Intro to SMT

Sara Stymne

2018-09-10

Partly based on slides by Jörg Tiedemann and Fabienne Cap

The revolution of the empiricists

Classical approaches require lots of manual work!

- long development times
- low coverage, not robust
- disambiguation at various levels $\rightarrow$ slow!

Learn from translation data:

- example databases for CAT and MT
- bilingual lexicon/terminology extraction
- **statistical translation models**

Motivation for Data-Driven MT

How do we learn to translate?

- grammar vs. examples
- teacher vs. practice
- intuition vs. experience

Is it possible to create an MT engine without any human effort?

- no writing of grammar rules
- no bilingual lexicography
- no writing of preference & disambiguation rules

Motivation for Data-Driven MT

Learning to translate:

- there is a bunch of translated stuff (collect all)
- learn common word/phrase translations from this collection
- look at typical sentences in the target language
- learn how to write a sentence in the target language

Motivation for Data-Driven MT

Learning to translate:

- there is a bunch of translated stuff (collect all)
- learn common word/phrase translations from this collection
- look at typical sentences in the target language
- learn how to write a sentence in the target language

Translation:

- try various translations of words/phrases in given sentence
- put them together, shuffle them around
- check which translation candidate looks best

Motivating example

Imagine a spaceship with aliens coming to earth, telling you:

keipu kaj mei cloy ?

Translation? Anyone?

Motivating example

Imagine a spaceship with aliens coming to earth, telling you:

keipu kaj mei cloy ?

Translation? Anyone?

Problem:

- Human translators may not be available
- Human translators are expensive

Motivating example

Imagine a spaceship with aliens coming to earth, telling you:

keipu kaj mei cloy ?

Translation? Anyone?

Problem:

- Human translators may not be available
- Human translators are expensive

Possible solution:
We found a collection of translated text!

Practical exercise

15–20 minutes

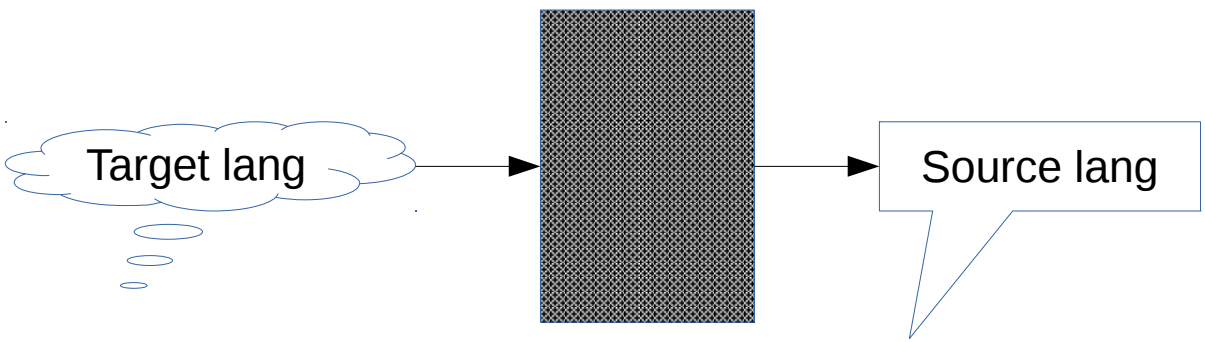
Try to learn to translate the alien language!

