Introduction to Statistical Machine Translation

Fabienne Cap

UPPSALA
UNIVERSITY
SWEDEN

Where are we?

Type	Date	Time	Place	Topic	Reading / Assignments
F	2016-03-30	10-12	6-K1031	Introduction (SS)	Koehn 1; JM 25.1-2; Hutchins; CFMF
F	2016-03-30	14-16	6-K1031	MT evaluation (SS)	Koehn 8; JM 25.9
F	2016-04-04	10-12	2-0076	MT in practice (Convertus) - guest lecture	
L	2016-04-06	10-12	Chomsky	MT in practice (AS)	lab report 1
F	2016-04-11	10-12	2-0076	Introduction to SMT (FC)	Koehn Ch 4, Ch 7, KK97
L	2016-04-13	10-12	Chomsky	Word-based SMT (SS)	lab report 2
L	2016-04-18	10-12	Chomsky	Word-based SMT (SS)	lab report 2
F	2016-04-18	14-16	2-1077	Machine translation at Semantix, a translation provider - guest lecture	
F	2016-04-20	10-12	6-K1031	Parallel Corpora, Alignment (AS)	Koehn 2-4, JT 3-4, KK97, KK99
L	2016-04-25	10-12	Chomsky	Parallel corpora & alignment (AS)	lab report 3
F	2016-04-27	10-12	2-0076	Phrase-based SMT (FC)	Koehn Ch 5
L	2016-05-02	10-12	Chomsky	Phrase-based SMT (AS)	lab report 4
F	2016-05-04	10-12	6-K1031	Decoding (CH)	Koehn Ch 6
L	2016-05-09	10-12	Chomsky	Phrase-based SMT (AS)	lab report 4
F	2016-05-11	10-12	2-0076	Tree-based SMT & MT for morphologically rich languages (SS, FC)	Koehn 10.2, 11
F	2016-05-16	10-12	2-0076	Document-wide decoding & Neural MT (CH)	
L	2016-05-18	10-12	Chomsky	Document-wide decoding lab (AS)	oral lab report 5
S	2016-05-23	10-12	2-0076	Seminar - master student presentations	
S	2016-05-25	10-12	6-K1031	Seminar - master student presentations	

Where are we?

Type	Date	Time	Place	Topic	Reading / Assignments
F	2016-03-30	10-12	6-K1031	Introduction (SS)	Koehn 1; JM 25.1-2; Hutchins; CFMF
F	2016-03-30	14-16	6-K1031	MT evaluation (SS)	Koehn 8; JM 25.9
F	2016-04-04	10-12	2-0076	MT in practice (Convertus) - guest lecture	
L	2016-04-06	10-12	Chomsky	MT in practice (AS)	lab report 1
F	2016-04-11	10-12	2-0076	Introduction to SMT (FC)	Koehn Ch 4, Ch 7, KK97
L	2016-04-13	10-12	Chomsky	Word-based SMT (SS)	lab report 2
L	2016-04-18	10-12	Chomsky	Word-based SMT (SS)	lab report 2
F	2016-04-18	14-16	2-1077	Machine translation at Semantix, a translation provider - guest lecture	
F	2016-04-20	10-12	6-K1031	Parallel Corpora, Alignment (AS)	Koehn 2-4, JT 3-4, KK97, KK99
L	2016-04-25	10-12	Chomsky	Parallel corpora & alignment (AS)	lab report 3
F	2016-04-27	10-12	2-0076	Phrase-based SMT (FC)	Koehn Ch 5
L	2016-05-02	10-12	Chomsky	Phrase-based SMT (AS)	lab report 4
F	2016-05-04	10-12	6-K1031	Decoding (CH)	Koehn Ch 6
L	2016-05-09	10-12	Chomsky	Phrase-based SMT (AS)	lab report 4
F	2016-05-11	10-12	2-0076	Tree-based SMT & MT for morphologically rich languages (SS, FC)	Koehn 10.2, 11
F	2016-05-16	10-12	2-0076	Document-wide decoding & Neural MT (CH)	
L	2016-05-18	10-12	Chomsky	Document-wide decoding lab (AS)	oral lab report 5
S	2016-05-23	10-12	2-0076	Seminar - master student presentations	
S	2016-05-25	10-12	6-K1031	Seminar - master student presentations	

Goals for Today

High-level introduction to Statistical Machine Translation

Word-based Translation Models

Noisy Channel Model

Language Models

The Dark Side (e.g. Siths)	[STAR WARS POSTER]	The Light Side (e.g. Jedis)

Tegu mus kelias antai kash.

