Machine Translation: IBM Model 1

Instructor: Jackie CK Cheung and David Adelani

COMP-550

Readings: J&M 13-13.1 (3rd edition)

Outline

Machine translation

- Why is it a hard problem?
- Ineffability and the Sapir-Whorf hypothesis

The Vauquois triangle

The Noisy-Channel Model for MT

IBM Model 1

Machine Translation (MT)

Remember from the first lecture? MT is how NLP got started!

Automatische Textverarbeitung gefällt mir.

↕

I like natural language processing.

Why is MT Difficult?

Think about the domains of language we have discussed so far in this course. How do they affect MT?

- Morphology
- Syntax
- Semantics
- Pragmatics and discourse

Lexical Gaps

There is usually no one-to-one mapping between lexical items in different languages.

- Commonly cited examples include colours, kinship terms, hands/legs, but this is pervasive throughout language.

e.g.:

Pirahã does not distinguish mother and father

- *baíxi* (mother or father) (Everett, 2005)

Chinese (of various kinds) have no term for “brother”, “sister”, “aunt”, “uncle”, “grandmother”, …

- Must specify relative age, side of family, blood relation, etc.
- 嬸嬸 (father’s younger brother’s wife)

Pragmatic Variation: Presupposition

Some words, such as *again*, *stop*, or *more*, **presuppose**, or contain an assumption about the world.

Mark called again. presupposes that Mark called.

In English, the use of these words typically assume the presupposed information is in the **common ground** between the speaker and the hearer.

A: *Mark phoned again.*

B: *Again? I didn't know he phoned in the first place!*
(presupposition failure → challenge as response)

In St'àt'imcets, such usages do not elicit a challenge from the hearer. (Matthewson, 2006)

Other Examples

Morphological:

- Different levels of, requirements for inflection
- Number, tense, aspect marking
- Noun classes/grammatical gender

Syntactic:

- Word order differences; word order variability

Semantic:

- How spatial relations are grammatically distinguished

Pragmatic:

- Grammatical politeness

World Atlas of Language Structures:

- https://wals.info/

In General

Different languages require or allow different morphological/syntactic/semantic/discourse properties to be expressed explicitly.

They interact in different ways with:

- other linguistic aspects
- non-linguistic aspects/the overall culture of the speakers of the language

Sapir-Whorf Hypothesis

Is it even possible to produce a perfect translation?

Sapir-Whorf hypothesis:

- The language you speak affects your thoughts
- *Strong* version: Language determines and constrains all human actions and thoughts
- *Weak* version: Language may influence human actions and thoughts slightly, in highly specific ways.

Very few linguists believe in the strong version of this hypothesis. Some linguists reject any version of it.

Spatial Organization

Kuuk Thaayorre uses an absolute system (north, east, west, south), rather than a relative system (left, right, ahead):

The cup is southwest of the dinner plate.

- Speakers very good at navigation, orienting themselves.

Ask people to arrange events by time:

- English speakers: left to right
- Hebrew speakers: right to left
- Kuuk Thaayorre speakers: east to west

(Boroditsky and Gaby, 2010)

Now, back to MT…

Machine Translation

(Un)fortunately, we are not at the point of worrying about Sapir-Whorf in MT:

- Translating about events and participants
- Focusing on conveying high-level, literal meaning is already a huge challenge

How do we measure progress?

Machine Translation Evaluation

Difficult, much like automatic summarization evaluation

Many of the same issues with human evaluation

One key difference: less variation in desired output content

Most automatic measure: BLEU (Papieni et al., 2002)

$$BLEU_n = \frac{\sum_{C \in \{Cands\}} \sum_{ngram \in C} Count_{match}(ngram)}{\sum_{C \in \{Cands\}} \sum_{ngram \in C} Count(ngram)}$$

- BLEU is precision-oriented:
  - For each n-gram in the proposed translation, check if it is found in the reference translation.
  - In practice, BLEU incorporates an additional brevity penalty, and the geometric mean over several values of n is taken.

