Statistical Natural Language Processing

ELEC-E5550 P, Spring 2024

12 Mar 2024

Statistical Machine Translation & Machine Translation Evaluation

Jaakko Väyrynen

Lecturer: Lead Data Scientist, Utopia Analytics

D.Sc. (Tech.)

Slides: Jaakko Väyrynen, Philipp Koehn, Mathias Creutz, Krista Lagus, Timo Honkela etc.

About the lecturer

- -2012, Aalto University
- 2013–2015, Joint Research Centre, European Commission
- 2016-, Utopia Analytics

JRC
EUROPEAN COMMISSION
Joint Research Center -
Institute for the Protection and Security of the Citizen - JRC (EU)

UTOPIA
ANALYTICS

EMM
Europe Media Monitor

1. Statistical Machine Translation

Lecture based on:

- Chapter 13.2-13.4 in Manning & Schütze
- Chapter 21 in Jurafsky & Martin: Speech and Language Processing (An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition) (**Ch. 11 in 3rd edition**)
- Koehn: "Statistical Machine Translation", http://www.statmt.org/book/

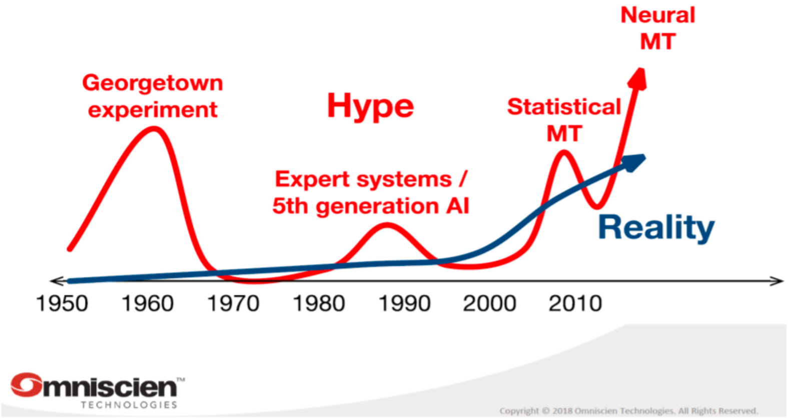
See also:

- CLT310 (2016) slides from University of Helsinki
- CS 224N / Ling 284 slides from Stanford University

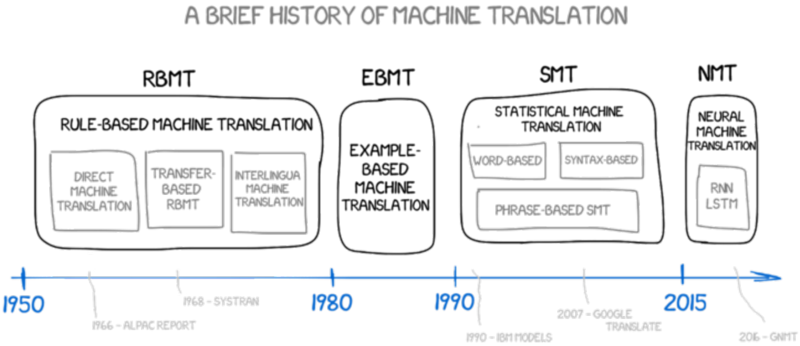
1.1 MT Applications

- Connect people/companies (The European Single Market).
  - Patent translation (European Patent Office)
  - Communication (Google, FB, NSA, Customs, Military, ...)
  - Text, speech, augmented reality translation (Google, FB, ...)
- Multilingual organizations (UN, EU, AU, Finland, India, ...)
  - In 2020, DGT translated 2.3M pages, eTranslation delivered 70M pages
  - Technical documents (Microsoft), user manuals
- Different goals, different standards
  - Understanding (gist translation)
  - Dissemination (publishable quality, authored by humans)

1.2 Historical context



Thanks to Philipp Koehn for this nice summary


A BRIEF HISTORY OF MACHINE TRANSLATION
RBMT
RULE-BASED MACHINE TRANSLATION
DIRECT MACHINE TRANSLATION
TRANSFER-BASED RBMT
INTERLINGUA MACHINE TRANSLATION
EBMT
EXAMPLE-BASED MACHINE TRANSLATION
SMT
STATISTICAL MACHINE TRANSLATION
WORD-BASED
SYNTAX-BASED
PHRASE-BASED SMT
NMT
NEURAL MACHINE TRANSLATION
RNN LSTM
1950
1966 - ALPAC REPORT
1968 - SYSTRAN
1980
1990
1990 - IBM MODELS
2007 - GOOGLE TRANSLATE
2015
2016 - GNMT

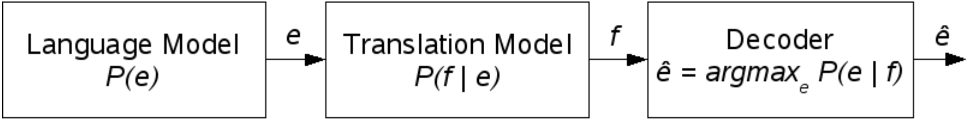
by Ilya Pestov

1.3 Statistical approach

- In 1949, Warren Weaver suggested applying statistical and cryptanalytical techniques from the field of communication theory to the problem of using computers to translate text from one natural language to another.

- However, computers at that time were far too inefficient, and the availability of language data (text) in digital form was very limited.

- The idea of the **noisy channel** model: The language model generates an English sentence e. The translation model transmits e "noisily" as the foreign sentence f. The decoder finds the English sentence $\hat{e}$ which is most likely to have given rise to f.