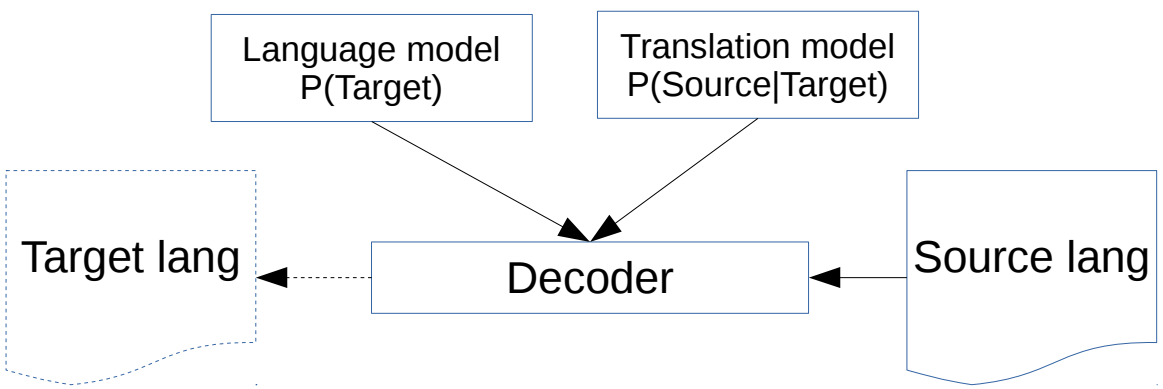
What can we learn from this exercise?

- We can learn, based on translated text
- 1-to-1 translations are easier to identify than 1-to-n n-to-1 or n-to-m
- unseen words cannot be translated
- ambiguity: some words have more than one correct translation $\rightarrow$ the context helps determine which one
- sometimes words need to be reordered

Statistical Machine Translation

Noisy channel for MT: "What could have been the sentence that has generated the observed source language sentence?"