Translation? Anyone?

Problem:

→ Human translators may not be available

→ Human translators are expensive

Possible solution:

We found a collection of translated texts!

Practical Exercise

15min - 20min

May the force be with you!

Lessons Learned

What do we learn from this exercise?

- 1-to-1 translations 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 determines which one
- sometimes words need to be re-orderend

Goals for Today

High-level introduction to Statistical Machine Translation

Word-based Translation Models

Noisy Channel Model

Language Models

Word-based Translation Models

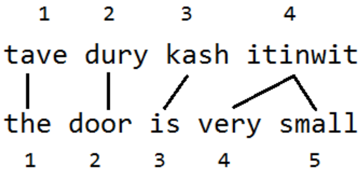
Today, word-based translation models are **outdated**, but they introduce some **important concepts** which are still relevant for state-of-the-art SMT models:

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

$\rightarrow$ **more details in Aaron's lecture on word alignment!**

Generative Modelling

Generative Model: source language words are generated by target language words

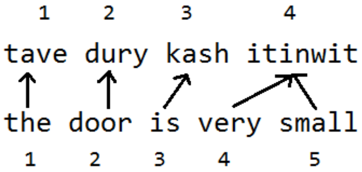


Introduce an **alignment function** a: i $\rightarrow$ j

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

Generative Modelling

Generative Model: source language words are generated by target language words



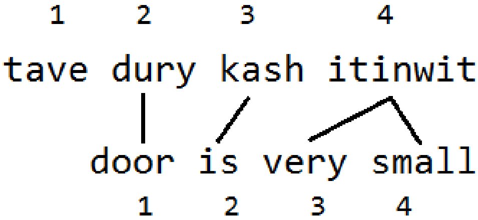
Introduce an **alignment function** a: i $\rightarrow$ j

a:$\{1\rightarrow2,\ 2\rightarrow2,\ 3\rightarrow3,\ 4\rightarrow4,\ 5\rightarrow4\}$

Translation: Decode what kind of English word sequence has generated the Sith word sequence

Special Cases in Word Alignment

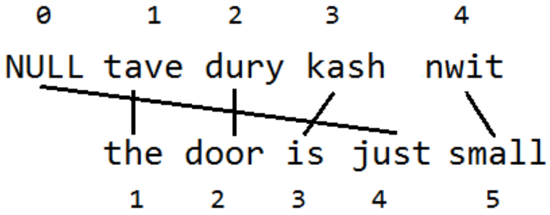
Dropping words:



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

Special Cases in Word Alignment

Inserting words: Introduce a special NULL word



a{1→1, 2→2, 3→3, 4→0, 5→4}

Translation Model Parameters

Lexical Translations

- tave $\rightarrow$ the
- dury $\rightarrow$ door
- kash $\rightarrow$ is, in
- nwit $\rightarrow$ smal

In case of multiple translation options:

- use the most common on in that context

Context-Independent Models

Count translation statistics:

How often is **dury** translated into...

Translation of **dury**	Count
door	8,000
portal	1,600
entrance	200
doorway	150
gate	50

Context-Independent Models

Estimate translation probabilities:

- Maximum Likelihood Estimation (MLE)

  $t(english|sith) = \frac{count(english,sith)}{count(sith)}$

- for sith = dury:

$$t(english|sith) = \begin{cases} 0.8 & \text{if english} = \text{door} \\ 0.16 & \text{if english} = \text{portal} \\ 0.02 & \text{if english} = \text{entrance} \\ 0.015 & \text{if english} = \text{doorway} \\ 0.005 & \text{if english} = \text{gate} \end{cases}$$

Goals for Today

High-level introduction to Statistical Machine Translation

Word-based Translation Models

Noisy Channel Model

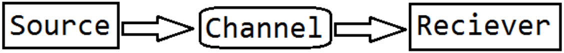
Language Models

Noisy Channel Model

What is a noisy channel?

https://www.youtube.com/watch?v=0MUsVcYhERY

Noisy Channel Model



- origin in acoustics and information theory
- idea: foreign language sentence is a message distorted through a noisy channel
- decode distorted message and restore original message
- use two models:
  - source model p(Source) (= language model)
  - channel model p(Recieved|Source) (= translation model)

Caution Confusing Terminology!!!

