Machine Translation - Decoding

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。

MT

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.

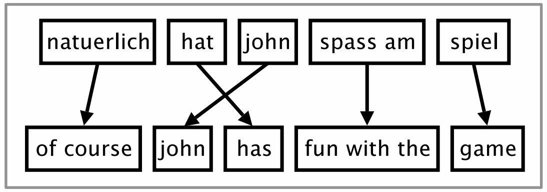
Jörg Tiedemann
University of Helsinki

Fabienne Cap
Uppsala University

Phrase-Based SMT in a Nutshell

Training:

millions of aligned parallel sentences



Translation Model

Reordering Model

billions of target language sentences

of course john has
course john has fun
john has fun with
.....

Target Language Model

translation = decoding: $\mathbf{e}^* = \underset{\mathbf{e}}{argmax}\, P(\mathbf{e}|\mathbf{f})$

Model Size

Phrase tables

- 2 phrase translation probabilities
- 2 lexical translation weights

Language models

- 1 probability per known N-gram
- backoff probabilities, unknown word probabilities

Example: English - French trained on Europarl

- 114 million phrase translations
- 113 million 5-gram probabilities in language model
- 133 million backoff probabilities in language model

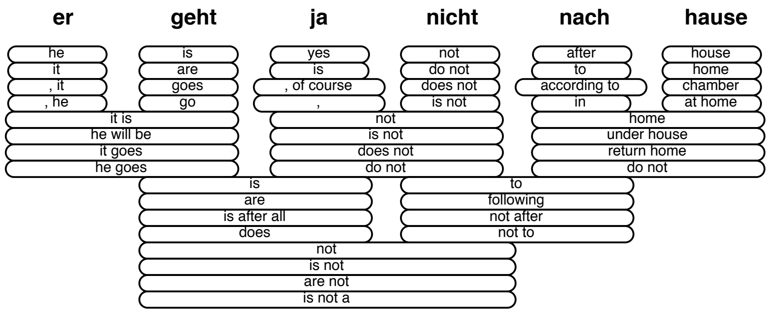
Decoding Complexity

Naively, in a sentence of N words with T translation options for each phrase, we can have

- $O(2^N)$ phrase segmentations,
- $O(T^N)$ sets of phrase translations,
- $O(N!)$ word reordering permutations

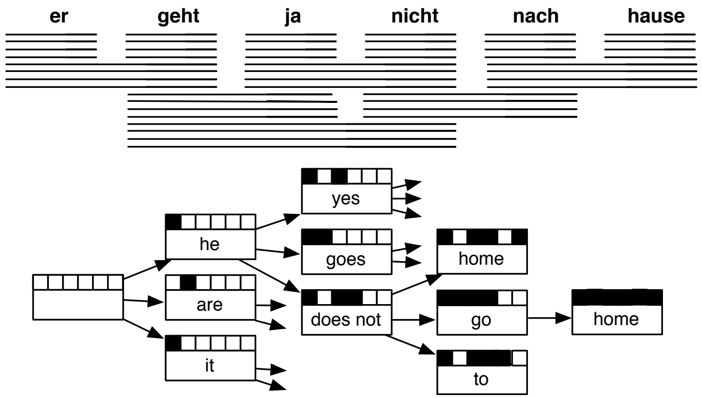
Translation Options

Get (all) translation options for all possible segmentations of the input sentence to be translated!

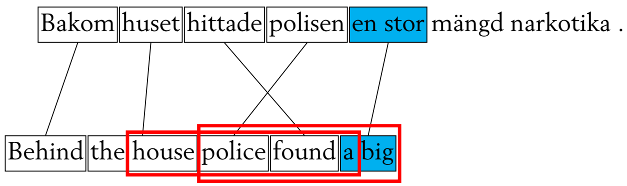


Decoding by Hypothesis Expansion

Using the available translation options
create translation hypothesis from left to right:



Exploiting Model Locality



What we need to score a new hypothesis is

- the score of the previous hypothesis
- the translation model score
- the new language model score

given

context independent

limited window

Choices are independent of everything beyond this window

Hypothesis Recombination

Example:

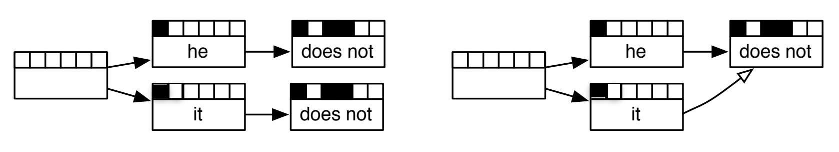
- Three hypotheses with the same coverage
- trigram language model

~~After the house police~~	~~Score = –12.5~~
Behind the house police	Score = –11.2
~~, the house police~~	~~Score = –22.0~~

Competing hypotheses can be discarded because they will never beat the winning one later on!

Hypothesis Recombination

In the search graph: Combine branches

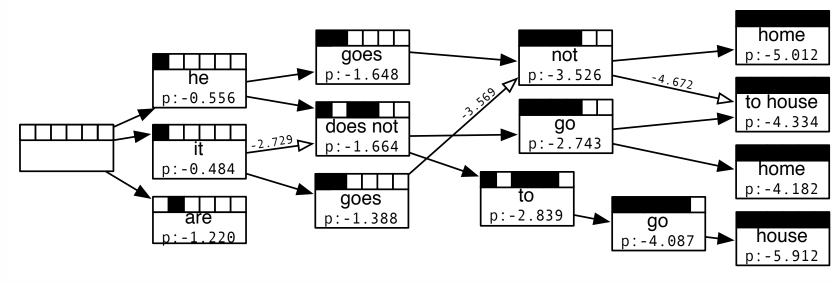


It is a form of dynamic programming

- substantial reduction of the search space
- search is still optimal

Hypothesis Recombination

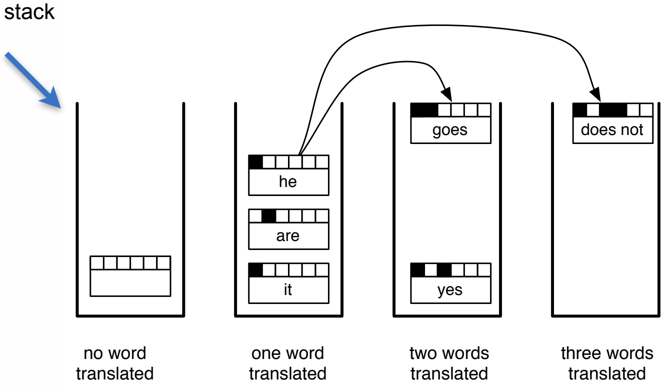
Combine branches greatly reduces the search space



But decoding is still exponential

Stack Decoding

define stack limits



sorted by the number of input words covered by the given hypothesis

Pruning

Histogram Pruning

- keep no more that *n* hypotheses per stack
- *Parameter:* **Stack size *n***

Threshold Pruning

- discard hypotheses with low scores compared to the score of the best hypothesis on the same stack *h**
- *Score(h) < a * Score(h*)*
- *Parameter:* **threshold factor *a***

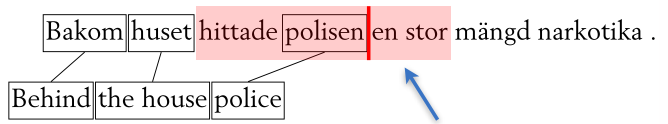
Distortion Limits

Limit reordering reduces search space dramatically

- most partial hypotheses cover the same input
- search complexity: linear in sentence length

Is it OK to do that?