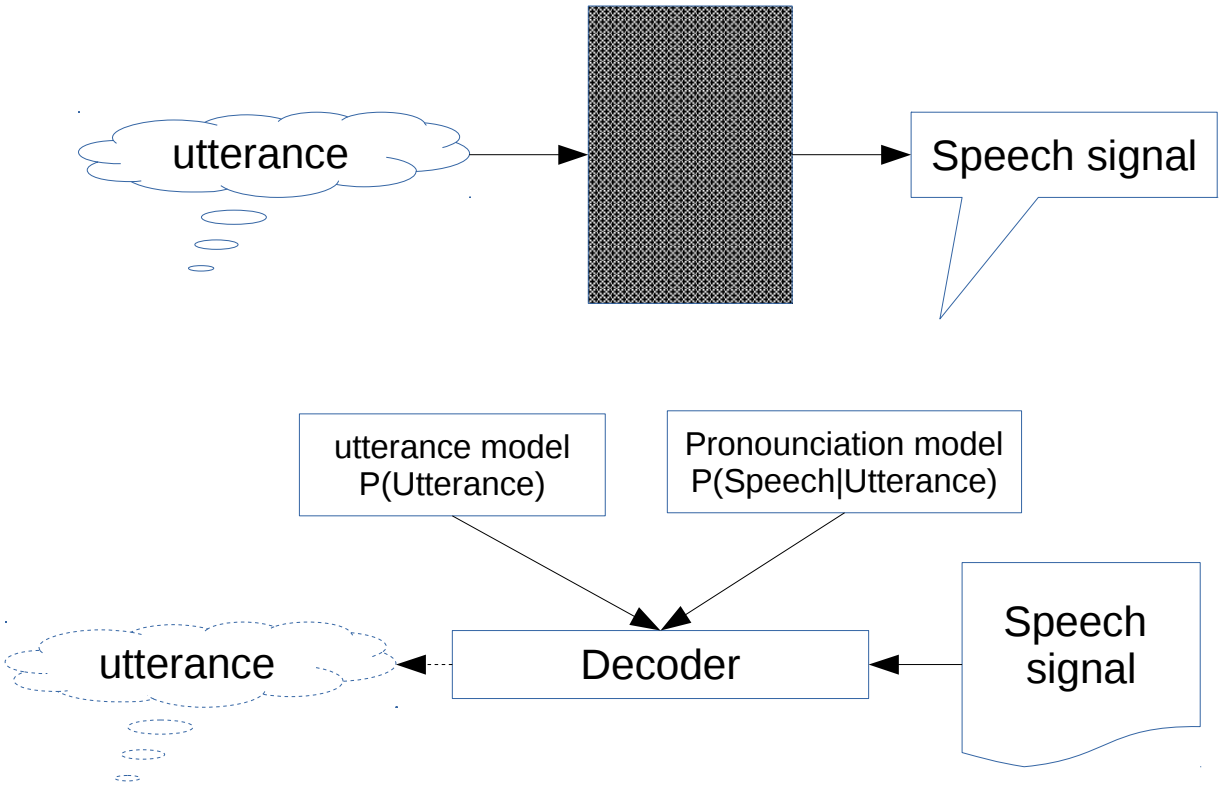




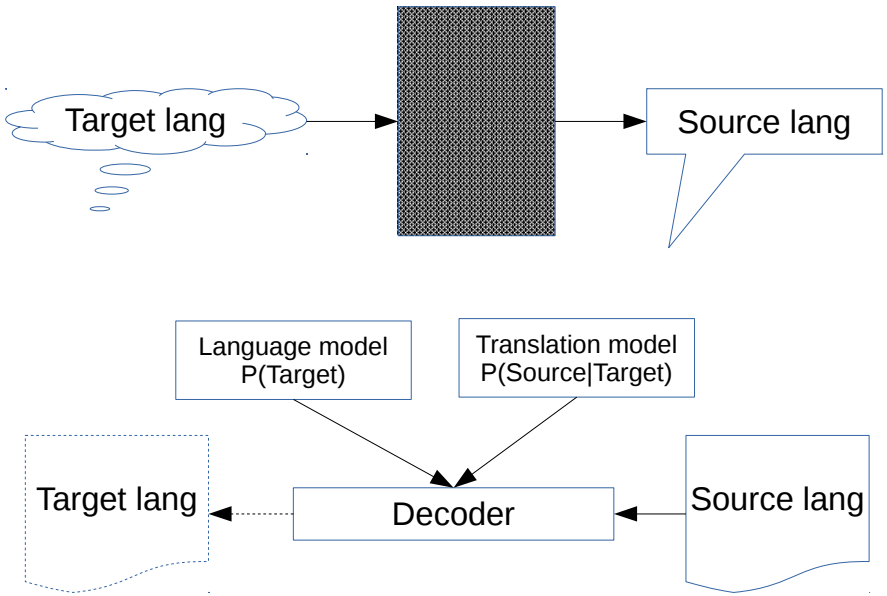
... what a strange idea!

Statistical Machine Translation

Ideas borrowed from Speech Recognition (and in turn from information theory):



Statistical Machine Translation



Probabilistic view on MT (T = target language, S = source language):

$$\begin{aligned} \widehat{T} &= argmax_T P(T|S) \\ &= argmax_T P(S|T)P(T) \end{aligned}$$

Noisy Channel Model vs SMT

Noisy Channel Model	**SMT**	**Example**
Source signal	(desired) SMT output target language text	English text
(noisy) Channel	Translation model	
Reciever (distorted message)	SMT input source language text	Foreign text

Statistical Machine Translation

Informally:

- model translation as an optimization (search) problem
- look for the most likely translation T for a given input S
- use a probabilistic model that assigns these conditional likelihoods
- use Bayes theorem to split the model into 2 parts:
  - a language model (for the target language)
  - a translation model (source language given target language)

Some (very) basic concepts of probability theory

- probability $P(X)$ maps event X to number between 0 and 1
- $P(X)$ represents the likelihood of observing event X in some kind of experiment (trial)
- discrete probability distribution: $\sum_i P(X = x_i) = 1$

Some (very) basic concepts of probability theory

- probability $P(X)$ maps event X to number between 0 and 1
- $P(X)$ represents the likelihood of observing event X in some kind of experiment (trial)
- discrete probability distribution: $\sum_i P(X = x_i) = 1$
- $P(X|Y)$ = conditional probability (likelihood of event X given that event Y has been observed before)