Noisy Channel Model vs. SMT

Noisy Channel Model	SMT	our example
Source signal	(desired) SMT output target language text	English text
(noisy) Channel	Translation model	
Reciever (distorted message)	SMT input source language text	Sith text

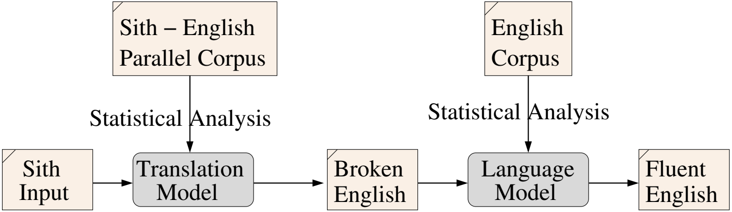
Modelling Statistical Machine Translation

Use Bayes' rule to decompose P(english|sith) into

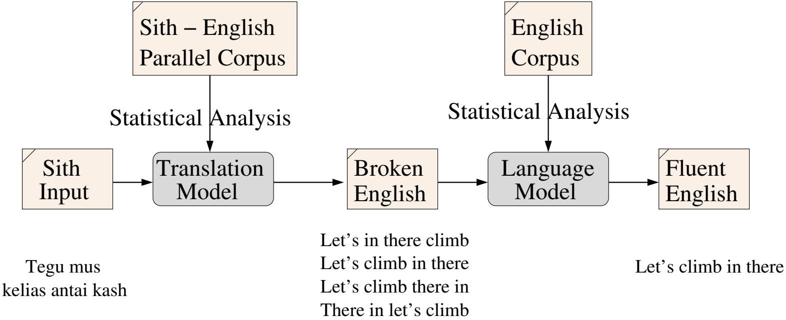
- Translation Model P(sith|english)
- Language Model P(english)

$$
\begin{aligned}
argmax_e P(e|s) &= argmax_e \frac{P(s|e) * P(e)}{P(s)} \\
&= argmax_e P(s|e) * P(e)
\end{aligned}
$$

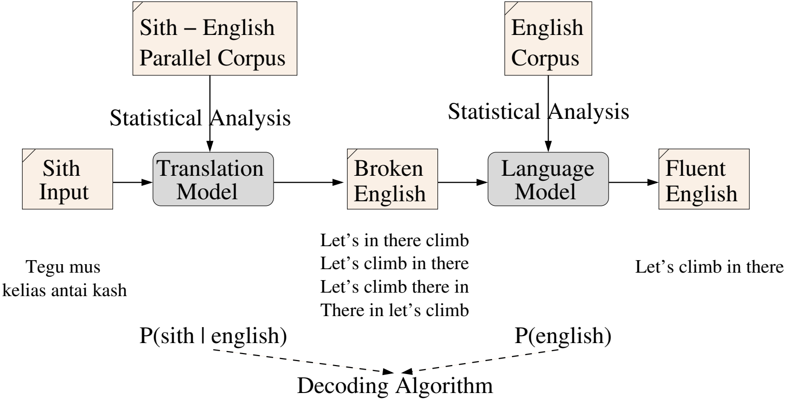
Modelling Statistical Machine Translation



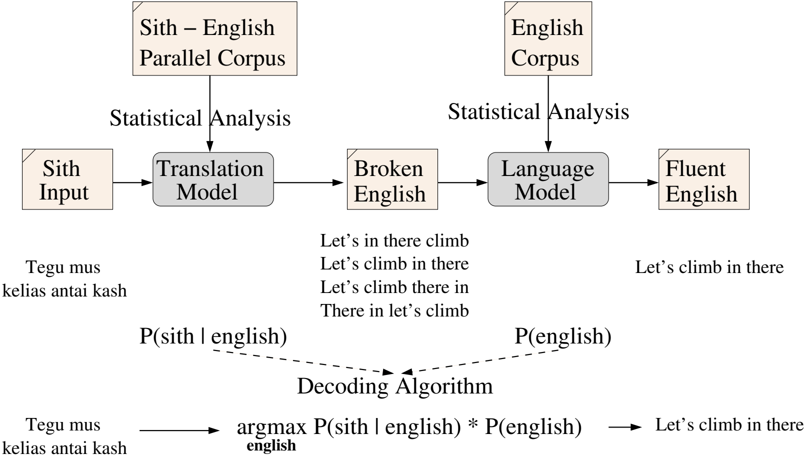
Modelling Statistical Machine Translation



