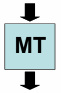
Machine Translation - Foundations

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Jörg Tiedemann
University of Helsinki

Fabienne Cap
Uppsala University

Overview

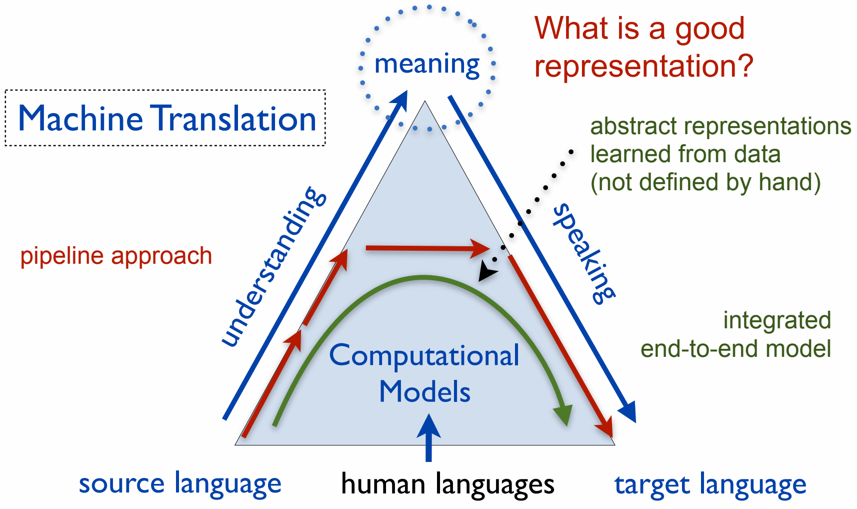
What is this course about?

- introduction of **statistical** machine translation (SMT)
- training SMT models and using them for translation
- identifying MWEs in parallel texts and using them in MT

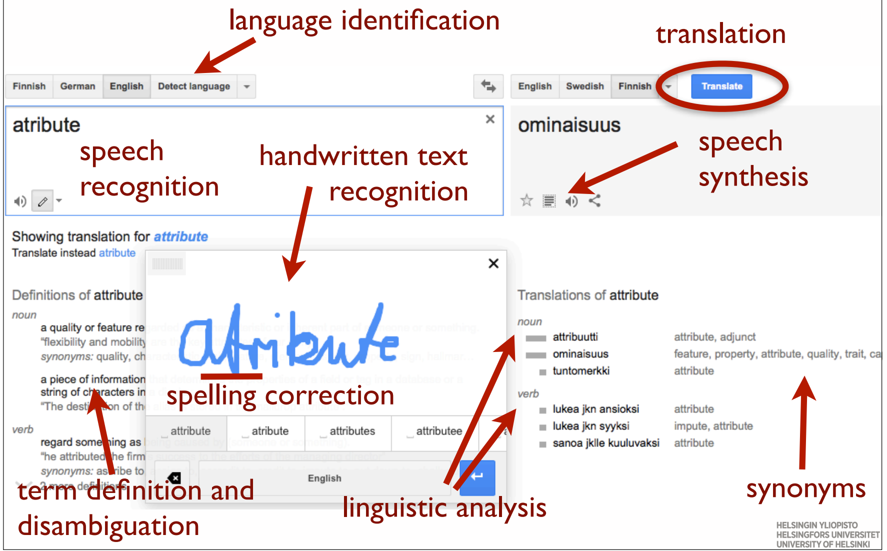
Setup

- **Foundations** of statistical MT
- Translation as **decoding**
- **Multi-word expressions** and MT
- Lab-session on phrase-based SMT

Computational Linguistics and MT



MT and Other Language Technology



atribute

ominaisuus

Showing translation for *attribute*
Translate instead atribute

Definitions of attribute

noun

a quality or feature re
"flexibility and mobility
synonyms: quality, ch

a piece of information
string of characters in
"The desti on of th

verb

regard some ing as
"he attributed he firm
synonyms: as ibe to

attribute | atribute | attributes | attributee

Translations of attribute

noun	
attribuutti	attribute, adjunct
ominaisuus	feature, property, attribute, quality, trait, ca
tuntomerkki	attribute
verb	
lukea jkn ansioksi	attribute
lukea jkn syyksi	impute, attribute
sanoa jklle kuuluvaksi	attribute

Multi-Word Expressions and MT

What are multi-word expressions?

- non-compositionality

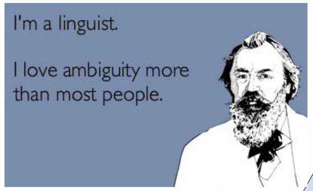
Machine translation models

- must generalise (using composition)

Everything depends on the context

- *I'll **get** a cup of coffee*
- *I didn't **get** the joke*
- *I **get** up at 8am*
- *I **get** nervous*
- *Yeah, I **get** around*

Expert-Driven Rule-Based Systems



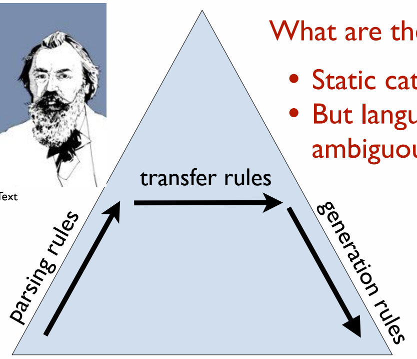
I'm a linguist.

I love ambiguity more than most people.