Some (very) basic concepts of probability theory

- probability $P(X)$ maps event X to number between 0 and 1
- $P(X)$ represents the likelihood of observing event X in some kind of experiment (trial)
- discrete probability distribution: $\sum_i P(X = x_i) = 1$
- $P(X|Y) =$ conditional probability (likelihood of event X given that event Y has been observed before)
- joint probability: $P(X, Y)$ (likelihood of seeing both events)
- $P(X, Y) = P(X) * P(Y|X) = P(Y) * P(X|Y)$

Some (very) basic concepts of probability theory

- probability $P(X)$ maps event X to number between 0 and 1
- $P(X)$ represents the likelihood of observing event X in some kind of experiment (trial)
- discrete probability distribution: $\sum_i P(X = x_i) = 1$
- $P(X|Y) =$ conditional probability (likelihood of event X given that event Y has been observed before)
- joint probability: $P(X, Y)$ (likelihood of seeing both events)
- $P(X, Y) = P(X) * P(Y|X) = P(Y) * P(X|Y)$, therefore:

$$\text{Bayes Theorem: } P(X|Y) = \frac{P(X) * P(Y|X)}{P(Y)}$$

Some quick words on probability theory & Statistics

Where do the probabilities come from? $\rightarrow$ Experience!

Use experiments (and repeat them often ....)
Maximum Likelihood Estimation (rely on N experiments only):

$$P(X) \approx \frac{count(X)}{N}$$

Some quick words on probability theory & Statistics

Where do the probabilities come from? $\rightarrow$ Experience!

Use experiments (and repeat them often ....)
Maximum Likelihood Estimation (rely on N experiments only):

$$P(X) \approx \frac{count(X)}{N}$$

For conditional probabilities:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \approx \frac{count(X,Y) * N}{count(Y) * N} = \frac{count(X,Y)}{count(Y)}$$

Translation Model Parameters

Lexical translations:

- das $\rightarrow$ the
- haus $\rightarrow$ house, home, building, household, shell
- ist $\rightarrow$ is
- klein $\rightarrow$ small, low

Multiple translation options:

- learn translation probabilities from data
- Use relative frequencies to estimate probabilites

Context-independent models

Count translation statistics:

- How often is *Haus* translated into:

Translation of *Haus*	Count
house	8,000
building	1,600
home	200
household	150
shell	50
	10,000

Context-independent models

- Maximum likelihood estimation (MLE)

$$t(s|t) = \frac{\text{count(s,t)}}{\text{count(t)}} \tag{1}$$

- for s = *Haus*:

$$\begin{aligned}
\text{t(s|t)} &= 0.8 && \text{if t = house} \\
\text{t(s|t)} &= 0.16 && \text{if t = building} \\
\text{t(s|t)} &= 0.2 && \text{if t = home} \\
\text{t(s|t)} &= 0.015 && \text{if t = household} \\
\text{t(s|t)} &= 0.005 && \text{if t = shell}
\end{aligned}$$

(Classical) Statistical Machine Translation

$$
\begin{aligned}
\hat{T} &= argmax_T P(T|S) \\
&= argmax_T \frac{P(S|T)P(T)}{P(S)} \\
&= argmax_T P(S|T)P(T)
\end{aligned}
$$