Modelling Statistical Machine Translation



Modelling Statistical Machine Translation



The role of the Translation and the Language Model

Translation Model: prefers **adequate** translations

- P(Tegu mus kelias antai kash|Let's climb in there) >
- P(Tegu mus kellias antai kash|Let's climb in **here**) >
- P(Tegu mus kelias antai kash|Let's **clamber** in there)

Language Model: prefers grammatical/**fluent** sequences

- P(Let's climb in there) > P(Let's there climb in)

Goals for Today

High-level introduction to Statistical Machine Translation

Word-based Translation Models

Noisy Channel Model

Language Models

Statistical Language Models

Prefer one string over another (ensure fluency)

- "small step": 5,880,000 hits on Google
- "little step": 1,780,000 hits on Google

Language Model:

estimate how likely a string is in a given language:

P_{LM}(the door is small) $>$ P_{LM}(small the is door)
P_{LM}(let's climb in there) $>$ P_{LM}(let's climb is there)

N-Gram Language Models

Markov chain

$p(w_1, w_2, ..., w_n) = p(w_1)p(w_2|w_1)...p(w_n|w_1, w_2, ..., w_{n-1})$

Markov assumption

$p(w_n|w_1, w_2, ..., w_{n-1}) = p(w_n|w_{n-m}, ..., w_{n-2}, w_{n-1})$

Maximum likelihood estimation (e.g. 3-gram)

$p(w_3|w1, w2) = \frac{count(w_1, w2, w3)}{\sum_w count(w_1, w_2, w)}$

N-Gram Language Models

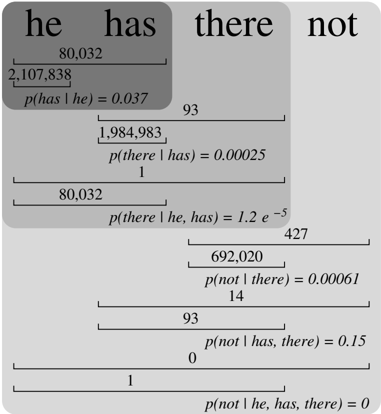
Add **special markers** at the start and the end of the sentence!
$\rightarrow$ certain tokens often appear at the start or at the end

Smoothing

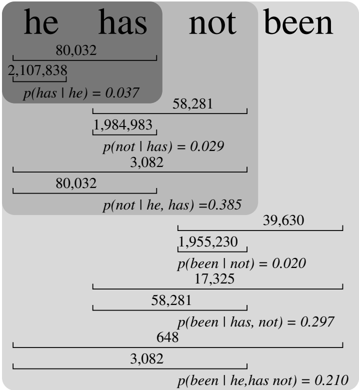
- big problem: unseen n-grams $\rightarrow$ p(e) $= 0$
- smoothing: reserve probability mass for unseen events

Interpolation and backoff
$\rightarrow$ combine higher-order and lower-order models

N-Gram Language Models - Example



$p(he, has, there, not) =$
$0.037 * 0.00025 * 1.2\ e^{-5} * 0.00061 * 0.15 * 0 = 0$



$p(he, has, not, been) =$
$0.037 * 0.029 * 0.385 * 0.020 * 0.297 * 0.210 = 5{,}15e^{-7}$

Goals for Today

High-level introduction to Statistical Machine Translation

Word-based Translation Models

Noisy Channel Model

Language Models

IBM Models for Word Alignment

A short overview on IBM Models for Word Alignment

IBM Model 1

IBM Model 1 only uses lexical translation

das

e	$t(e\|f)$
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus

e	$t(e\|f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

ist

e	$t(e\|f)$
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

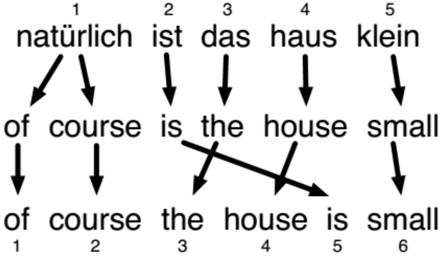
klein

e	$t(e\|f)$
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

$$
\begin{aligned}
p(e,a|f) &= \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein}) \\
&= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4 \\
&= 0.0028\epsilon
\end{aligned}
$$