Other Metrics

METEOR (Banerjee & Lavie, 2005)

A score based on explicit word-to-word matches between output and reference

Translation Error Rate (TER) (Snover et al., 2006)

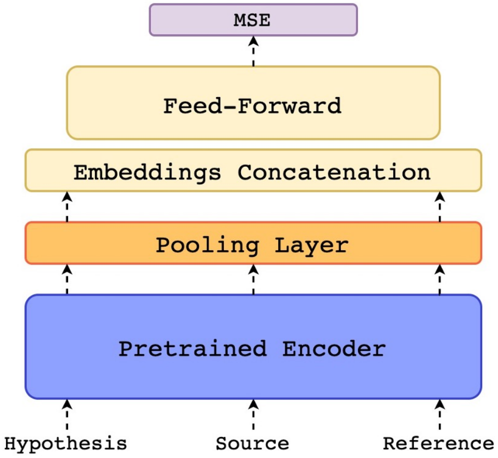
Number of edits required to change an MT output into the reference.

Character level F1-score (ChrF / ChrF++) (Popović, 2015)

Compares MT output and a reference translation using character n-grams

COMET (Rei et al., 2020)

An embedding-based evaluation

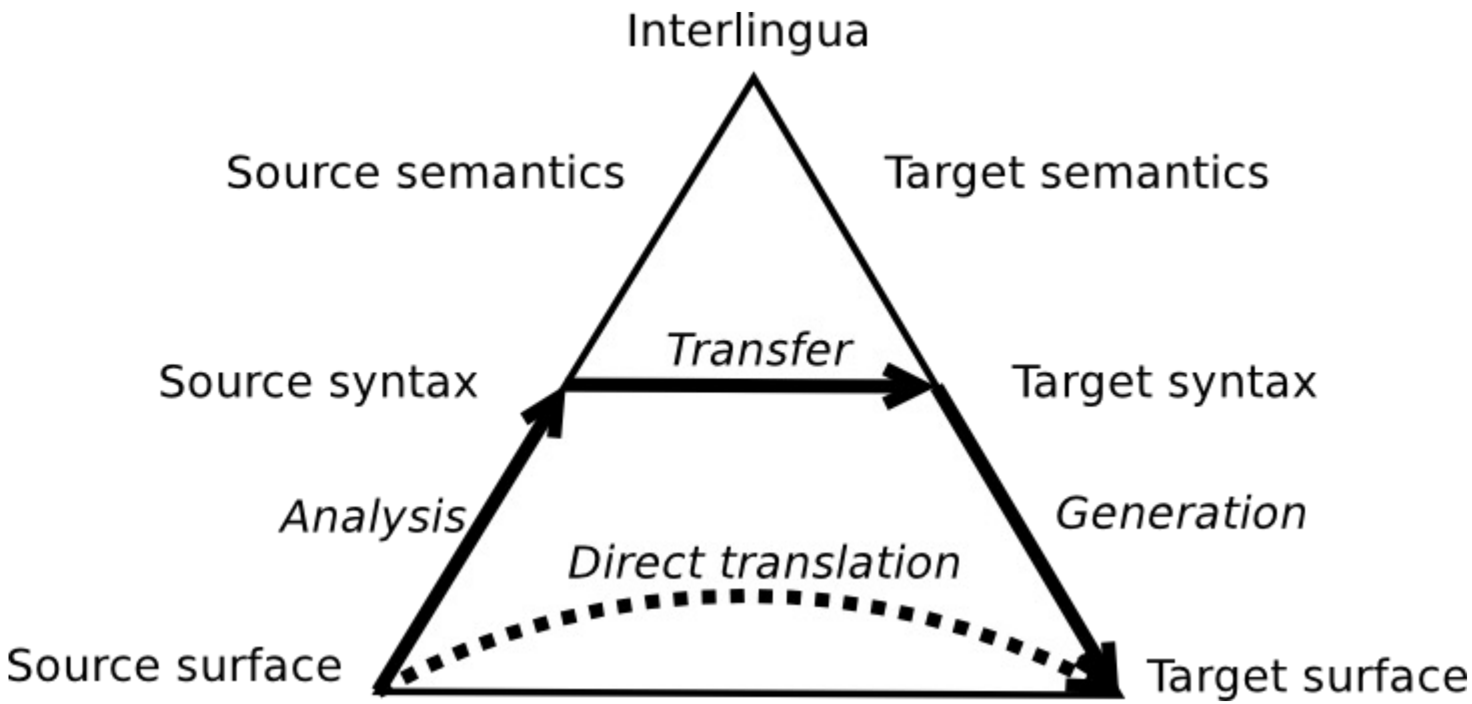


COMET Estimator

https://machinetranslate.org/metrics
https://github.com/mjpost/sacreBLEU

The Vauquois Triangle

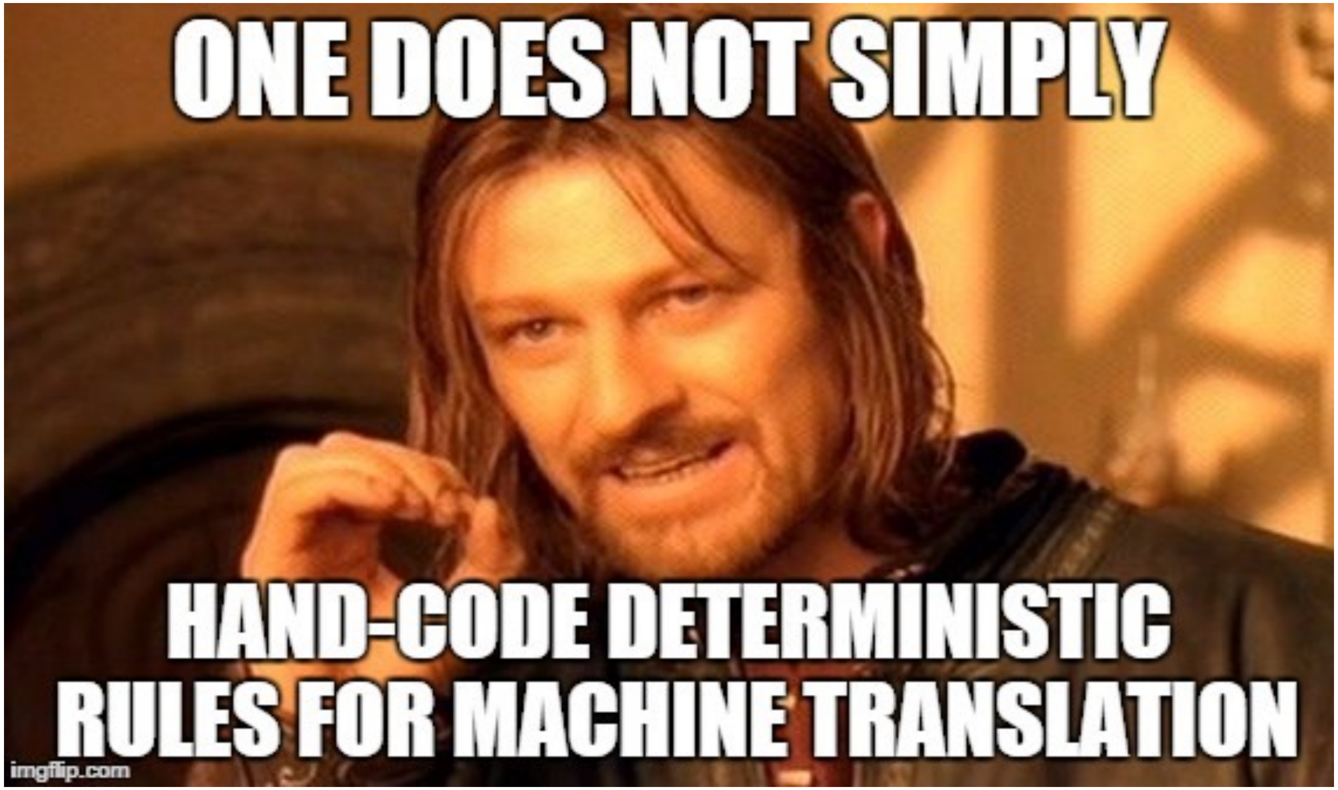
One important way to distinguish systems:



http://mttalks.ufal.ms.mff.cuni.cz/images/f/f1/Pyramid.png

Early Efforts

Early MT researchers developed a set of bilingual dictionary rules to map from one language to another.