VP → PP[+Goal] V ⇒ VP → V PP[+Goal] Text

N → house
ADJ → red
Det → the
NP → Det ADJ N

$$\begin{bmatrix} \text{EVENT} & \text{SLAPPING} \\ \text{AGENT} & \text{MARY} \\ \text{TENSE} & \text{PAST} \\ \text{POLARITY} & \text{NEGATIVE} \\ \text{THEME} & \begin{bmatrix} \text{WITCH} & \\ \text{DEFINITENESS} & \text{DEF} \\ \text{ATTRIBUTES} & \begin{bmatrix} \text{HAS-COLOR} & \text{GREEN} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$



meaning

transfer rules

parsing rules

generation rules

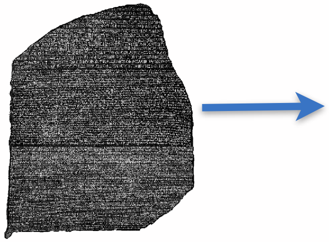
source language

target language

What are the problems?

- Static categorical models
- But languages are dynamic, ambiguous, distributional

Machine Translation as Decoding



learn the unknown code

When I look at an article in Russian, I say:
This is really written in English,
but it has been coded in some strange symbols.
I will now proceed to decode.
[Weaver, 1947, 1949]

Finding Patterns (Knight, 1997)

Your Assignment: Translate **Klingon** to **Acturan**

- **farok crrok hihok yorok clok kantok ok-yurp**

No expert around ...

No grammar at hand ...

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

translated sentence

Klingon

Arcturan

Database of example translations

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat **jjat** bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat **jjat** quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok **crrrok** hihok yorok zanzanok . ??? 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: **farok** crrrok **hihok** yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok **yorok** ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** **yorok** zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

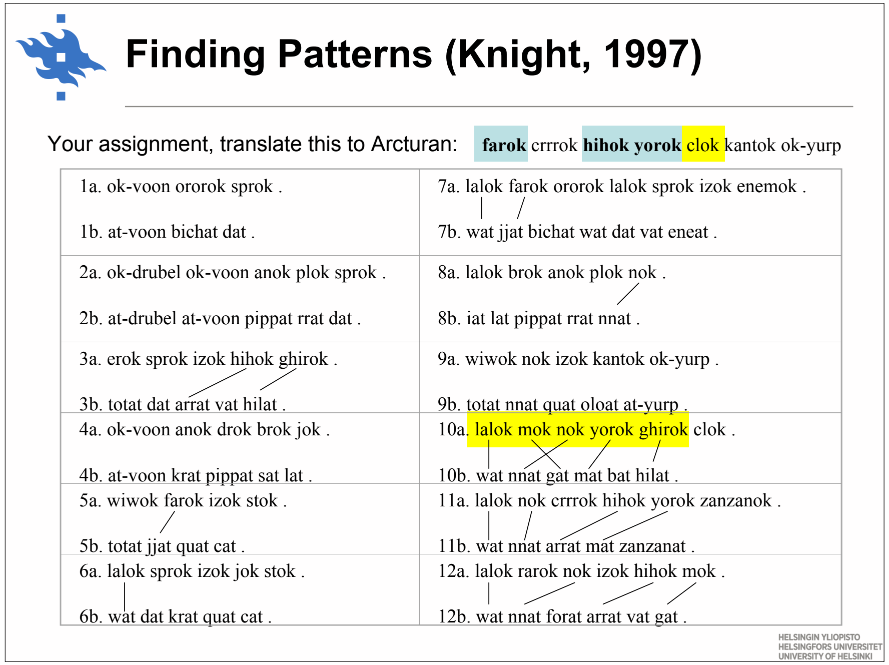
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . ??? 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

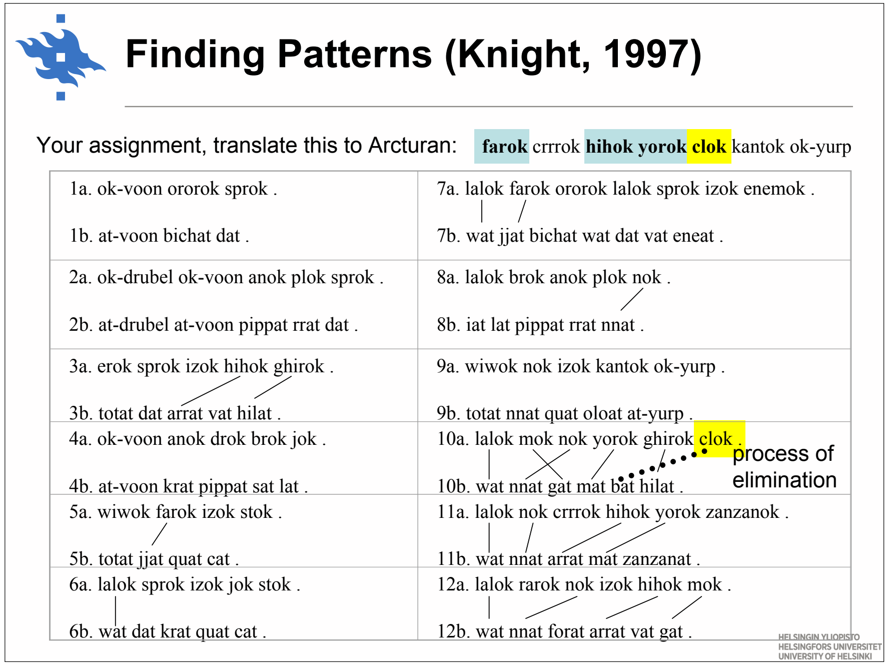
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI



Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

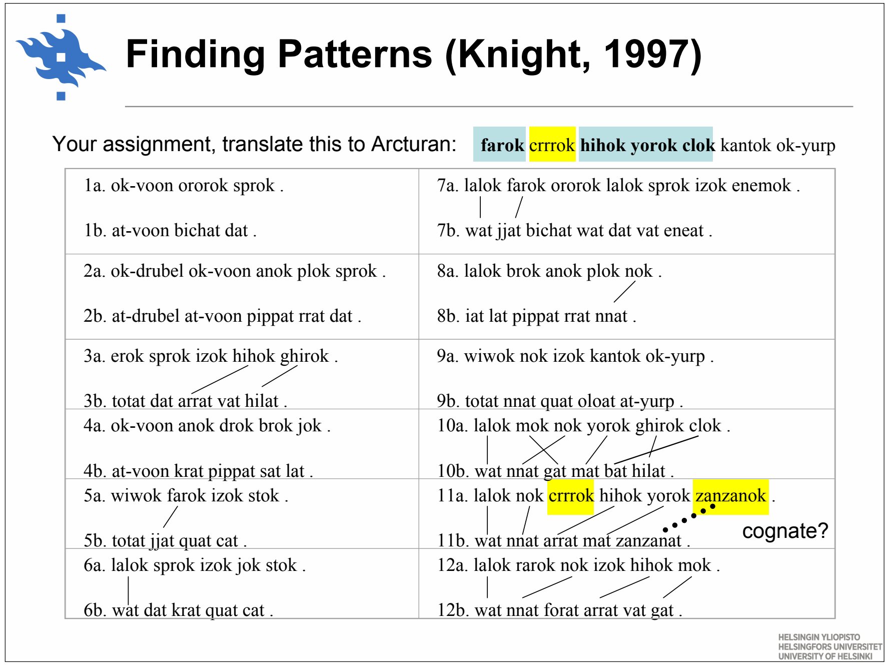


Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

process of elimination

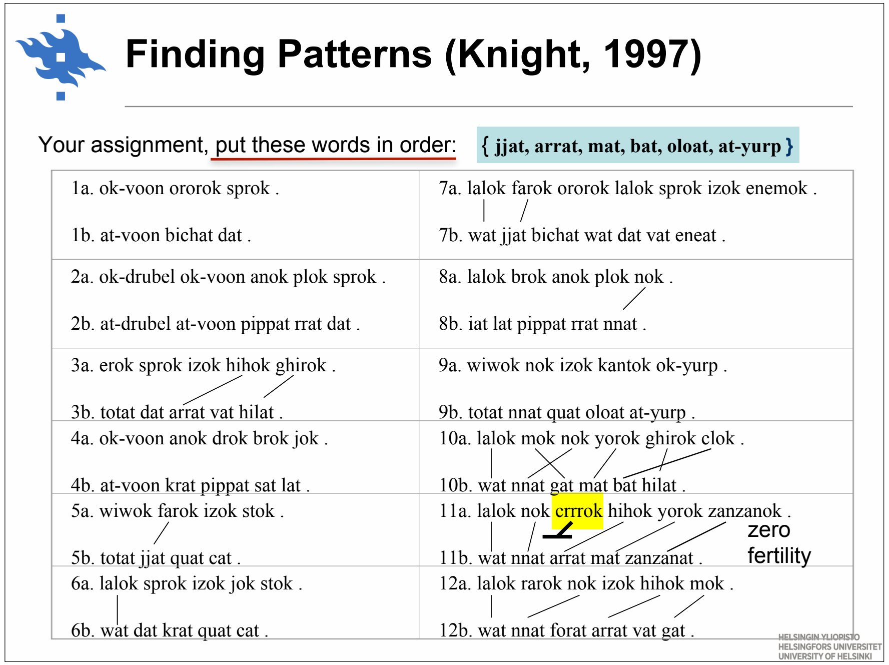


Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

cognate?



Finding Patterns (Knight, 1997)

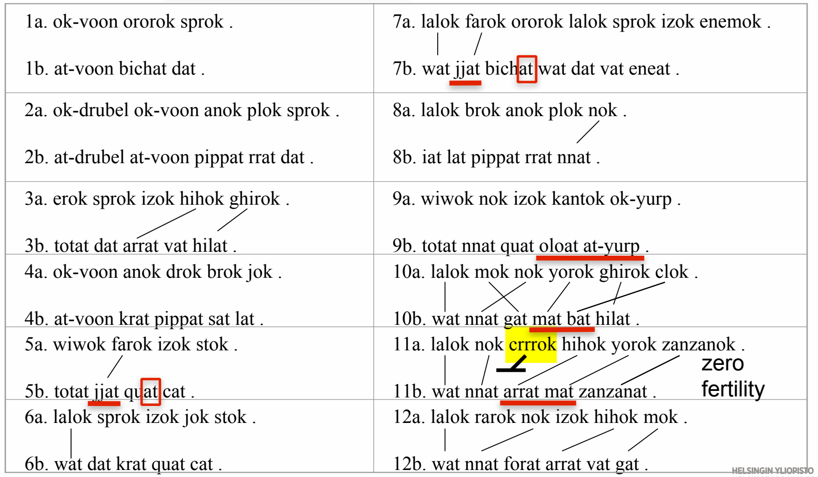
Your assignment, put these words in order: { jjat, arrat, mat, bat, oloat, at-yurp }

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

zero fertility

Finding Patterns (Knight, 1997)