(Classical) Statistical Machine Translation

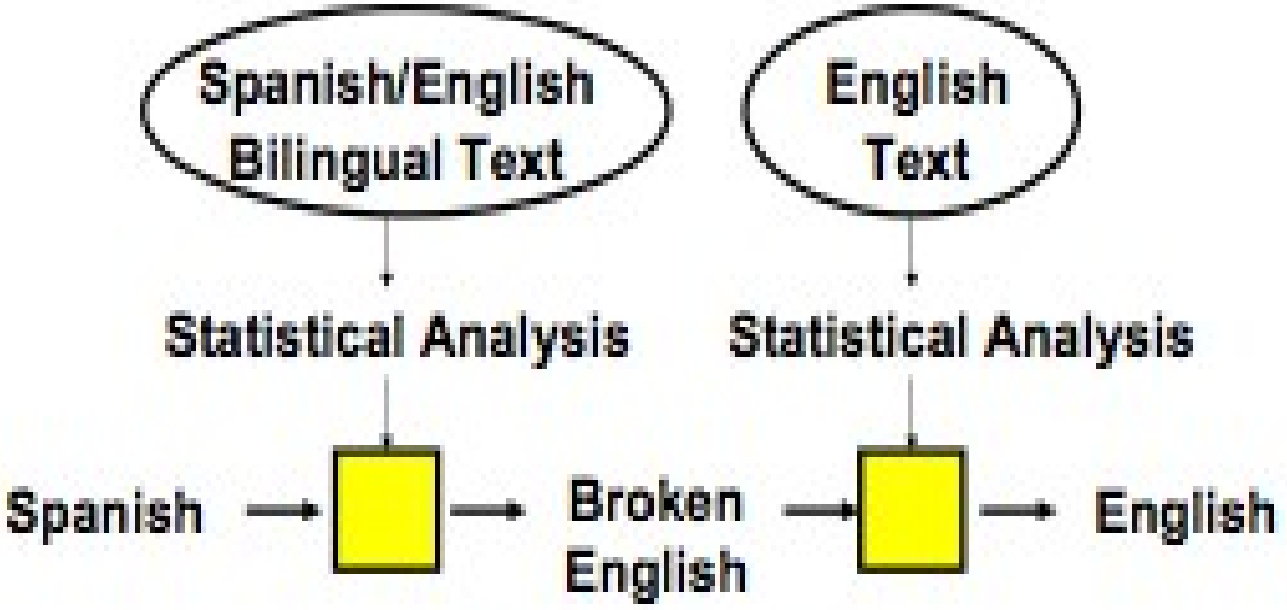
$$\begin{aligned}
\hat{T} &= argmax_T P(T|S) \\
&= argmax_T \frac{P(S|T)P(T)}{P(S)} \\
&= argmax_T P(S|T)P(T)
\end{aligned}$$

Translation model: $P(S|T)$, estimated from (big) parallel corpora, takes care of adequacy

Language model: $P(T)$, estimated from (huge) monolingual target language corpora, takes care of fluency

Decoder: global search for $argmax_T P(S|T)P(T)$ for a given sentence S

Modelling Statistical Machine Translation



Que hambre tengo yo → What hunger have I, Hungry I am so, I am so hungry, Have I that hunger … → I am so hungry

The role of the translation and language model

- Translation model: prefer **adequate** translations
  - P(Das Haus ist klein—The house is small) >
  - P(Das Haus ist klein—The **building** is small) >
  - P(Das Haus ist klein—The **shell** is **low**)
- Language model: prefer **fluent** translations:
  - P(The house is small) >
  - P(The is house small)

Word-based SMT models

Why do we need word alignment?

- Cannot directly estimate $P(S|T)$ ... Why not?

Word-based SMT models

Why do we need word alignment?

- Cannot directly estimate $P(S|T)$ ... Why not?

- almost all sentences are unique
- sparse counts! $\rightarrow$ no good estimations

$\rightarrow$ decompose into smaller chunks!

Word-based SMT models

Why do we need word alignment?

- Cannot directly estimate $P(S|T)$ ... Why not?