Language Model
$P(e)$

e

Translation Model
$P(f \mid e)$

f

Decoder
$\hat{e} = argmax_e \; P(e \mid f)$

$\hat{e}$

- In the examples, we usually translate from a foreign language f into English e. (The Americans want to figure out what is written or spoken in Russian, Chinese, Arabic...) In the first publications in the field (the so-called IBM model), f referred to French, but to think of f as any foreign language is more general.

- Using **Bayes' rule**, or the noisy channel metaphor, we obtain:

$$P(e|f) = \frac{P(e)P(f|e)}{P(f)}. \tag{1}$$

  Since the denominator is independent of e, finding $\hat{e}$ is the same as finding e so as to make $P(e)P(f|e)$ as large as possible:

$$\hat{e} = \arg\max_{e} P(e)P(f|e)/P(f) = \arg\max_{e} P(e)P(f|e). \tag{2}$$

- This can be interpreted as maximizing the **fluency** of the English sentence $P(e)$ as well as the **faithfulness** of the translation between English and the foreign language $P(f|e)$:

$$\text{best translation } \hat{e} = \arg\max_{e} \text{ fluency}(e) \cdot \text{faithfulness}(f|e). \tag{3}$$

- The language model probability (or measure of fluency) $P(e)$ is typically decomposed into a product of n-gram probabilities (see Lecture on statistical language models).

- The translation model (or measure of faithfulness) $P(f|e)$ is typically decomposed into a product of word-to-word, or phrase-to-phrase, translation probabilities. For instance, $P(\textit{Angleterre}|\textit{England})$ should be high, whereas $P(\textit{Finlande}|\textit{England})$ should be low.

- It maybe strange to think of a human translator that would divide the task into first (1) enumerating a large number of fluent English sentences, and then (2) choosing one, where the words translated into French would match the French input sentence well.

- The IBM model also comprises **fertility** and **distortion** probabilities. We will get back to them shortly.

- The success of statistical machine translation depends heavily on the quality of the **text/word alignment** that is produced.

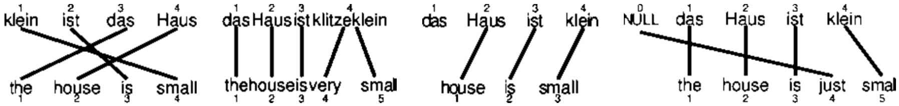
1.4 Word-based models

- **Lexical translation probabilities** $P(f = Haus|e = house)$ as maximum likelihood estimates from a parallel corpus.

- **Alignment model** needs to handle word reordering, multiple alignments per word, dropping words, inserting words.

- Basic idea for training: Expectation Maximization (EM) alternating between finding most likely alignments for the parallel corpus and estimating lexical translation probabilies from the alignments.

- IBM models are still relevant for phrase-based models for creating a starting point for word aligment.


klein 1 ist 2 das 3 Haus 4
the 1 house 2 is 3 small 4
das 1 Haus 2 ist 3 klitzeklein 4
the 1 house 2 is 3 very 4 smal 5
das 1 Haus 2 ist 3 klein 4
house 1 is 2 small 3
NULL 0 das 1 Haus 2 ist 3 klein 4
the 1 house 2 is 3 just 4 smal 5

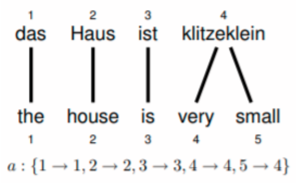
IBM alignment models

- Foreign sentence $f = (f_1, \ldots, f_{l_f})$ of length l_f
- English sentence $e = (e_1, \ldots, e_{l_e})$ of length l_e
- Each output word is linked only to one input word with alignment $a : j \to i$ of each English word e_j to a foreign word f_i
- Handles many-to-one alignments, but not one-to-many alignments.
- Gradually increase model complexity, use output from last step as input to bootstrap model training.
- Lexical translation probabilities $t(f|e)$

IBM (alignment) model 1

$$p(f,a|e,l_e) = \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(f_j|e_{a(j)})$$

$$p(f|e,l_e) = \sum_a p(f,a|e,l_e)$$


1 das
2 Haus
3 ist
4 klitzeklein
the 1
house 2
is 3
very 4
small 5
$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$

- Parameter ϵ is a normalization constant

IBM (alignment) models 2-5

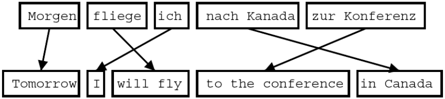
- IBM model 2: adds absolute alignment/reordering model $a(i|j, l_e, l_f)$
- HMM alignment model: add condition on alignment of the previous word: $p(a(j)|a(j-1), l_e)$
- IBM model 3 adds fertility model: how many output words ϕ each foreign word usually translates to: $n(\phi|f)$, distortion instead of absolute alignment, NULL insertion parameter
- IBM model 4: adds relative reordering model: each word is dependent on the previously aligned word and on the word classes of the surrounding words
- IBM model 5: fixes decifiency by reformulating IBM Model 4 by enhancing the alignment model with more training parameters
- IBM model 1 has a global maximum: other models build on previously trained lower-level models; IBM model 3 cannot do exact estimation: sampling over high-probability alignments

1.5 Phrase-based translation model

$$\arg\max_e p(e|f) = \arg\max_e p(f|e)p(e) = \arg\max_e \phi(f|e)p_{LM}(e)\omega^{\texttt{length(e)}} \tag{4}$$

- Components: phrase translation model $\phi(f|e)$, reordering model d, language model $p_{LM}(e)$, length bonus $\omega^{\texttt{length(e)}}$
- Sentence f is decomposed into I phrases $\bar{f}_1^I = \bar{f}_1, \ldots, \bar{f}_I$
- Decomposition of $\phi(f|e)$

$$\phi(\bar{f}_1^I, \bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)d(a_i - b_{i-1}) \tag{5}$$



Log-linear translation model

- Log-linear translation model

$$p(e|f) = \frac{\exp\left[\sum_{m=1}^{M} \lambda_m h_m(f,e)\right]}{\sum_{e'} \exp\left[\sum_{m=1}^{M} \lambda_m h_m(f,e')\right]}, \tag{6}$$

  with weights λ_m and feature functions h_m.