Your assignment, put these words in order: { jjat, arrat, mat, bat, oloat, at-yurp }



1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

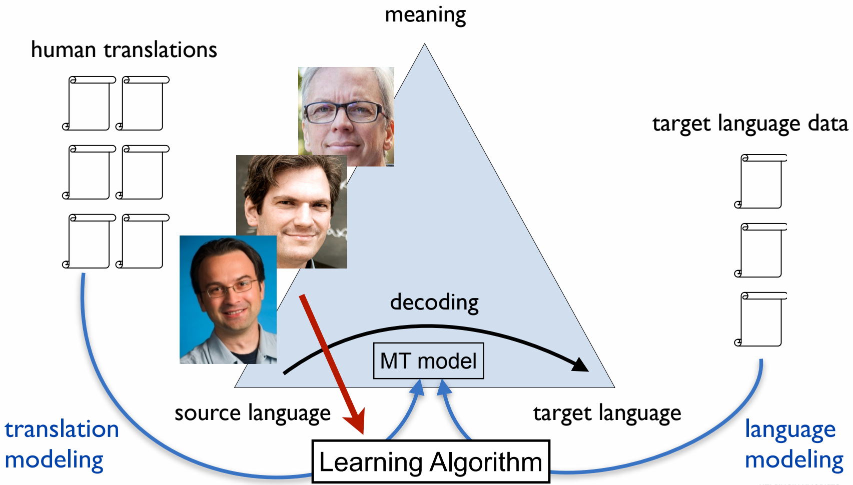
zero fertility

Finding Patterns (Knight, 1997)

Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Data-Driven Machine Translation



Statistical Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

⬇

Probabilistic Model: P(**e**|**f**)

Search problem (decoding): $\mathbf{e}^* = \mathit{argmax}_{\mathbf{e}}\, P(\mathbf{e}|\mathbf{f})$

⬇

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Statistical Machine Translation

Use Bayes' Rule to Decompose *p(**e**|**f**)* into

- Translation Model *p(**f** |**e**)*
- Target Language Model *p(**e**)*

$$\text{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) = \text{argmax}_{\mathbf{e}} \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{p(\mathbf{f})}$$

$$= \text{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

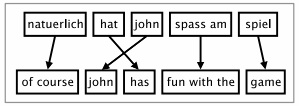
What's in a translation model?

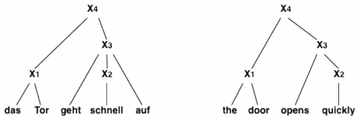
What's in a language model?

Statistical Machine Translation

Modeling Translation

- word-based models
- phrase-based models
- hierarchical models





Learning Model Parameters

- sentence and word alignment
- rule extraction and scoring
- parameter tuning

Decoding

Language Models

Probabilistic 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

$$p_{\text{LM}}(\textit{the house is small}) > p_{\text{LM}}(\textit{small the is house})$$

$$p_{\text{LM}}(\textit{I am going home}) > p_{\text{LM}}(\textit{I am going house})$$

N-Gram Language Models

Parameter estimation

- *p("the house is small") = …?*

the
the house
the house is
house is small
is small .

Markov assumption: Limit context, e.g. trigrams only

- *p(the) * p(house | the) * p(is | the, house) * p(small | house, is)*

Maximum likelihood estimation

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

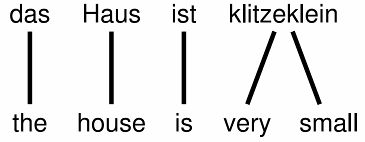
trigram frequency counts in large data sets

frequency of anything else following w1 and w2

Translation Models

Word-Based Translation Models

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



Translation: Find the most likely word sequences that may have generated the foreign string of words.

Context-Independent Models

Words translate without looking at any context

Count translation statistics in aligned training data:

- How often is *Haus* translated into ....

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

Context-Independent Models

Estimate Translation Probabilities:

- Maximum Likelihood Estimation (MLE)

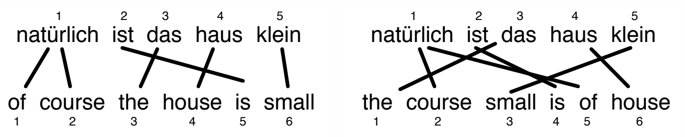
$$t(f|e) = \frac{count(f,e)}{count(e)}$$

- for f = Haus:

$$t(f|e) = \begin{cases} 0.8 & \text{if } e = \text{house},\\ 0.16 & \text{if } e = \text{building},\\ 0.02 & \text{if } e = \text{home},\\ 0.015 & \text{if } e = \text{household},\\ 0.005 & \text{if } e = \text{shell}. \end{cases}$$

Context-Independent Models

What does this mean for *p(**f** |**e**)*?



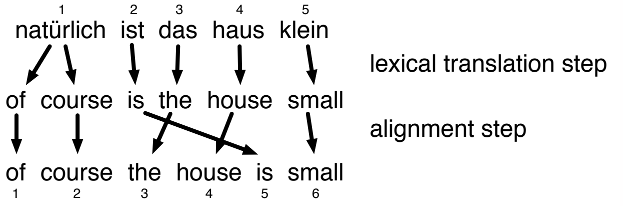
1 natürlich 2 ist 3 das 4 haus 5 klein
of 1 course 2 the 3 house 4 is 5 small 6
1 natürlich 2 ist 3 das 4 haus 5 klein
the 1 course 2 small 3 is 4 of 5 house 6

What does this mean for *p(**e**|**f**) = p(**f** |**e**) p(**e**)*?