Interlingua

A conceptual space common to all languages

Advantage: Can develop a general MT system

- Direct translation implies a system them is trained on and works for a specific pair of languages
- With interlingua, adding a new language only requires translating it into the interlingua

Disadvantage: What should an interlingua look like?

- It might be difficult to work with such an expressive formalism.

Statistical Machine Translation

Let’s look at a popular direct-transfer approach to statistical machine translation: the **noisy channel model**.

$$\begin{array}{ccc} \text{English} & & \text{Russian} \\ P(E) & \xrightarrow[P(F|E)]{} & \end{array}$$

When I look at an article in Russian, I say:
‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’

Warren Weaver, 1955

Which Direction

Suppose we are translating from Russian to English. Which of the following is correct?

$$E^* = \mathrm{argmax}_{\mathrm{E}}\, P(F)P(E|F)$$

$$E^* = \mathrm{argmax}_{\mathrm{E}}\, P(E)P(F|E)$$

Language Modelling, Again

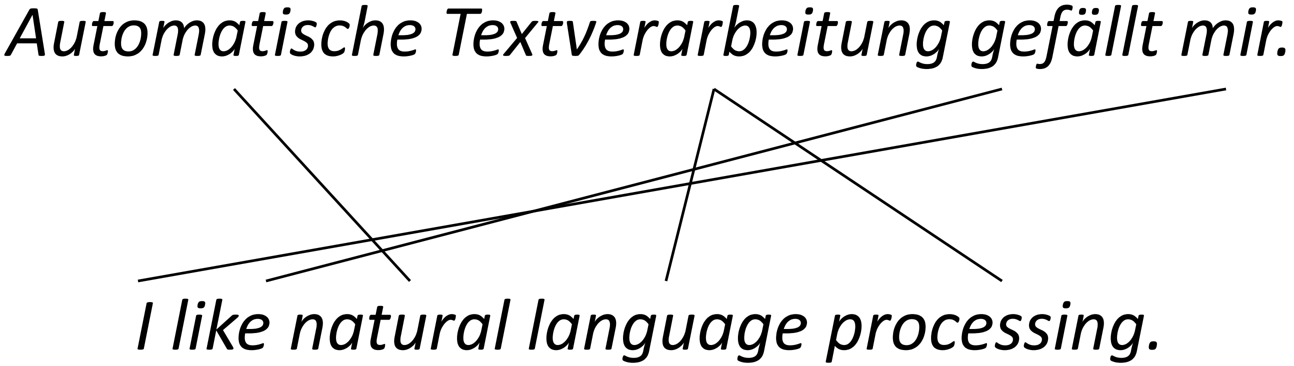
Noisy channel model:

$$E^* = \text{argmax}_{\mathrm{E}}\, \underset{\substack{\big\uparrow \\ \text{Language model}}}{P(E)} P(F|E)$$

What about $P(F|E)$? That is the **translation model**.

Translation as Word Alignment

Train a model of $P(F|E) = P(\text{Source}|\text{Target})$, as a **word alignment** model.



Automatische Textverarbeitung gefällt mir.

I like natural language processing.

First, need a source of data to train such a model…

Parallel Corpus

Contains the “same” text in two or more languages

- e.g., **Canadian Hansard** – parliamentary debates in English and French

E: Canada should therefore drop any reference to any system other than the metric system in ads, on signs, and on packaging. The petitioners are also calling for containers to be standardized to the metric system in units of 100 grams or 100 millilitres.

F: Le Canada devrait donc abandonner toute référence à un système autre que le système métrique dans la publicité, l'affichage et sur les contenants. Les pétitionnaires demandent aussi l'uniformisation des contenants au système métrique par tranche de 100 grammes ou de 100 millilitres.