- Optimize weights: maximize likelihood, or more typically some automatic translation quality measure (such as BLEU).

- Tune on a development set, which is often from a different domain than most of the training data.

- Possible feature functions (anything based on e, f, and/or a)
  - Direct and inverse translation scores: $\log P(f|e)$, $\log P(e|f)$.
  - Direct and inverse lexical scores for phrases: $lex(f|e)$, $lex(e|f)$.

- Additional language models: $\log P(e)$
- Word count: $wc(e) = \log |e|$; Phrase count: $pc(e) = \log |I|$
- Reordering models, e.g. $\log |start_i - end_{i-1} - 1|$
- Phrase pair frequency
- Sparse features: translation, word deletion/insertion, phrase length, count bins, domain features, soft-matching features (similar to tagging)
- Bilingual language model
- Reordering operations

Tools for the derivation of the log-linear model

- Viterbi assumption: take best alignment, do not sum over all
- Feature functions $h(f,a|e)$, e.g. language model $P_{LM}(e)$, distortion model $P_D(a)$, translation model $P_{TM}(f,a|e)$
- Assume independence of features so product of feature functions
- Add a weight for each feature, e.g. $P_{LM}(e)^{\lambda_{LM}}$
- Softmax normalizes scores into probabilities: $\theta(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$
- Work in log space: $\arg\max P = \arg\max \log P$, $\log(P^\lambda) = \lambda \log P$, $\log(P_1 P_2) = \log(P_1) + \log(P_2)$, $\log e^{f(x)} = e^{\log f(x)} = f(x)$

Group discussion

Discuss 5 minutes in groups of three or four. Consider different levels of language and different kinds of source-target pairs.

- What would be easy/hard to translate with MT?
- Have you seen failed/succesful usage or applications of MT?

Phrase-based SMT system

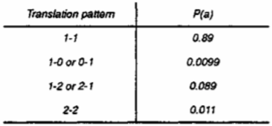
- Training data and data preprocessing
- Word aligment, phrase aligment
- Estimation of translation model scores
- Estimation of reordering model scores
- Estimation of language model scores
- Decoding algorithm and optimization of the model weights
- Translation, recasing, detokenization
- Evaluation, quality estimation
- Operational management

1.6 Training data

- Leverage existing translations.
- Parallel texts: Bible, UN/EU documents, subtitles, dictionaries, translation memories, user manuals.
- Quasi-parallel texts: Wikipedia pages, news articles, home pages.
- Sentence alignment: find parallel sentences from parallel texts
- More challenging problems: How to find translations for words and phrases? How to translate new sentences?
- Preprocessing: casing, tokenization, normalization

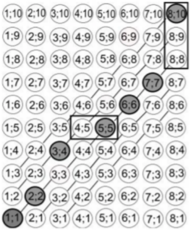
1.7 Sentence Alignment

- Simplifying assumptions: monotonic, break on paragraphs
- Sentence beads: $1 : n$, $n : 1$
- Gale&Church algorithm: model sentence lengths
- Other features: (automatic) dictionaries, cognates
- Dynamic programming (Similar to Viterbi)
- Find reliable alignments?



Translation pattern	*P(a)*
1-1	*0.89*
1-0 or 0-1	*0.0099*
1-2 or 2-1	*0.089*
2-2	*0.011*

Table 1: a priori probabilities of translation patterns. Source: [Gale and Church, 1991].



1.8 Word Alignment

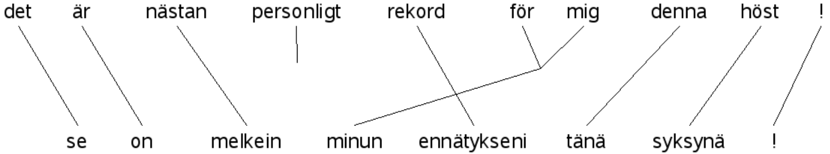
- In the alignment of entire sentences and sections, cross-alignments are not considered. If there were differences in the order in which the message was conveyed in the two languages, we created large enough beads that included multiple sentences on both sides. In this way, we didn't have to rearrange the order of the sentences in either language, while each bead still contained approximately the same thing in both languages.

- The sentence alignment was just a first step to facilitate a complete word-level alignment. In the word-level alignment, we do take into account the reordering (*distortion*) and *fertility* of the words.

- Distortion means that word order differs across languages.

- The fertility of a word in one source language with respect to another target language measures how many words in the target language the word in the source language is translated to on average.

- For instance,



Personligt was not aligned at all, and the two words *för mig* were aligned with one word *minun* (and the morpheme *-ni* if we analyze the words into parts). (Compare to tagging: no monotonic mapping)

- The basic approach in word level alignment: alternate between the two steps (after initialization):

  1. Generate a word level alignment using estimated translation probabilities
  2. Estimate translation probabilities for word pairs from the alignment.

  This is a form of Expectation-Maximization (EM) algorithm.

The bilingual dictionary will contain (finally) only word pairs that provide enough evidence, i.e., enough samples for the equivalent of those words.