- for closely related languages: most reordering is local
- could do ***pre-ordering*** if necessary



limited distortion window

Take Home Messages

Translation as decoding

- optimise the search problem
- hypothesis expansion and recombination
- pruning and beam search

Links

- Moses decoder: http://www.statmt.org/moses/
- other tools: http://www.statmt.org

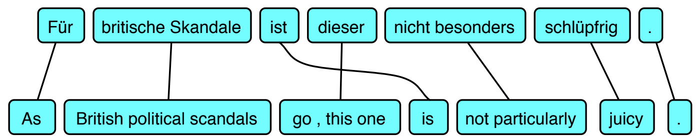
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Hierarchical Models

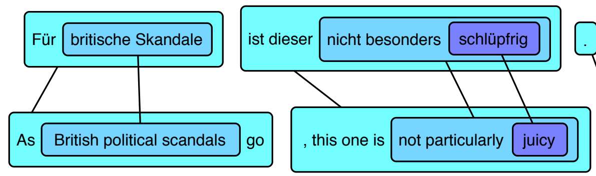
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Hierarchical Phrase-Based Models

Like phrase pairs. . .



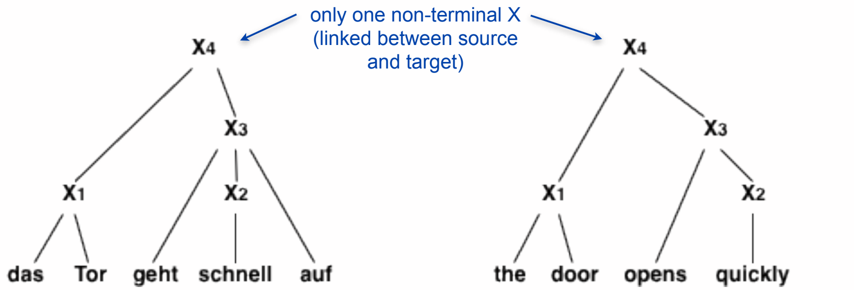
But with nesting:



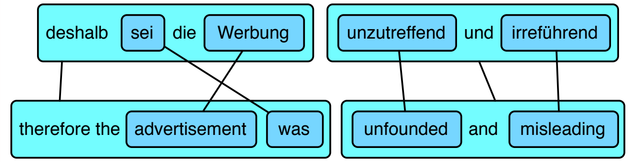
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Hierarchical Phrase-Based Models

standard phrase-based models = one level of hierarchy
HIERO = any kind of tree depth
Represented as Synchronous Context-Free Grammars (SCFGs)



Hierarchical Phrase-Based SMT



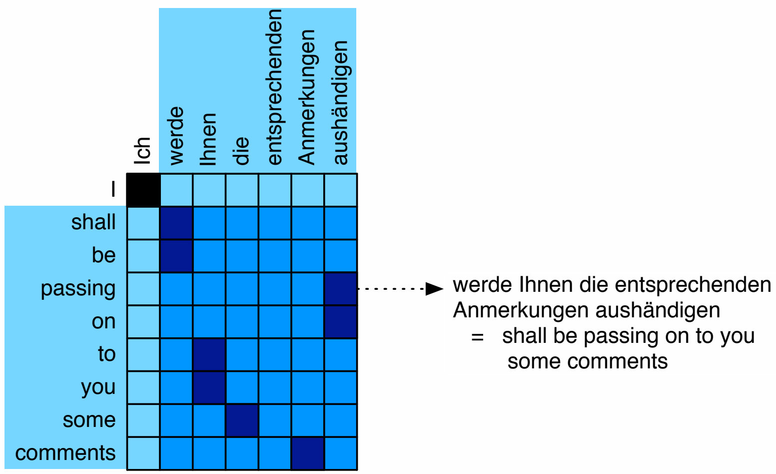
Synchronous re-write rules:

$$\mathrm{X} \rightarrow \textit{deshalb}\ \mathrm{X}_1\ \textit{die}\ \mathrm{X}_2 \mid \textit{therefore the}\ \mathrm{X}_2\ \mathrm{X}_1$$