Ms. Ève Péclet, 41st Parliament, #232

Sentence Alignment

Make use of various tricks to get sentence alignment

- Sentence lengths (Gale and Church, 1993)
- Cognate words (if languages use similar orthographies)
- Longest common subsequence of characters

Define a similarity function between sentences using these factors.

Search for an optimal alignment using a dynamic programming algorithm (e.g., **edit distance** variant such as **dynamic time warping**)

Word Alignment

Even after sentence alignment, we don’t have words that are aligned.

Factors to consider for word alignment:

- Plausibility of translation
- Many-to-many (or none) mapping
- Regularities in rearranging word orders

Unsupervised Word Alignment

Let’s play a language decoding game:

Swahili	English
Atacheza	*He/she will play*
Mlifahamu	*Y’all understood*
Mnapika	*Y'all cook*
Nilicheza	*I played*
Ninapika	*I cook*
Nitapika	*I will cook*
Tulifahamu	*We understood*
Unacheza	*You play*
Utapika	*You will cook*
Wanafahamu	*They understand*
Watapika	*They will cook*
Walicheza	*They played*

What are the Swahili morphemes for *play, understand, cook, I, you, he/she, we, y’all, they, PAST, PRES, FUT*?

(2014 North American Computational Linguistics Olympiad)

Unsupervised Word Alignment

Let’s play a language decoding game:

Swahili	English
Atacheza	*He/she will play*
Mlifahamu	*Y’all understood*
Mnapika	*Y'all cook*
Nilicheza	*I played*
Ninapika	*I cook*
Nitapika	*I will cook*
Tulifahamu	*We understood*
Unacheza	*You play*
Utapika	*You will cook*
Wanafahamu	*They understand*
Watapika	*They will cook*
Walicheza	*They played*

What are the Swahili morphemes for *play, understand, cook, I, you, he/she, we, y’all, they, PAST, PRES, FUT*?

(2014 North American Computational Linguistics Olympiad)

IBM Model 1

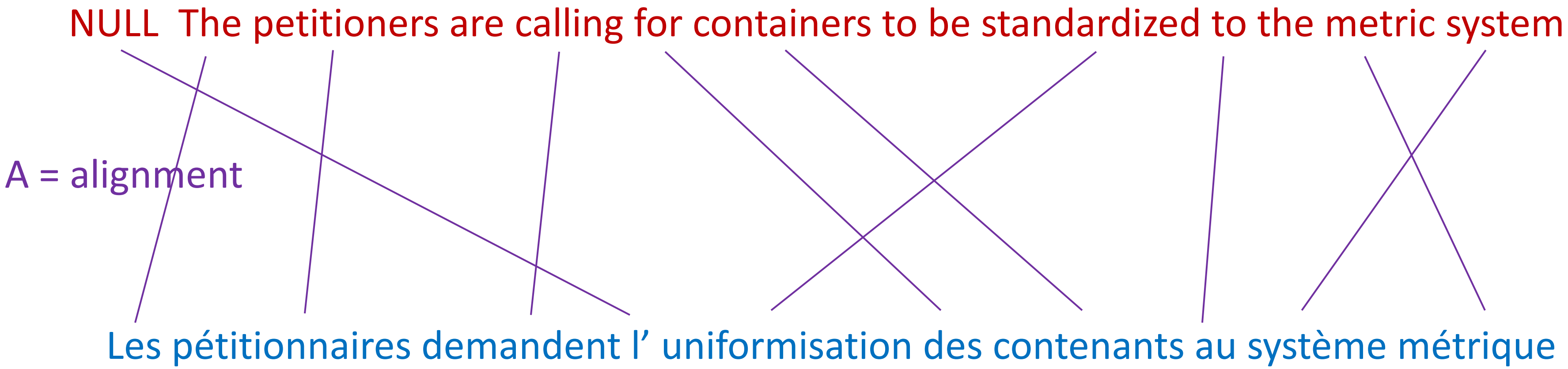
IBM developed a series of five influential models that make increasingly powerful assumptions.