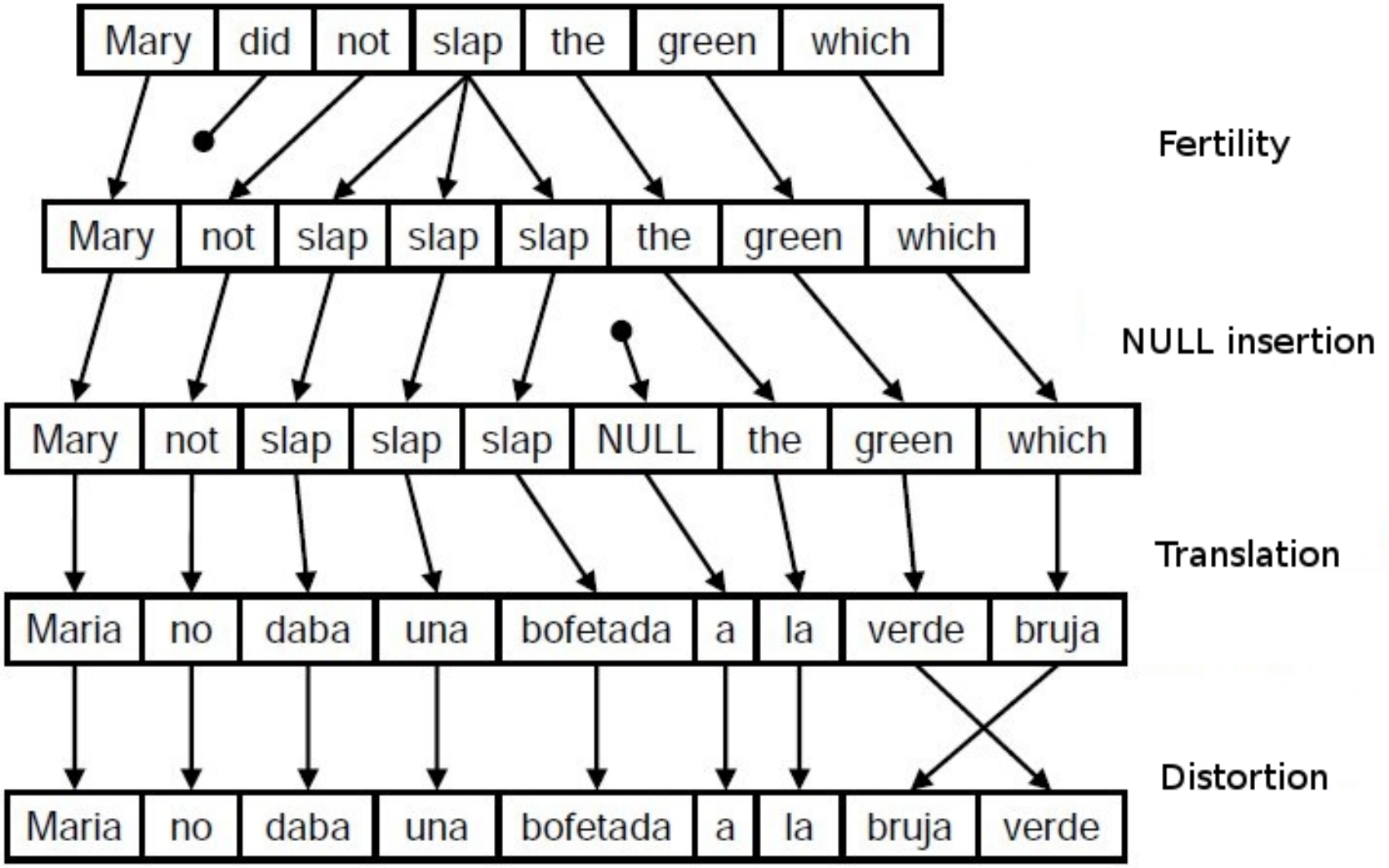
- almost all sentences are unique
- sparse counts! $\rightarrow$ no good estimations

$\rightarrow$ decompose into smaller chunks!

Word-based model: Assume that words in one language have been generated by words in another!

$\rightarrow$ a (hidden) word alignment explains this process

Word-based Translation Models



Word-based Translation Models

What do we need to estimate model parameters?

- lexical translation
- distortion/re-ordering
- fertility
- NULL insertion

→ We need a word-aligned parallel corpus!

Word alignment

How do we formalize word alignment? A simple example:

1	2	3	4
das	Haus	ist	klein
↑	↑	↑	↑
the	house	is	small
1	2	3	4

Word alignment

How do we formalize word alignment? A simple example:

1	2	3	4
das	Haus	ist	klein
↑	↑	↑	↑
the	house	is	small
1	2	3	4

Define alignment function a based on positions:

$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$$

Word alignment

Natural languages are not that easy ...