- The translation probability of a sentence is then obtained as follows: Let f be a sentence in foreign language and e in English. Then the translation probability is

$$P(f|e) = \sum_{a} P(f,a|e) = \frac{1}{Z} \sum_{a_1=0}^{l_e} \cdots \sum_{a_{l_f}=0}^{l_e} \prod_{j=1}^{l_f} P(f_j|e_{a_j}), \qquad (7)$$

where l_e and l_f are the word counts in sentences e and f; $P(f_j|e_{a_j})$ is the probability of a word in the sentence in foreign language in position j being generated from a word in English in position a_j (0 stands for empty set). Z is a normalization factor.

Nested summations thus sum over all possible alternative alignments and the product over the words in the sentence f.

- The word-level translation probability can be constructed so as to take into account distortion and fertility probabilities (IBM models).

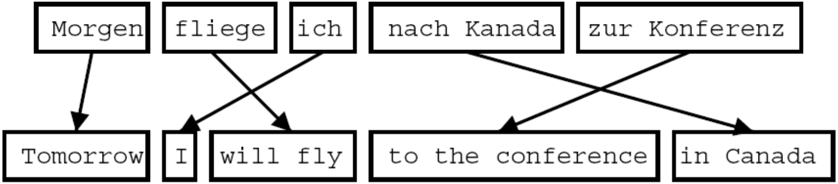
- The program GIZA++ implements IBM alignment models.

- The program fast_align is a reparameterization of IBM Model 2.

1.9 Phrase Alignment

- Translation problems with words as units:
  - "Cut-and-paste" translation (no syntax or semantics): it is probable that when words are "cut" from one context and "pasted" into another context mistakes occur, despite the language model.
  - The distortion (reordering) probability typically penalizes more, if several words have to be reordered. However, usually larger multi-word chunks (subphrases) need to be moved.

- Example:



- Phrase-to-phrase translation is an alternative to the IBM phrase-to-word model, and the phrase-models can be constructed starting from the IBM phrase-to-word models in both directions.

- Although we still rely on the "cut-and-paste" philosophy, we deal with larger chunks, so there are fewer "seams" between chunks combined in a new way. The word sequence within a phrase has been attested before in real texts, so it should be more or less correct. Phrases can also capture non-compositional word sequences, such as *it's anyone's guess* = *on mahdoton tietää*. In short, better use is made of the **local context**.

- The more data is available, the longer phrases can be learned. In translation, phrases typically consist of 1–3 words.

- Compare to ASR where sub-word units (morphs) are more suitable.

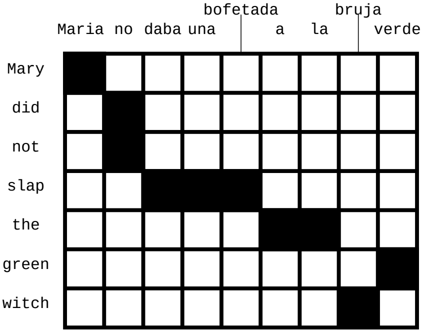
Phrase translation table

- Phrase translations for *den Vorschlag*

English	ϕ**(e|f)**	**English**	ϕ**(e|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	...	...

How to learn the phrase translation table?

- Start with the *word alignment*:



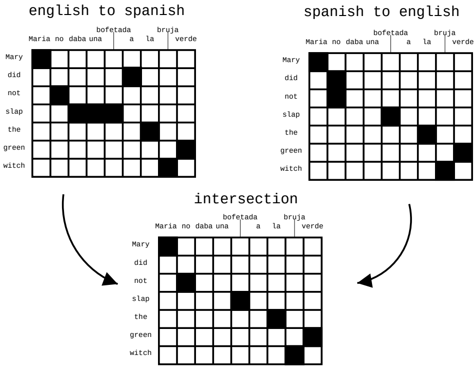
	Maria	no	daba	una	bofetada	a	la	bruja	verde
Mary	■								
did		■							
not		■							
slap			■	■	■				
the						■	■		
green									■
witch								■	

- Collect all phrase pairs that are **consistent** with the word alignment

Word alignment with IBM models

- IBM Models create a *many-to-one* mapping
  - words are aligned using an **alignment function**
  - a function may return the same value for different input (one-to-many mapping)
  - a function can not return multiple values for one input (*no many-to-one* mapping)
- But we need *many-to-many* mappings

Symmetrizing word alignments



english to spanish

	Maria	no	daba	una	bofetada	a	la	bruja	verde
Mary	■								
did						■			
not		■							
slap			■	■	■				
the							■		
green									■
witch								■	

spanish to english

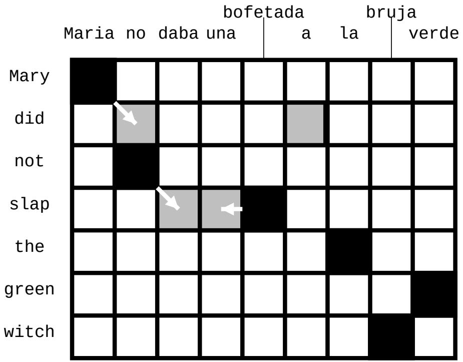
	Maria	no	daba	una	bofetada	a	la	bruja	verde
Mary	■								
did		■							
not		■							
slap					■				
the							■		
green									■
witch								■	

intersection

	Maria	no	daba	una	bofetada	a	la	bruja	verde
Mary	■								
did									
not		■							
slap					■				
the							■		
green									■
witch								■	

