, natural)

- Human annotators to evaluate?
  - rate fluency/adequacy on a scale
  - ranking: given two sentences, which one is better?
- High-quality evaluation
  but: training and guidelines needed, expensive and slow

Automatic Evaluation

- Automatic metrics: less accurate, but fast
- Test potential system improvements,
  automatic loss function when training

- General idea for automatic metrics: compare with reference sentence(s)
- Intuition: a good translation contains characters and/or words
  occurring in a human translation
- Test set consists of source sentence, a gold target translation
  (reference) and an MT output (hypothesis)

Character Overlap: chrF

- **chrF: character F-score**: character n-gram overlaps with reference

  Popović (2015)

- Parameter k: length of the n-ngrams´
- **chrP**: percentage of character 1-grams, 2-grams, ..., k-grams in the hypothesis that occur in the reference, averaged
- **chrR**: percentage of character 1-grams, 2-grams,..., k-grams in the reference that occur in the hypothesis, averaged
- $chrF\beta = (1+\beta^2)\frac{chrP \cdot chrR}{\beta^2 \cdot chrP + chrR}$ for $\beta = 2$ (higher weight to recall)
- chrF is simple, robust and correlates well with human judgments in many languages

BLEU

- BLEU: word-based overlap metric — Papineni et al. (2002)
- Precision-based metric
- **Modified unigram precision**:
  - MT systems have a tendency to overgenerate reasonable words
  - a reference word is considered exhausted after matching with a candidate word

    Candiate: `the` `the` `the the the the the`
    Reference: `the` `cat sat on` `the` `mat`
- n-gram precision favors short sentences → **brevity penalty** to discount MT output shorter than the reference
- Word-based metric → sensitive to tokenization
- Computed at document-level

Limitations of Overlap Metrics

- chrF is very local: large phrase moved around does not change much
- BLEU does not work well with morphologically rich languages; cannot capture inflectional variants
- Dependent on reference (→ lexical choices, syntactic structure): cannot (sufficiently) capture synonyms or other valid variations
- METEOR: considers matches of synonyms
- Very strict criteria → a good translation may differ substantially from the reference

Embedding-Based Models

- Use BERT or other embeddings to measure similarity between reference and MT output
- Given a dataset with human assessments of translation quality $(x, \tilde{x}, r)$
  - reference translation $x = (x_1, ..., x_n)$
  - candidate translation $\tilde{x} = \tilde{x}_1, ..., \tilde{x}_m$
  - human rating score r
- Metrics like COMET or BLEURT: train a predictor Rei et al. (2020); Sellam et al. (2020)
  - pass x and $\tilde{x}$ through a version of BERT
  - linear layer that is trained to predict r
  - output correlates highly with human labels
- Without human-labeled data sets: measure similarity of x and $\tilde{x}$ by the similarity of their embeddings (BERTScore) Zhang et al. (2020)

Outline

Translation in Large Language Models

- LLMs implicitly learn a wide range of language tasks, including machine translation
- Translation study with GPT — Robinson et al. (2023)
  - high-resource languages: GPT models approach or exceed performance of MT models
  - low-resource languages: consistently worse than traditional MT models
  - resource level is the most important feature in determining GPT's relative translation ability

Outline

Credits

Content based on:

- Slides from Philipp Koehn:

  Statistical Machine Translation

  `https://aclanthology.org/www.mt-archive.info/Koehn-2008.pdf`

  Phrase-based models

  `https://www2.statmt.org/book/slides/05-phrase-based-models.pdf`

- Dan Jurafsky and James H. Martin (2024)
  Speech and Language Processing: Chapter 13
  `https://web.stanford.edu/ jurafsky/slp3/`

- Lecture slides from Alexander Fraser (Machine Translation; Computational Morphology and Electronic Dictionaries 2017)

References

- Nathaniel Robinson, Perez Ogayo, David R. Mortensen, Graham Neubig (2023). *ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages.* Proceedings of WMT 2023.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi (2020). *BERTScore: Evaluating Text Generation with BERT*. ICLR 2020.
- Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie (2020). *COMET: A Neural Framework for MT Evaluation*. EMNLP 2020.
- Thibault Sellam, Dipanjan Das, Ankur Parikh (2020). *BLEURT: Learning Robust Metrics for Text Generation*. ACL 2020.
- Maja Popović (2015). *chrF: character n-gram F-score for automatic MT evaluation.* WMT 2015.
- Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu (2002). *Bleu: a Method for Automatic Evaluation of Machine Translation.* ACL 2002.