- not always 1:1 relation between words
- some words may be dropped
- word order can be quite different

Word alignment

Example of "reordering"

1	2	3	4
klein	ist	das	Haus
the	house	is	small
1	2	3	4

What does the alignment function look like?

Word alignment

Example of "reordering"

1	2	3	4
klein	ist	das	Haus
the	house	is	small
1	2	3	4

What does the alignment function look like?

$$a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$$

Word alignment

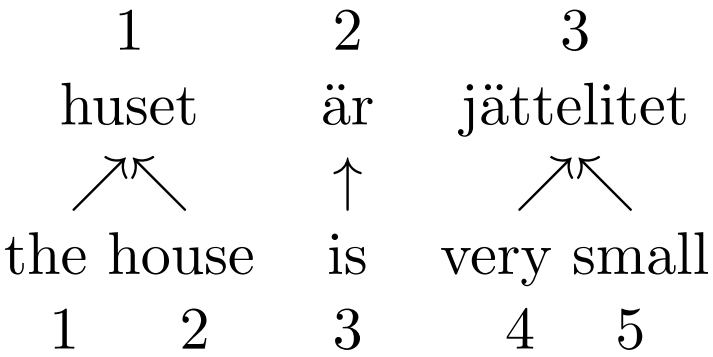
One-to-many alignments

1	2	3
huset	är	jättelitet

the	house	is	very	small
1	2	3	4	5

Word alignment

One-to-many alignments



$$a : \{1 \to 1, 2 \to 1, 3 \to 2, 4 \to 3, 5 \to 3\}$$

Word alignment

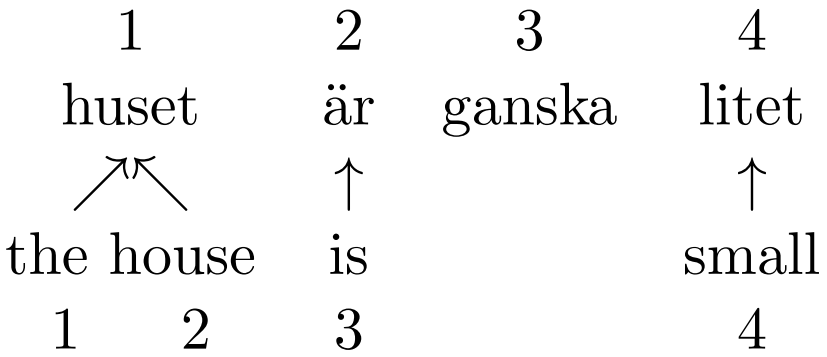
Dropping words:

1	2	3	4
huset	är	ganska	litet

the	house	is		small
1	2	3		4

Word alignment

Dropping words:



$$a : \{1 \to 1, 2 \to 1, 3 \to 2, 4 \to 4\}$$

Word alignment

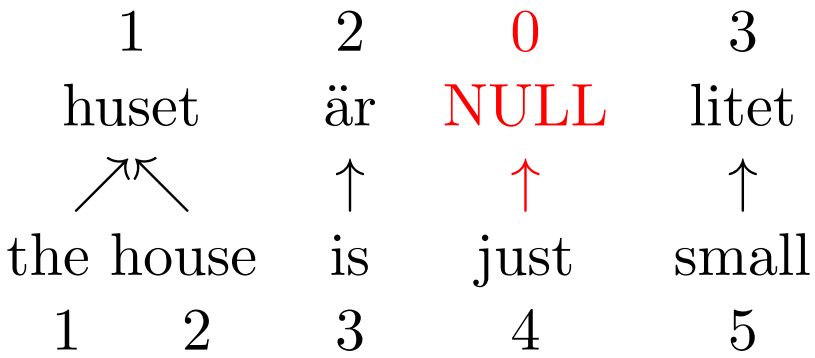
Inserting words:

1	2		3
huset	är		litet

the	house	is	just	small
1	2	3	4	5

Word alignment

Inserting words:



$$a : \{1 \rightarrow 1, 2 \rightarrow 1, 3 \rightarrow 2, 4 \rightarrow 0, 5 \rightarrow 3\}$$

Statistical word alignment models

Standard word-based translation models:

- IBM 1: lexical translation probabilities
- IBM 2: add absolute reordering
- IBM 3: add fertility
- IBM 4: relative reordering
- IBM 5: fix deficiency

Statistical word alignment models

Standard word-based translation models:

- IBM 1: lexical translation probabilities
- IBM 2: add absolute reordering
- IBM 3: add fertility
- IBM 4: relative reordering
- IBM 5: fix deficiency

How can we learn model parameters from parallel corpora without explicit word-alignment?