- *Intersection* of GIZA++ bidirectional alignments

Symmetrizing word alignments



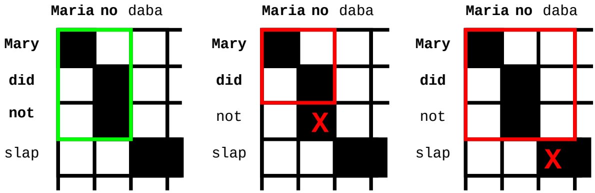
- *Grow* additional alignment points [Och and Ney, CompLing2003]

Growing heuristic

```
GROW-DIAG-FINAL(e2f,f2e):
  neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned and f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )
FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
           ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```

Consistent with word alignment



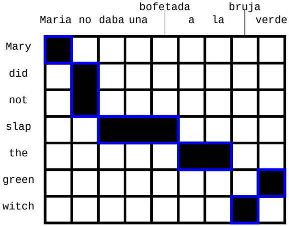
consistent inconsistent inconsistent

- **Consistent with the word alignment** :=

  phrase alignment has to *contain all alignment points* for all covered words

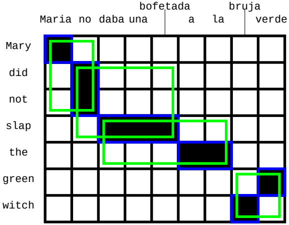
$$\begin{aligned} (\overline{e}, \overline{f}) \in BP \Leftrightarrow \quad & \forall e_i \in \overline{e} : (e_i, f_j) \in A \rightarrow f_j \in \overline{f} \\ \text{AND} \quad & \forall f_j \in \overline{f} : (e_i, f_j) \in A \rightarrow e_i \in \overline{e} \end{aligned}$$

Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

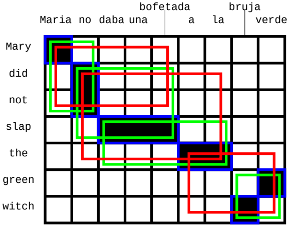
Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

Word alignment induced phrases



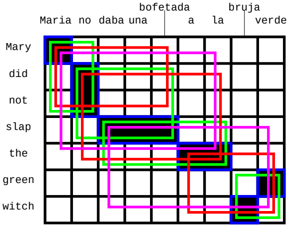
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),

(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),

(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

Word alignment induced phrases

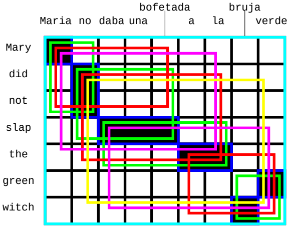


(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch), **(Maria no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),**

(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch)

Word alignment induced phrases (5)



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch), (Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch), (no daba una bofetada a la bruja verde, did not slap the green witch), (Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Probability distribution of phrase pairs

- We need a **probability distribution** $\phi(\bar{f}|\bar{e})$ over the collected phrase pairs

⇒ Possible *choices*

  - *relative frequency* of collected phrases: $\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f},\bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f},\bar{e})}$
  - or, conversely $\phi(\bar{e}|\bar{f})$
  - use *lexical translation probabilities*

Reordering

- *Monotone* translation
  - do not allow any reordering

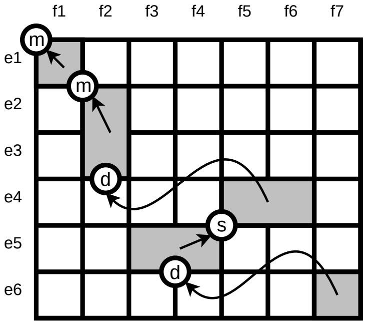
  → worse translations

- *Limiting* reordering (to movement over max. number of words) helps

- *Distance-based* reordering cost
  - moving a foreign phrase over n words: cost ω^n

- *Lexicalized* reordering model

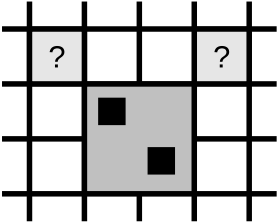
Lexicalized reordering models



[from Koehn et al., 2005, IWSLT]

- Three **orientation** types: **monotone**, **swap**, **discontinuous**
- Probability $p(swap|e, f)$ depends on foreign (and English) *phrase* involved

Learning lexicalized reordering models



[from Koehn et al., 2005, IWSLT]

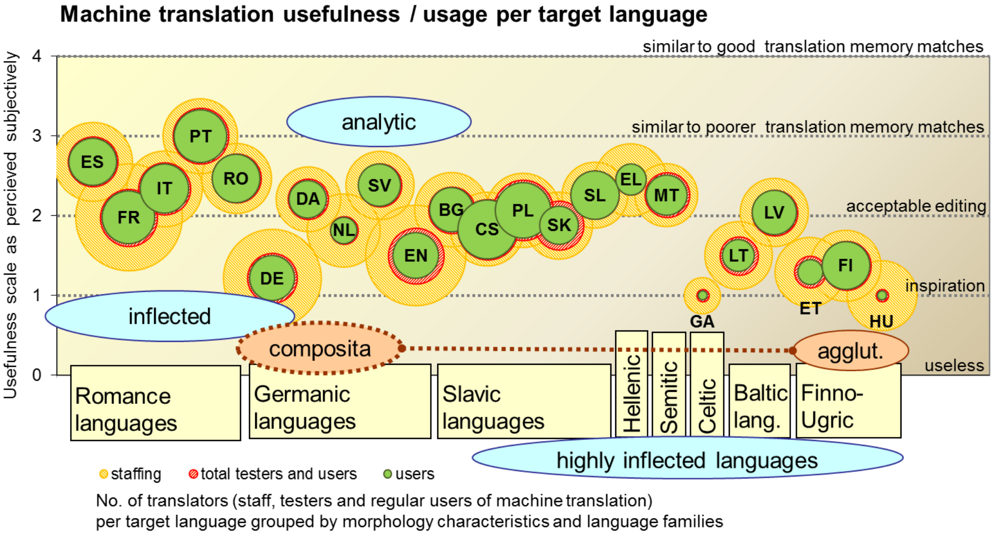
- Orientation type is *learned during phrase extractions*
- *Alignment point* to the *top left* (monotone) or *top right* (swap)?
- For more, see [Tillmann, 2003] or [Koehn et al., 2005]

1.10 Translation methods

There are different model families and decoders for phrase-based statistical translation. Typically models require either word or phrase alignment as input.

- Phrase-based methods with e.g. beam-search decoder (e.g. Moses) performs search similar to ASR.
- (Weighted) finite state transducer (FST) based translation models implement a bilingual language model learned from word alignment.
- Extended word-level representations, e.g., hierarchical phrase-based models and factored translation models with words augmented with POS tags, lemmas, etc.
- Syntax-based translation models, which take syntax parse trees as input.
- Feature-based models, where translation is performed between features, not words or phrases. E.g., discriminative training or conditional exponential family distribution over feature vectors

MT@EC: annual evaluation