Model 1 is the most basic:

- Each source word is aligned to **zero or one** target word
- Don’t try to model different **distortions** of word order (e.g., completely flipping word order vs. just swapping the orders of one or two words)
- Don’t try to model likelihood of **fertility** (some phrases, e.g., *take a walk*, might be translated as one unit)

Word Alignment

E = target sentence



NULL The petitioners are calling for containers to be standardized to the metric system

A = alignment

Les pétitionnaires demandent l’ uniformisation des contenants au système métrique

F = source sentence

- NULL node allows words in F to align to nothing in E.
- Since each source word is aligned to **zero or one** target word, |A| = |F|.
- Can represent A as indices: {1, 2, 4, 0, 9, 5, 6, 10, 13, 12}

Word Alignment Probabilities

$$P(F|E) = \sum_{A} P(F, A|E) = \sum_{A} P(F|E, A) \times P(A|E)$$

Probability of source sentence, given the target sentence, and knowing which words are aligned with which.

Probability of the alignment, given the target sentence.

$P(A|E)$

IBM Model 1 makes a very strong simplifying assumption:

- Uniform probability of translation length (i.e., length of A)
- Uniform probability for each possible alignment

$$P(A|E) \propto C$$

or

$$P(A|E) = \frac{\epsilon}{(I+1)^J}$$

, where I is the number of target words, J is the number of source words, ϵ is there to make sure things normalize across different possible values of J.

Why the + 1?

$\boldsymbol{P(F|E,A)}$

Decompose this into individual word alignments

$$P(F|E,A) = \prod_{j=1}^{J} P(f_j|e_{a_j})$$

How do we learn $P(f_j|e_{a_j})$?

- If we had observed word alignments in the training corpus, we could simply do MLE:

$$P(f|e) = \frac{\text{Count}(f,e)}{\text{Count}(e)}$$

- We don't, so it's time for …?

Expectation-Maximization

1. Initialize the parameters $P(f|e)$ randomly
2. Iterate for a while:
   - **E-step**: Given the current parameters, compute the expected value of $\mathrm{Count}(f,e)$ over the training data
   - **M-step**: Given the current $\mathrm{Count}(f,e)$, compute the new MLE $\theta_k = P(f|e)$

Probability of Alignments

To get the expected counts, what we really need is the probability of an alignment: $P(A|E,F)$

$$P(A|E,F) = \frac{P(A,E,F)}{P(E)P(F|E)} = \frac{P(F,A|E)}{P(F|E)} = \frac{P(F,A|E)}{\sum_A P(F,A|E)}$$

Since $P(F,A|E) = P(F|E,A)\times P(A|E)$, and $P(A|E)$ is the same for all alignments, we get:

$$P(A|E,F) = \frac{P(F|E,A)}{\sum_A P(F|E,A)}$$

Recall that $P(F|E,A) = \prod_{j=1}^{J} t(f_j|e_{a_j})$.

Thus, we're set, given some initial model of $t(f|e)$.

Example

Let’s do one round of EM training for the following mini-corpus:

red house	*the house*
maison rouge	*la maison*

Initialize the model $t(f|e)$ uniformly:

$t(maison\|red) = \frac{1}{3}$	$t(rouge\|red) = \frac{1}{3}$	$t(la\|red) = \frac{1}{3}$
$t(maison\|house) = \frac{1}{3}$	$t(rouge\|house) = \frac{1}{3}$	$t(la\|house) = \frac{1}{3}$
$t(maison\|the) = \frac{1}{3}$	$t(rouge\|the) = \frac{1}{3}$	$t(la\|the) = \frac{1}{3}$

Details, Details

In practice, don’t initialize $t(f|e)$ uniformly:

- Given reasonable sizes of lexicon, too many parameters = too much memory and computation!
- Rather, restrict it to word pairs e’, f’, where e’ and f’ occur is some aligned sentence pair in the training set.

When sentence lengths are high, need to efficiently compute probabilities of all possible alignments.

- Can adapt our algorithm to implicitly sum over all alignments

References

Boroditsky and Gaby. 2010. Remembrances of Times East: Absolute Spatial Representation of Time in an Australian Aboriginal Community. *Psychological Sciences.*

Matthewson. 2006. Presuppositions and cross-linguistic variation.