- *p(the house is small | das Haus ist klein)*
- *p(the is small house | das Haus ist klein)*
- *p(the house is small | das Haus ist klein)*

Distortion Models

Add a model for positional alignment:



1 natürlich 2 ist 3 das 4 haus 5 klein
of course is the house small
of 1 course 2 the 3 house 4 is 5 small 6
lexical translation step
alignment step

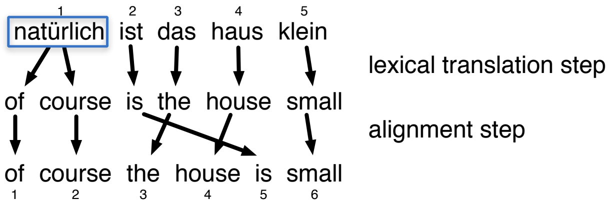
New parameters (different variants):

- *d(pos(e)* = **5** | *pos(f)* = **2**, *length(e)* = **6**, *length(f)* = **5**)
- *d(pos(e)* = **5** | *pos(e-1)* = **2**, *length(e)* = **6**)
- *d(pos(e)* = **5** | *pos(f)* = **2**, *length(e)* = **6**, *f = "ist", e = "is")*

Fertility Models

Add a fertility parameter:

- What is the likelihood that *"natürlich"* generates 2 words?
- What is the likelihood that some words are dropped?



1 natürlich 2 ist 3 das 4 haus 5 klein
of course is the house small
of 1 course 2 the 3 house 4 is 5 small 6
lexical translation step
alignment step

new parameters:

- *n(* **2** *| "natürlich") = 0.9*
- *n(* **1** *| "natürlich") = ...*

Training Word-Based Models

Need word aligned parallel training data

- large quantities of translated documents
- automatic sentence alignment

Parameter estimation

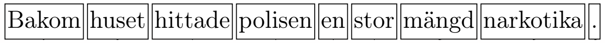
- word alignments as latent (hidden) variables
- expectation-maximisation algorithm

Language modeling

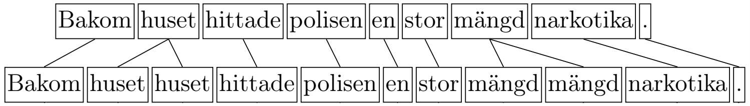
- probabilistic n-gram models trained on large monolingual data

Phrase-Based Models

Generative Model of Word-Based SMT

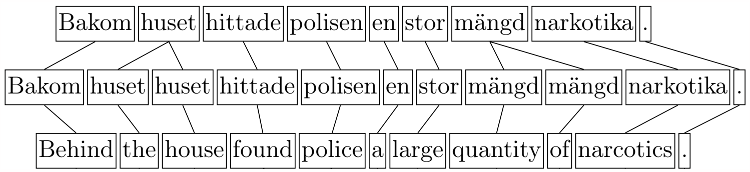


Generative Model of Word-Based SMT



Fertility (and NULL insertion)

Generative Model of Word-Based SMT



Bakom huset hittade polisen en stor mängd narkotika .

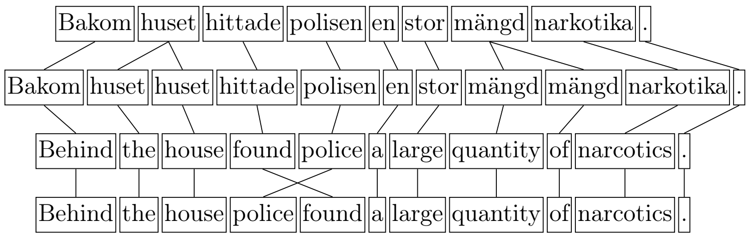
Bakom huset huset hittade polisen en stor mängd mängd narkotika .

Behind the house found police a large quantity of narcotics .

Fertility (and NULL insertion)

Word translation

Generative Model of Word-Based SMT



Bakom huset hittade polisen en stor mängd narkotika .

Bakom huset huset hittade polisen en stor mängd mängd narkotika .

Behind the house found police a large quantity of narcotics .

Behind the house police found a large quantity of narcotics .

Fertility (and NULL insertion)

Word translation

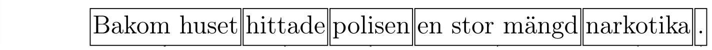
Re-ordering (distortion)

Generative Model of Phrase-Based SMT

Same example sentence:

- Bakom huset hittade polisen en stor mängd narkotika .

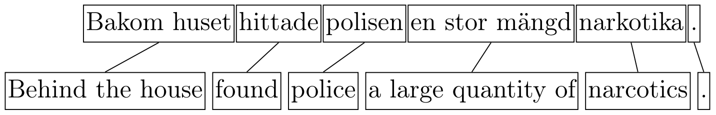
Generative Model of Phrase-Based SMT



Bakom huset hittade polisen en stor mängd narkotika .