No. of translators (staff, testers and regular users of machine translation) per target language grouped by morphology characteristics and language families

1.11 Example: A Phrase-Based Translation System

The open-source statistical machine translation system **Moses** has been used (`http://www.statmt.org/moses/`). Moses was trained on text data in which the words had been split into morphs by Morfessor. The training set contained circa 900,000 sentences, or 20 million words (including punctuation). The borders of the phrases used are marked using a vertical bar |. Morph boundaries are not marked. Work was done in 2007.

- **Source 1:** det är nästan personligt rekord för mig denna höst !
- **Translation 1:** se on melkein | henkilökohtainen | ennätys | minulle | tämän | vuoden syksyllä | !
- **Reference 1:** se on melkein minun ennätykseni tänä syksynä !
- **Google 1:** Se on melkein henkilökohtainen tietue minulle tänä syksynä!
- **Google 1 (2024):** Se on lähes henkilökohtainen ennätys minulle tänä syksynä!

- **Source 2:** det är fullständigt utan proportioner och hjälper inte till i fredsprocessen på något sätt .

- **Translation 2:** se on täysin | ilman | suhteelli|suudentaju | ja auttaa | rauhanprosessissa | ei | millään | tavalla .

- **Reference 2:** tämä on täysin suhteetonta eikä se edistä rauhanprosessia millään tavoin .

- **Google 2:** Se on täydellinen ilman osuuksia ja ei millään tavoin auta rauhanprosessissa.

- **Google 2 (2024):** Se on täysin suhteeton eikä edistä rauhanprosessia millään tavalla.

- **Source 3:** jag går in på denna punkt därför att den är mycket intressant .

- **Translation 3:** en | käsittele | tätä kohtaa | , koska se | on hyvin mielenkiintoinen .

- **Reference 3:** puutun tähän kohtaan , koska se on hyvin mielenkiintoinen .

- **Google 3:** Menen tähän, koska se on erittäin mielenkiintoinen.

- **Google 3 (2024):** Menen tähän kohtaan, koska se on erittäin mielenkiintoinen.

- **Source 4:** vad konkurrensen anbelangar så är marknaden avgörande för utvecklingen i kusthamnarna .

- **Translation 4:** mitä | tulee | niin | kilpailu|t | markkinat ovat | ratkaisevan tärkeitä | kehitykse|n | merisatamiin | .

- **Reference 4:** mitä kilpailuun tulee , markkinat vaikuttavat ratkaisevasti merisatamien kehitykseen .

- **Google 4:** Kilpailun osalta markkinat ovat ratkaisevia rannikkosatamien kehittämisessä.

- **Google 4 (2024):** Kilpailun kannalta markkinat ovat ratkaisevia rannikkosatamien kehitykselle.

- **Source 5:** denna prioritering är emellertid skadlig för miljön och innebär ett socialt slöseri .
- **Translation 5:** tämän | ensisijaisena tavoitteena on | kuitenkin | vahingoittaa | ympäristöä ja aiheuttaa | yhteiskunnallista | tuhlausta .
- **Reference 5:** tällainen suosiminen on kuitenkin ekologisesti vahingollista ja sosiaalisesti epäonnistunutta .
- **Google 5:** Tämä prioriteetti on kuitenkin haitallista ympäristölle ja siihen liittyy yhteiskunnallista jätettä.
- **Google 5 (2024):** Tämä priorisointi on kuitenkin haitallista ympäristölle ja aiheuttaa sosiaalista tuhlausta.

Some Weaknesses of the System

- No modeling of syntax or semantics.
- Sensitivity to training data: small changes in training data (or test data) selection cause significant changes to resulting rates. The correspondence between training and testing data should be high for this kind of word level translation model to work well.
- Efficiency: computationally heavy for long sentences.
- Data sparseness (inadequency). For rare words the estimates are bad (read: quite random).
- In morphologically rich languages the data sparseness is emphasized unless the words are segmented or otherwise taken into account.
- If the language model is local (e.g., an n-gram model), it won't help even if the translation model could provide translations utilizating long distance dependencies. The assumptions made by different models should be consistent.

1.12 Rise of the SMT?

- Faster computers, more memory, networks
- Free resources (see OPUS, `http://opus.lingfil.uu.se/`)
- Open source software (eg. Moses, `http://www.statmt.org/moses/`)
- Generalization (IBM models with words, language independence)
- Instantiation (phrase based translation)
- Increased efficiency and complexity (heuristics, new algorithms)
