Statistical and Adaptive Natural Language Processing

T-61.5020 (5 cr) L, Spring 2008

Tenth lecture

Statistical Machine Translation

Lecturer: **Jaakko Väyrynen**

Slides: Krista Lagus, Philipp Köhn, Mathias Creutz, Timo Honkela, Jaakko Väyrynen

10. Statistical Machine Translation

Lecture based on:

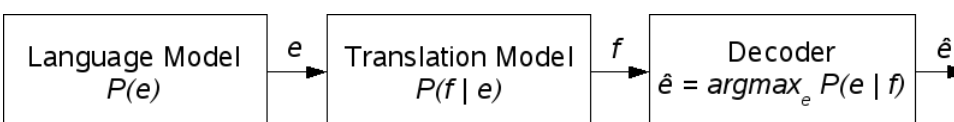
- Chapter 13.2-13.4 in Manning & Schütze
- Chapter 21 in Jurafsky & Martin: Speech and Language Processing (An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition)
- Article on the IBM model: Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer (1993). The Mathematics of Statistical Machine Translation. *Computational Linguistics* 19 (2), pp. 263–312.
- Slides by Philipp Köhn (Koehn), Lecturer (Assistant Professor) at the University of Edinburgh.

10.1 Statistical Approach

- In 1949, Warren Weaver suggested applying statistical and cryptanalytic 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.



- 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:

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

- 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 9).
- 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.
- 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.

10.2 Word Alignment

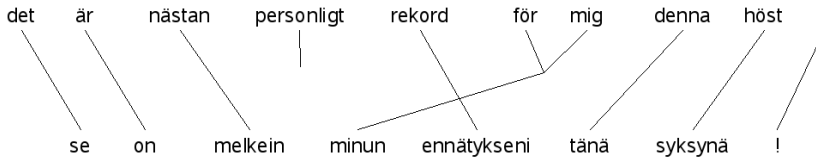
- In the alignment of entire sentences and sections, we did not identify cross-alignments. If there were differences in the order in which the message was conveyed in the two languages, we created large enough beads that comprised multiple sentences on both sides. In this way, we didn't have to rearrange 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 (called *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).

- 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

$$P(f|e) = \frac{1}{Z} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(f_j|e_{a_j}), \tag{4}$$

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

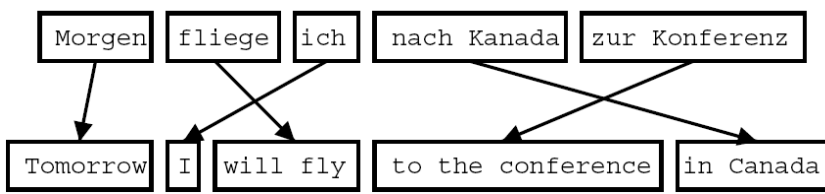
  Nested summations thus sum over all possible alternative alignments and the product over the words in the sentence f.

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

10.3 Phrase Alignment

- Problems with word-to-word translation:
  - "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 word-to-word model, and the phrase-models can be constructed starting from the IBM word-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, the longer phrases can be learned.

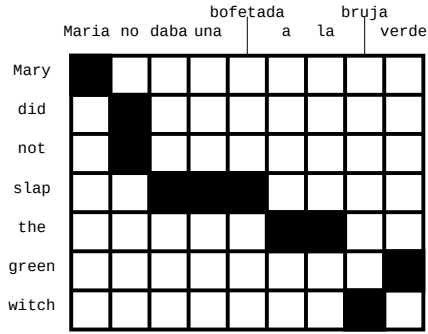
Phrase translation table

- 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*:



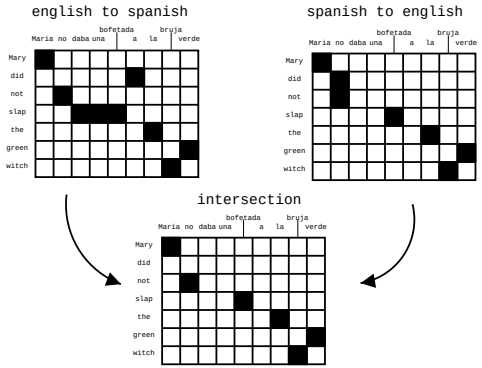
- Collect all phrase pairs that are **consistent** with the word alignment

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)

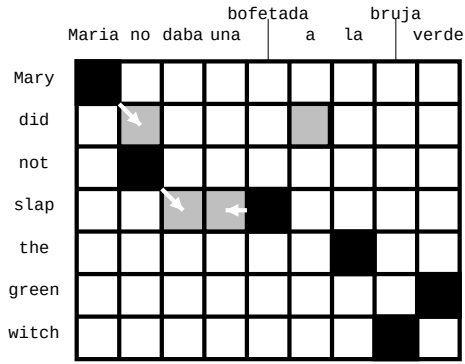
- But we need *many-to-many* mappings

Symmetrizing word alignments



- *Intersection* of GIZA++ bidirectional alignments

Symmetrizing word alignments



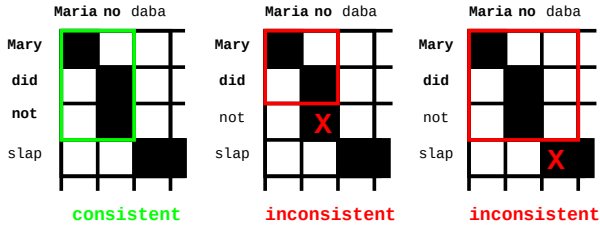
- *Grow* additional alignment points [Och and Ney, CompLing2003]