Segmentation (into “phrases”)

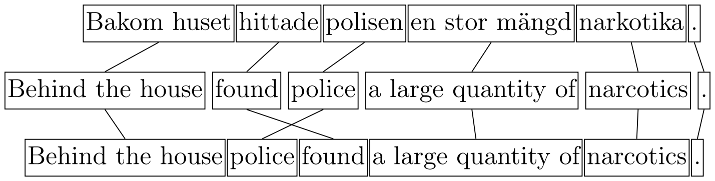
Generative Model of Phrase-Based SMT



Segmentation (into "phrases")

Phrase translation

Generative Model of Phrase-Based SMT

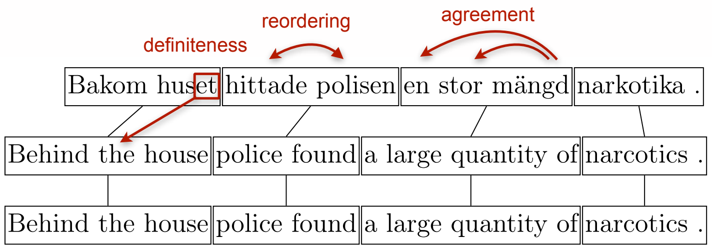


Segmentation (into "phrases")

Phrase translation

Phrase reordering (distortion)

Alternative Segmentation

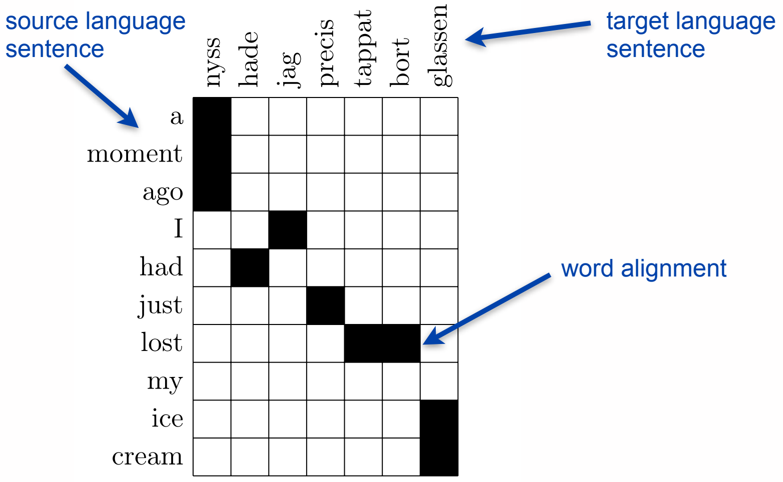


• can handle non-compositional expressions
• lexical disambiguation based on local context

Additional challenge in PB-SMT: Any segmentation is possible!

- in practice: fixed maximum phrase length (typically 7 words)
- fixed distortion limit (typically 6)

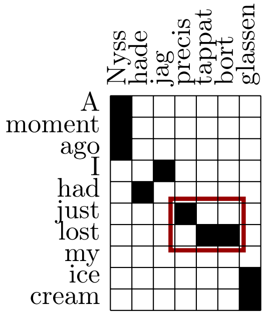
Training Phrase-Based Models



	nyss	hade	jag	precis	tappat	bort	glassen
a	■						
moment	■						
ago	■						
I			■				
had		■					
just				■			
lost					■	■	
my							
ice							■
cream							■

Phrase Extraction

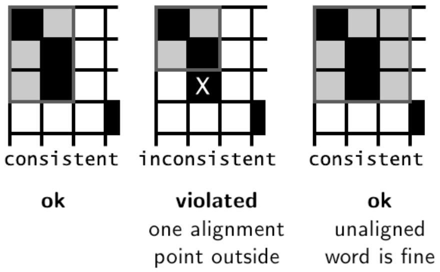
Extract ALL phrase pairs that are **consistent** with underlying **word alignment**



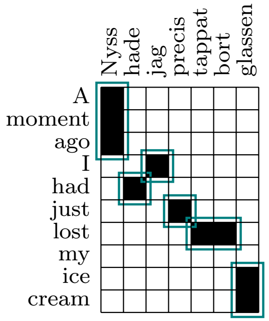
just lost–precis tappat bort

Consistent Phrase Pairs

No word is aligned to any word outside of the phrase pair!

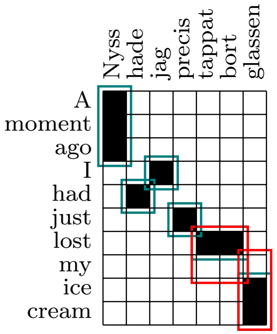


Phrase Extraction



a moment ago–nyss, I–jag, had–hade, just–precis lost–tappat bort, ice cream–glassen

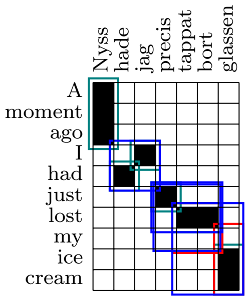
Phrase Extraction



a moment ago–nyss, I–jag, had–hade, just–precis lost–tappat bort, ice cream–glassen

lost my–tappat bort, my ice cream–glassen

Phrase Extraction

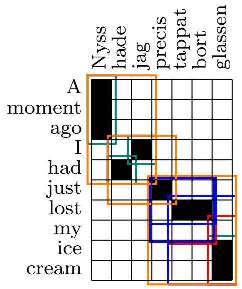


a moment ago–nyss, I–jag, had–hade, just–precis
lost–tappat bort, ice cream–glassen

lost my–tappat bort, my ice cream–glassen

I had–hade jag, lost my ice cream–tappat bort glassen
just lost–precis tappat bort, just lost my–precis tappat bort

Phrase Extraction



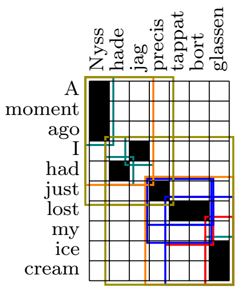
a moment ago–nyss, I–jag, had–hade, just–precis
lost–tappat bort, ice cream–glassen

lost my–tappat bort, my ice cream–glassen

I had–hade jag, lost my ice cream–tappat bort glassen
just lost–precis tappat bort, just lost my–precis tappat bort

a moment ago I had–nyss hade jag, I had just–hade jag precis
just lost my ice cream–precis tappat bort glassen