- Incorporating knowledge (patterns which can be leveraged)
- Learn complex dependencies (neural machine translation)

SMT pros and cons

- Reuses translations of word groups
- Does not generalize from an observation to “similar” cases
- Can handle very large vocabularies, but no complex linguistic construction
- Good in adequacy, not so good in fluency
- Modular: models focusing on certain aspects can be improved separately
- Incorporating new data via incremental training is complicated

2. Machine Translation Evaluation

Part II

Machine Translation Evaluation

Outline

Human evaluation

Automatic evaluation

Meta-evaluation

How to evaluate MT systems?

Final evaluation should depend on the intended application

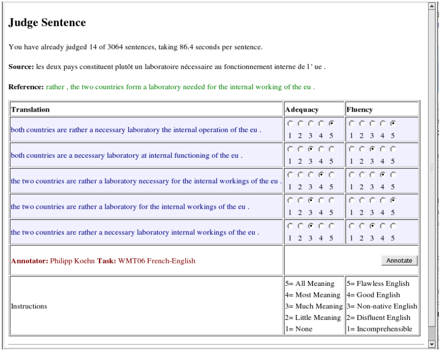
- Understanding text as it is; skimming/gisting → Human evaluation
- Aid for human translations → Decrease in translation time
- Multilingual information retrieval → IR evaluation

Human evaluation: Direct assessment

Given translation output and source and/or reference translation, how good the translation is?

Adequacy: Does the output convey the same meaning?

Fluency: Is the output good and fluent language?



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

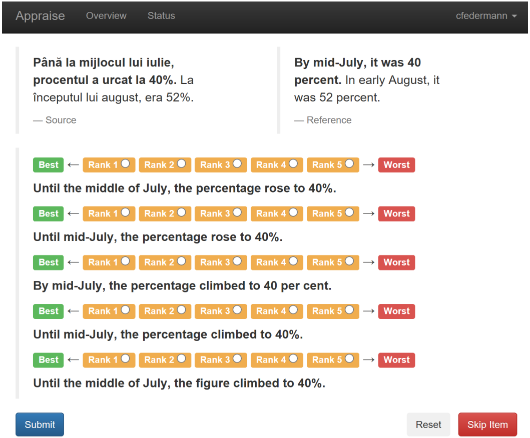
Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	○ ○ ○ ○ ● 1 2 3 4 5	○ ○ ○ ○ ● 1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	○ ○ ● ○ ○ 1 2 3 4 5	○ ○ ● ○ ○ 1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	○ ○ ○ ● ○ 1 2 3 4 5	○ ○ ○ ● ○ 1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	○ ○ ● ○ ○ 1 2 3 4 5	○ ○ ○ ○ ● 1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	○ ○ ● ○ ○ 1 2 3 4 5	○ ○ ● ○ ○ 1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Human evaluation: Ranking

Given *N* translation output and source, order them from best to worst.



Până la mijlocul lui iulie, procentul a urcat la 40%. La începutul lui august, era 52%.

— Source

By mid-July, it was 40 percent. In early August, it was 52 percent.

— Reference

Until the middle of July, the percentage rose to 40%.

Until mid-July, the percentage rose to 40%.

By mid-July, the percentage climbed to 40 per cent.

Until mid-July, the percentage climbed to 40%.

Until the middle of July, the figure climbed to 40%.

Human evaluation: Agreement

Evaluators disagree in their assessments.

Inter-evaluator agreement can be measured with Kappa coefficient:

$$K = \frac{p(A) - p(E)}{1 - p(E)}$$

$p(A)$ = proportion of agreement

$p(E)$ = agreement by chance

Ranking provides more consistent results than direct assessment.

Evaluating translator efficiency gain

How does the average translation time per sentence change?

- From scratch
- Using only translation memory
- Between different MT systems

Challenges:

- Translators have different experience and ways of working
- High variability between translation segments
- Easiest cases often solved by translation memories
- How to present the translation in the UI

Needs lots of data or complicated setup and advanced analysis (e.g. mixed-effect regression models).

Why automatic evaluation?

Manual evaluation is expensive

MT researchers rarely have the resources.

Annual competitions (WMT shared tasks) help somewhat.

Manual evaluation is slow

Cannot be used during development.

Especially not for optimization of model parameters and hyperparameters.

Challenges in automatic evaluation

Why is MT evaluation more difficult than ASR evaluation?

Why can we not use word error rate (WER)?

Challenges in automatic evaluation

Multiple correct answers: Ideally there should be several reference translations made by different persons.

Graded correctness: Word choices, grammatical correctness, emphasis (“koira jahtasi kissaa” vs. “kissaa koira jahtasi”), style (“kick the bucket” vs. “die”), ...

Usefulness depends on intended use.

- Translator’s tool: Long segments that require no changes
- Skimming: Meaning should be correct; fluent enough for easy understanding
- Information retrieval: Terminology important; fluency and grammatical correctness do not matter

Global edit distance metrics

Word and letter error rates do not account for possible variations in word order.