→ Next time more about this ...

A note on word-based SMT

Today, word-based translation models are **outdated**, but they introduce some **important concepts** which are still relevant for state-of-the-art SMT models:

- generative modelling
- noisy-channel model
- **IBM models 1–5**
- **expectation-maximisation algorithm**

More details next lecture

Statistical Machine Translation

Remember:

$$\hat{T} = argmax_T P(S|T)P(T)$$

- aligned parallel corpora $\rightarrow$ translation model

What is missing?

Statistical Machine Translation

Remember:

$$\hat{T} = argmax_T P(S|T)P(T)$$

- aligned parallel corpora → translation model

What is missing?

- aligned parallel corpora → translation model $P(S|T)$
- we still need the language model $P(T)$

→ Standard N-gram language models

Statistical Machine Translation: Language Modeling

Language modeling:

- (probabilistic) LM = predict likelihood of any given string
- What is the likelihood $P(T)$ to observe sentence T?

$P_{LM}(\text{the house is small}) > P_{LM}(\text{small the is house})$
$P_{LM}(\text{small step}) > P_{LM}(\text{little step})$

N-gram language models

- Markov chain
  - $p(w_1, w_2, \ldots, w_n) =$
    $p(w_1)p(w_2|w_1)\ldots p(W_n|w_1, w_2, \ldots, w_{n-1})$
- Markov assumption
  - $p(w_1, w_2, \ldots, w_n) \simeq p(w_n|w_{n-m}, \ldots, w_{n-2}, w_{n-1})$
- Maximum likelihood estimation (3-grams)
  - $p(w_3|w_1, w_2) = \frac{\text{count}(w_1,w_2,w_3)}{\sum_w \text{count}(w_1,w_2,w)}$

N-gram language models

- Markov chain
  - $p(w_1, w_2, \ldots, w_n) =$
    $p(w_1)p(w_2|w_1)\ldots p(W_n|w_1, w_2, \ldots, w_{n-1})$
- Markov assumption
  - $p(w_1, w_2, \ldots, w_n) \simeq p(w_n|w_{n-m}, \ldots, w_{n-2}, w_{n-1})$
- Maximum likelihood estimation (3-grams)
  - $p(w_3|w_1, w_2) = \frac{\text{count}(w_1,w_2,w_3)}{\sum_w \text{count}(w_1,w_2,w)}$

- unigram model: $P(E) = P(e_1) * P(e_2)...P(e_n)$
- bigram model:
  $P(E) = P(e_1) * P(e_2|e_1) * P(e_3|e_2)...P(e_n|e_{n-1})$
- trigram model:
  $P(E) = P(e_1) * P(e_2|e_1) * P(e_3|e_1, e_2)...P(e_n|e_{n-2}e_{n-1},)$

Summary

- MT can be put into a probabilistic framework
- **translation models**: estimated from parallel corpora
- **language models**: estimated from monolingual corpora
- global search = **decoding** = translating

→ fully automatic (!!!)
→ various simplifications / assumptions necessary
→ probabilistic variant of direct translation

Coming up

- This week:
  - Assignment 2a: word-based SMT
    - Work in any pairs
    - Play with TM (fertility and translation) and LM to gain insights (not a natural working model)
- Next week:
  - Lecture: Alignment and LMs
  - Assignment 2b: word-based SMT and LMs
  - Lab: alignment (not Master 5)
  - Lecture: Phrase-based SMT