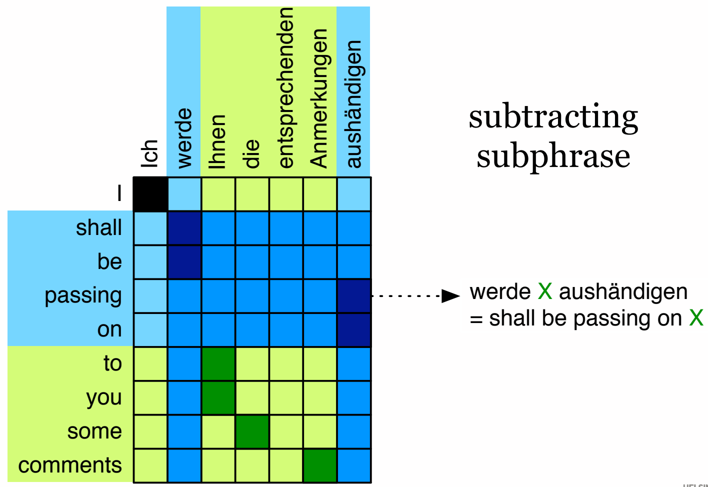
$$\mathrm{X} \rightarrow \mathrm{X}_1\ \textit{und}\ \mathrm{X}_2 \mid \mathrm{X}_1\ \textit{and}\ \mathrm{X}_2$$

Add probabilities derived from statistics for each rule

Hierarchical Phrase Extraction



Hierarchical Phrase Extraction



Linguistically Motivated Syntax

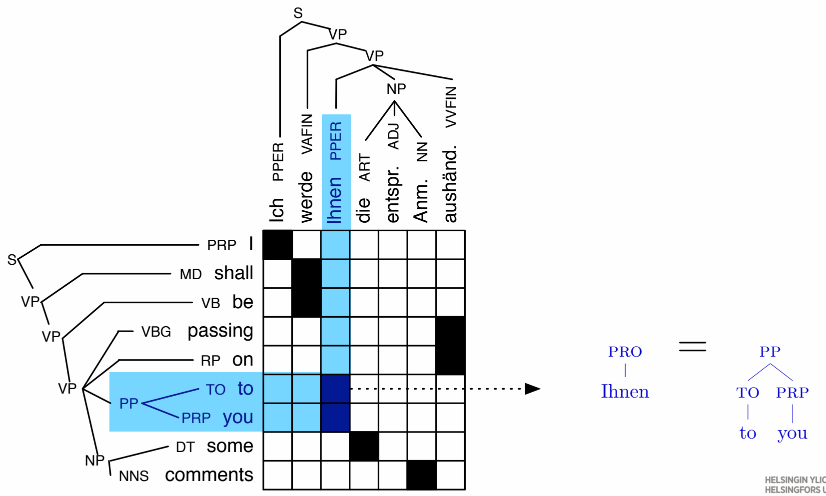
String-to-Tree models

- train on parsed parallel corpora (at least target language)
- extract hierarchical SCFG rules
- translate plain text input

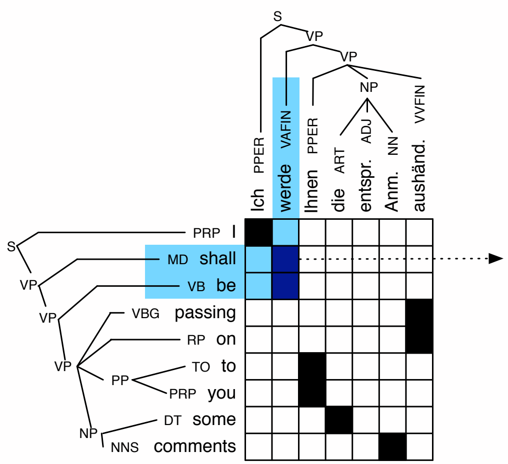
Tree-to-String and Tree-to-Tree models

- train on parsed parallel corpora
- extract synchronous tree-substitution rules (STSGs)
- translate parsed input