Edit distance with moves is an NP-hard problem.

Solutions:

- TER: Shift operation + greedy search (**snower2006ter**)
- SPEDE: Limited-distance word swapping (**wang-manning:2012:wmt**)

Local metrics

Concentrate on small parts of the full text at a time.

Similarity to IR metrics:

- **Precision:** Every item should be found in the reference.
- **Recall:** Anything in the reference should not be left out.

Observing individual words is not adequate (word order!)

Local metrics: BLEU

BLEU (“Bilingual Evaluation Understudy”) (**papineni2002bleu**) was one of the first metrics to report high correlation with human judgments of quality.

$$\text{BLEU} = \min\left(1, \frac{\texttt{output-length}}{\texttt{reference-length}}\right)\left(\prod_{i=1}^{4}\texttt{precision}_i\right)^{\frac{1}{4}}$$

Typically calculated over entire corpus (system-level evaluation).

Example in the excercise session.

Local metrics: Problems in BLEU

Does not work for languages with no word boundaries.

Single word or n-gram is scored 0 or 1.

- Inflections: “translation” vs. “translations”
- Derivations: “[he] made translations” vs. “[he] translated”
- Compounds: “Arbeits Geberverband” vs. “Arbeitgeberverband” *(employers’ organization)*

Poor measure of adequacy for morphologically rich languages.

Beyond word-based metrics

Preprocessing (stemming, morphological segmentation)

- METEOR (**banerjee2005meteor**; **denkowski-lavie:2011:wmt**)
- AMBER (**chen-kuhn:2011:wmt**)

Characted-based measures

- char-BLEU (**denoual2005bleu**)
- Weighted character F-score (chrF3) (**popovic:2015:WMT**)

Combine with word similarity calculation

- Alignment based on character similarity (**homola-kubovn-pecina:2009:wmt**)
- Tolerant BLEU (**libovicky-pecina:2014:W14-33**)
- LeBLEU (**virpioja2015wmt**)

Semantic similarity using contextual embeddings

- BERTscore (**zhang2019bertscore**)
- COMET (**rei-etal-2020-comet**)

How to evaluate evaluation metrics?

Goals

- Correct: better systems have higher scores
- Interpretable: intuitive interpretation of translation quality
- Consistent: repeated use gives the same results
- Low cost: efficient computation, no extra work or linguistic resources needed
- Tuning compatible: can be used to tune translation systems

WMT Metrics shared task

- Long-running comparison of evaluation metrics
- Correlation with human evaluation scores

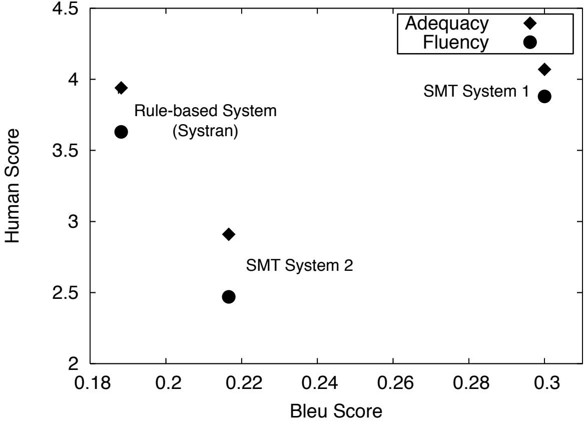
How to evaluate evaluation metrics?

Even if a metric works for comparing similar MT systems, it should not to be trusted for comparing very different ones.

Example from http://www.statmt.org/book/:

Evidence of Shortcomings of Automatic Metrics

Rule-based vs. statistical systems



NMT quality on par with human translators?

Sometimes human evaluation has indicated that NMT would be on the level of human translation.

E.g. paper by Microsoft Research:
“Achieving Human Parity on Automatic Chinese to English News Translation“ (**awadalla2018achieving**)

- Direct assessment (score 0-100) by bilingual humans.
- No statistically significant difference between NMT output and reference translations by humans!

NMT quality on par with human translators?

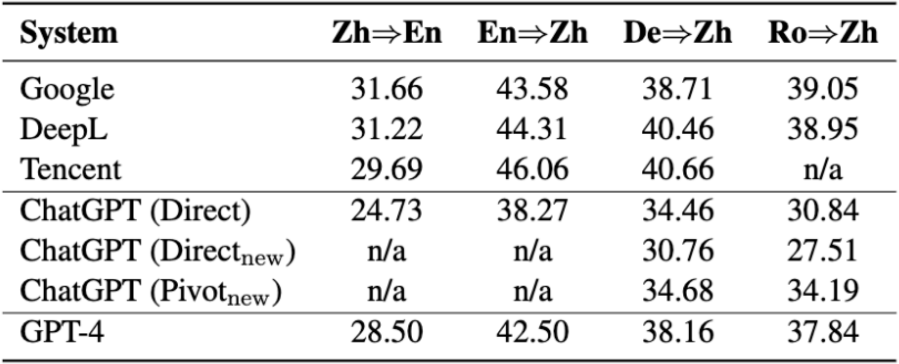
Caveats:

- Are the human translators professionals? Are they translating to their native language?
- How about the human evaluators?
  - Do they understand what to judge (e.g. fluency vs. adequacy)? Even bad NMT is fluent.
  - Skill and time spent: ability to notice subtle differences.
  - Bilingual vs evaluators only speaking target language (use source, or only reference?)
  - Is the document context available?

See e.g. https://www.linkedin.com/pulse/microsoft-mt-reaches-parity-bad-human-translation-tommi-nieminen
or (**toral2018attaining**; **laubli2020set**)

Bibliography

How good is ChatGPT at translation?



System	Zh⇒En	En⇒Zh	De⇒Zh	Ro⇒Zh
Google	31.66	43.58	38.71	39.05
DeepL	31.22	44.31	40.46	38.95
Tencent	29.69	46.06	40.66	n/a
ChatGPT (Direct)	24.73	38.27	34.46	30.84
ChatGPT ($\text{Direct}_{\text{new}}$)	n/a	n/a	30.76	27.51
ChatGPT ($\text{Pivot}_{\text{new}}$)	n/a	n/a	34.68	34.19
GPT-4	28.50	42.50	38.16	37.84

Figure 0: Translation performance of GPT-4 (Date: 2023.03.15).

(jiao2023chatgpt)