IBM Model 2

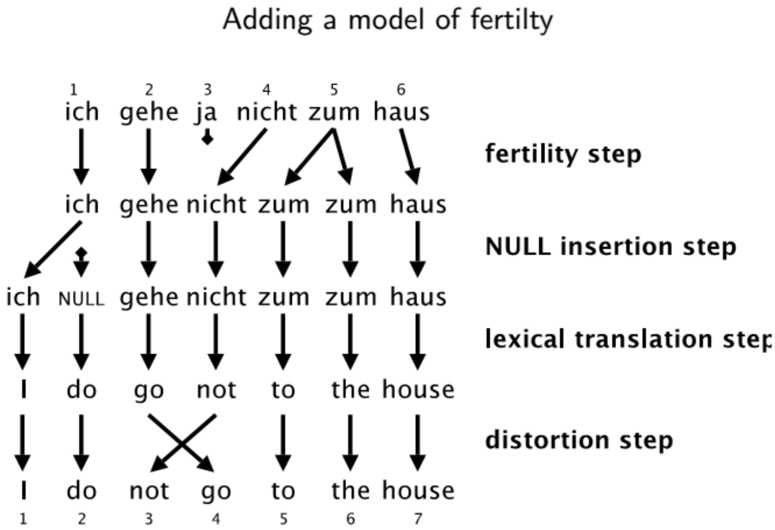
Adding a model of alignment



lexical translation step

alignment step

IBM Model 3



Adding a model of fertilty
1 2 3 4 5 6
ich gehe ja nicht zum haus
fertility step
ich gehe nicht zum zum haus
NULL insertion step
ich NULL gehe nicht zum zum haus
lexical translation step
I do go not to the house
distortion step
I do not go to the house
1 2 3 4 5 6 7

IBM Model 4

Motivation

- Absolute position for distortion feels wrong
- Words do not move independently
- Some words tend to move and some not

$\rightarrow$ Introduce a relative distortion model

$\rightarrow$ Introduce a dependence on word classes

IBM Model 5

IBM Models 1-4 are deficient

- some impossible translations have positive probabilities
- multiple output words may be placed in the same place
- probability mass is wasted!

IBM Model 5

- fix deficiency by keeping track of vacancies
- details: see text book

Summary IBM Models

Models with increasing complexity

Higher models include more information

IBM Model 1	lexical translation
IBM Model 2	adds absolute alignment model
IBM Model 3	adds fertility model
IBM Model 4	relative alignment model
IBM Model 5	fixes deficiency

Summary

Take Home Messages from Today??

Where do we go?

Type	Date	Time	Place	Topic	Reading / Assignments
F	2016-03-30	10-12	6-K1031	Introduction (SS)	Koehn 1; JM 25.1-2; Hutchins; CFMF
F	2016-03-30	14-16	6-K1031	MT evaluation (SS)	Koehn 8; JM 25.9
F	2016-04-04	10-12	2-0076	MT in practice (Convertus) - guest lecture	
L	2016-04-06	10-12	Chomsky	MT in practice (AS)	lab report 1
F	2016-04-11	10-12	2-0076	Introduction to SMT (FC)	Koehn Ch 4, Ch 7, KK97
L	2016-04-13	10-12	Chomsky	Word-based SMT (SS)	lab report 2
L	2016-04-18	10-12	Chomsky	Word-based SMT (SS)	lab report 2
F	2016-04-18	14-16	2-1077	Machine translation at Semantix, a translation provider - guest lecture	
F	2016-04-20	10-12	6-K1031	Parallel Corpora, Alignment (AS)	Koehn 2-4, JT 3-4, KK97, KK99
L	2016-04-25	10-12	Chomsky	Parallel corpora & alignment (AS)	lab report 3
F	2016-04-27	10-12	2-0076	Phrase-based SMT (FC)	Koehn Ch 5
L	2016-05-02	10-12	Chomsky	Phrase-based SMT (AS)	lab report 4
F	2016-05-04	10-12	6-K1031	Decoding (CH)	Koehn Ch 6
L	2016-05-09	10-12	Chomsky	Phrase-based SMT (AS)	lab report 4
F	2016-05-11	10-12	2-0076	Tree-based SMT & MT for morphologically rich languages (SS, FC)	Koehn 10.2, 11
F	2016-05-16	10-12	2-0076	Document-wide decoding & Neural MT (CH)	
L	2016-05-18	10-12	Chomsky	Document-wide decoding lab (AS)	oral lab report 5
S	2016-05-23	10-12	2-0076	Seminar - master student presentations	
S	2016-05-25	10-12	6-K1031	Seminar - master student presentations	

SMT and Music

https://www.youtube.com/watch?v=GMi4MtyDg40