Phrase Extraction



a moment ago–nyss, I–jag, had–hade, just–precis
lost–tappat bort, ice cream–glassen

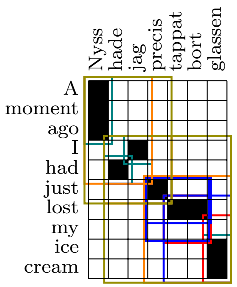
lost my–tappat bort, my ice cream–glassen

I had–hade jag, lost my ice cream–tappat bort glassen
just lost–precis tappat bort, just lost my–precis tappat bort

a moment ago I had–nyss hade jag, I had just–hade jag precis
just lost my ice cream–precis tappat bort glassen

a moment ago I had just–nyss hade jag precis
I had just lost my ice cream–hade jag precis tappat bort glassen

Phrase Extraction



a moment ago–nyss, I–jag, had–hade, just–precis
lost–tappat bort, ice cream–glassen

lost my–tappat bort, my ice cream–glassen

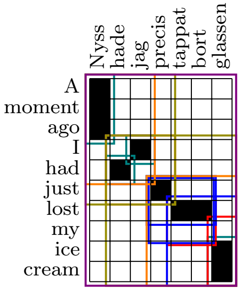
I had–hade jag, lost my ice cream–tappat bort glassen
just lost–precis tappat bort, just lost my–precis tappat bort

a moment ago I had–nyss hade jag, I had just–hade jag precis
just lost my ice cream–precis tappat bort glassen

a moment ago I had just–nyss hade jag precis
I had just lost my ice cream–hade jag precis tappat bort glassen

. . .

Phrase Extraction



a moment ago–nyss, I–jag, had–hade, just–precis
lost–tappat bort, ice cream–glassen

lost my–tappat bort, my ice cream–glassen

I had–hade jag, lost my ice cream–tappat bort glassen
just lost–precis tappat bort, just lost my–precis tappat bort

a moment ago I had–nyss hade jag, I had just–hade jag precis
just lost my ice cream–precis tappat bort glassen

a moment ago I had just–nyss hade jag precis
I had just lost my ice cream–hade jag precis tappat bort glassen

. . .

a moment ago I had just lost my ice cream–
nyss hade jag precis tappat bort glassen

Phrase Translation Probabilities

Extract all phrase pairs

- up to a certain size (typically: max 7 words)
- from all sentence pairs in the training data
- sort them and count

Probability estimation by MLE:

phrase pair count

$$\phi(\bar{t}|\bar{s}) = \frac{count(\bar{s},\bar{t})}{\sum_{\bar{t}_i} count(\bar{s},\bar{t}_i)}$$

count of source phrase
aligned to any other
target language phrase

Phrase Translation Tables

translated
EU speeches

Translations of “**begreppet**” extracted from Europarl

English	$\phi(\bar{t}\|\bar{s})$	English	$\phi(\bar{t}\|\bar{s})$
the	0.226415	the news	0.012816
told	0.169811	the report	0.008544
announcement	0.075472	the information	0.008544
message	0.056604	the back	0.004272
news	0.056604	the suspension	0.004272
information	0.037736	the death	0.004272
informed	0.037736	this announcement	0.002848
learnt	0.037736	this news	0.002136
peace of mind by ensuring	0.027778	a message	0.001539
insight	0.018868	his answer	0.000356
the announcement	0.017088	were told	0.000229
the message	0.012816	the back and	2.917e-05

translation
probability

- lexical variation (announcement, message, news, told, …)
- morphological variation (information, informed)
- including function words and **lots of noise**

Extension: Weighted Models

Components of Phrase-Based SMT

- phrase translation model
- reordering model
- language model

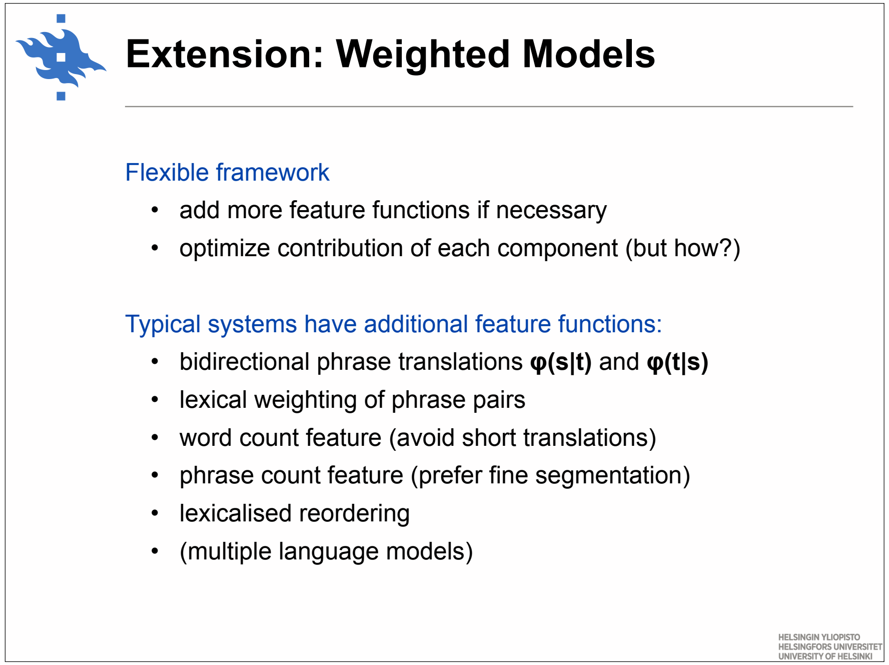
Log-Linear models with weighted feature functions:

- forget about generative models
- give components different weights

feature function
(component)

$$p(x) = \exp \sum_{i=1}^{n} \lambda_i h_i(x)$$

feature-specific
weight



Extension: Weighted Models

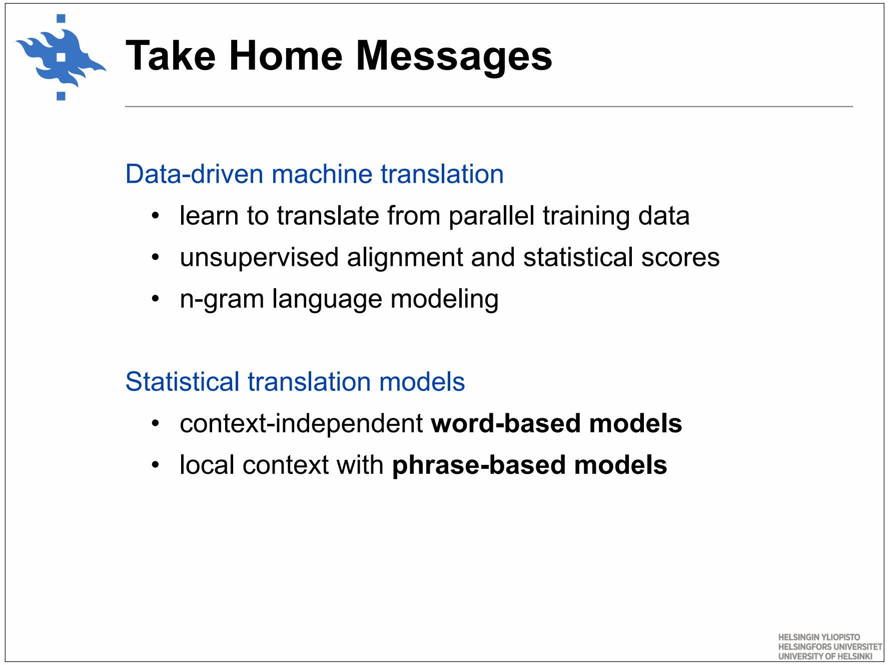
Flexible framework

- add more feature functions if necessary
- optimize contribution of each component (but how?)

Typical systems have additional feature functions:

- bidirectional phrase translations **φ(s|t)** and **φ(t|s)**
- lexical weighting of phrase pairs
- word count feature (avoid short translations)
- phrase count feature (prefer fine segmentation)
- lexicalised reordering
- (multiple language models)

Summing Up



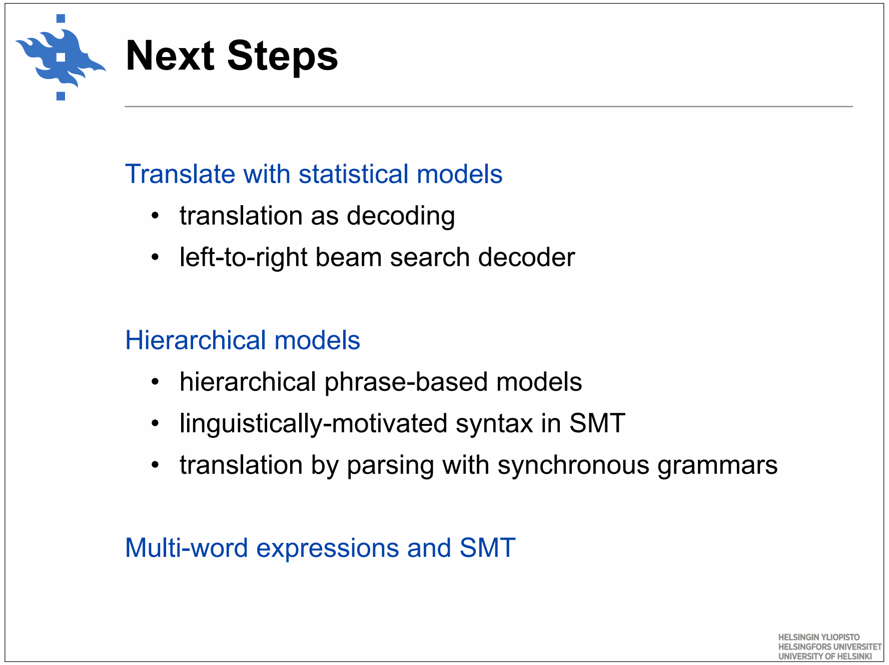
Take Home Messages

Data-driven machine translation

- learn to translate from parallel training data
- unsupervised alignment and statistical scores
- n-gram language modeling

Statistical translation models

- context-independent **word-based models**
- local context with **phrase-based models**



Next Steps

Translate with statistical models

- translation as decoding
- left-to-right beam search decoder

Hierarchical models

- hierarchical phrase-based models
- linguistically-motivated syntax in SMT
- translation by parsing with synchronous grammars

Multi-word expressions and SMT

Acknowledgments

Slides and images from various people

- Kevin Knight
- Philipp Koehn
- Sara Stymne
- Christian Hardmeier