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

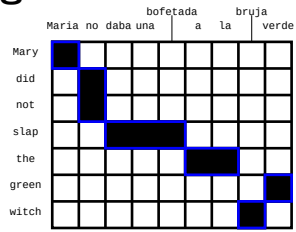


- **Consistent with the word alignment** :=

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

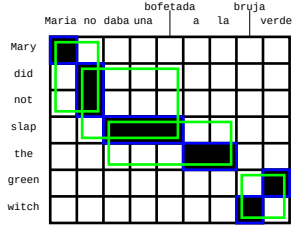
$$(\overline{e}, \overline{f}) \in BP \Leftrightarrow \quad \forall e_i \in \overline{e} : (e_i, f_j) \in A \rightarrow f_j \in \overline{f}$$
$$\text{AND} \quad \forall f_j \in \overline{f} : (e_i, f_j) \in A \rightarrow e_i \in \overline{e}$$

Word alignment induced phrases



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

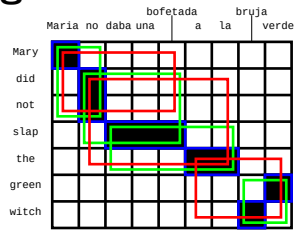
Word alignment induced phrases



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

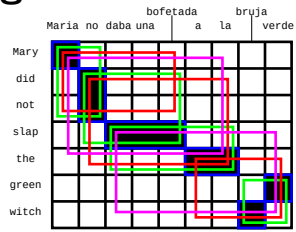
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

Word alignment induced phrases



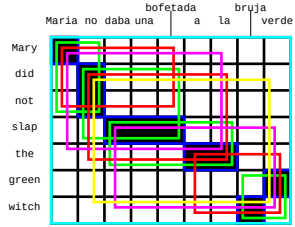
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada 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** $\phi(\overline{f}|\overline{e})$ over the collected phrase pairs

$\Rightarrow$ Possible *choices*

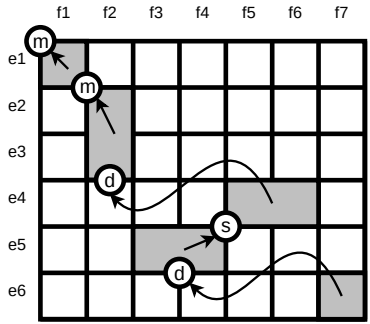
- *relative frequency* of collected phrases: $\phi(\overline{f}|\overline{e}) = \frac{\text{count}(\overline{f},\overline{e})}{\sum_{\overline{f}} \text{count}(\overline{f},\overline{e})}$
- or, conversely $\phi(\overline{e}|\overline{f})$
- use *lexical translation probabilities*

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

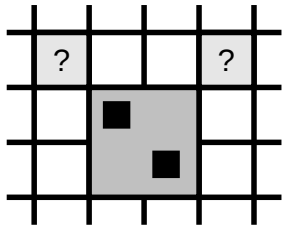
Lexicalized reordering models



[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]

10.4 Evaluation

- How to measure the quality of translations?
-
- Human scores: Subjective Sentence Error Rate (SSER), Information item Error Rate (IER), Information item Semantic Error Rate (ISER)
- But human labor is expensive and time consuming.
- Typically all you have is a held-out test set of sentences with reference translations, in the best case you have multiple reference translations.
- Automatic scores: Sentence Error Rate (SER), Word Error Rate (WER), Multi reference WER (mWER), Position-independent WER (PER), BLEU (Bilingual Evaluation Understudy), NIST, Metric for Evaluation of Translation with Explicit ORdering (METEOR)
- **BLEU** (Bilingual Evaluation Understudy) is a method for evaluating the quality of text which has been translated from one natural language to another using machine translation. BLEU was one of the first

software metrics to report high correlation with human judgments of quality.

- The metric calculates scores for individual fragments, generally sentences, and then averages these scores over the whole corpus in order to reach a final score.

- The metric works by measuring the n-gram (typically 1, 2, 3, and 4-gram) co-occurrence between a given translation and the set of reference translations and then taking the weighted geometric mean. BLEU is specifically designed to approximate human judgment on a corpus level and performs badly if used to evaluate the quality of isolated sentences.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005.

Source Language	Target Language										
	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	**31.2**	**32.1**	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	**30.1**	**31.1**	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	**30.5**	-	**40.2**	12.5	**32.3**	21.4	**35.9**	23.9
fr	23.7	18.5	26.1	**30.0**	**38.4**	-	12.6	**32.4**	21.1	**35.3**	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	**34.0**	**36.0**	11.0	-	20.0	**31.2**	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	**30.1**	**37.9**	**39.0**	11.9	**32.0**	20.2	-	21.9
sv	**30.3**	18.9	22.8	**30.2**	28.6	29.7	15.3	23.9	21.9	25.9	-

Table 2: BLEU scores for the 110 translation systems trained on the Europarl corpus

Examples of Phrase-Based Translation (Europarl Swedish-to-Finnish)

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 marks).

The borders of the phrases used are marked using a vertical bar |. Morph boundaries are not marked:

- **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 !

- **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 .

- **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 .

- **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 .

- **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 .

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 etc.
- If the language model is local (e.g., a 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.