Learning Syntactic Rules

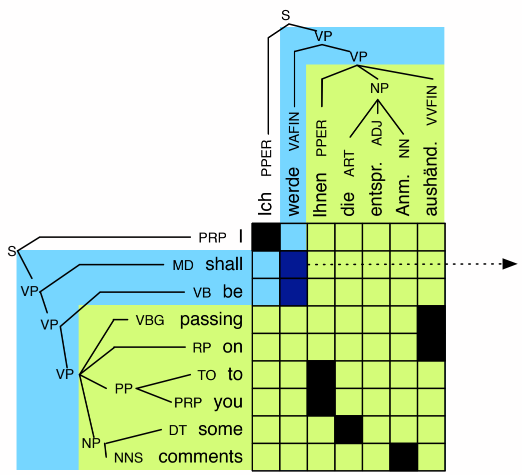


Learning Syntactic Rules

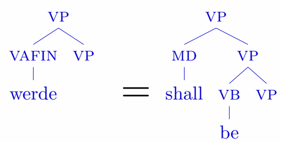


English span not a constituent
no rule extracted

Learning Syntactic Rules



Rule with this phrase pair
requires syntactic context



Parameter Estimation

Extract all rules

- from large aligned (possibly parsed) parallel data
- rule extraction heuristics

Score rules

- count statistics
- maximum-likelihood estimation

Decoding Hierarchical Models

Generating Strings with SCFGs

Input jemand mußte Josef K. verleumdet haben
someone must Josef K. slandered have

Grammar

r_1:	NP	$\rightarrow$	*Josef K.* | *Josef K.*	0.90
r_2:	VBN	$\rightarrow$	*verleumdet* | *slandered*	0.40
r_3:	VBN	$\rightarrow$	*verleumdet* | *defamed*	0.20
r_4:	VP	$\rightarrow$	*mußte* X_1 X_2 *haben* | *must have* VBN_2 NP_1	0.10
r_5:	S	$\rightarrow$	*jemand* X_1 | *someone* VP_1	0.60
r_6:	S	$\rightarrow$	*jemand mußte* X_1 X_2 *haben* | *someone must have* VBN_2 NP_1	0.80
r_7:	S	$\rightarrow$	*jemand mußte* X_1 X_2 *haben* | NP_1 *must have been* VBN_1 *by someone*	0.05

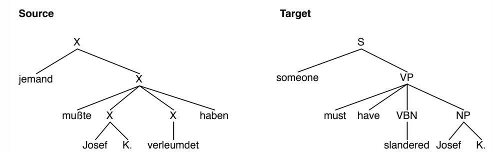
Generating Strings with SCFGs

Input jemand mußte Josef K. verleumdet haben
someone must Josef K. slandered have

Grammar

$\Rightarrow$	r_1:	NP	$\rightarrow$	*Josef K.* | *Josef K.*	0.90
$\Rightarrow$	r_2:	VBN	$\rightarrow$	*verleumdet* | *slandered*	0.40
	r_3:	VBN	$\rightarrow$	*verleumdet* | *defamed*	0.20
$\Rightarrow$	r_4:	VP	$\rightarrow$	*mußte* X_1 X_2 *haben* | *must have* VBN_2 NP_1	0.10
$\Rightarrow$	r_5:	S	$\rightarrow$	*jemand* X_1 | *someone* VP_1	0.60
	r_6:	S	$\rightarrow$	*jemand mußte* X_1 X_2 *haben* | *someone must have* VBN_2 NP_1	0.80
	r_7:	S	$\rightarrow$	*jemand mußte* X_1 X_2 *haben* | NP_1 *must have been* VBN_1 *by someone*	0.05

Derivation 1



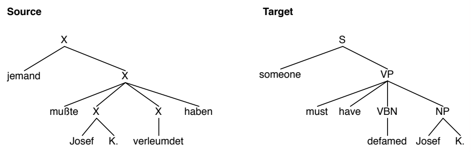
Generating Strings with SCFGs

Input jemand mußte Josef K. verleumdet haben
someone must Josef K. slandered have

Grammar

$\Rightarrow r_1$:	NP	$\rightarrow$	*Josef K.* | *Josef K.*		0.90
r_2:	VBN	$\rightarrow$	*verleumdet* | *slandered*		0.40
$\Rightarrow r_3$:	VBN	$\rightarrow$	*verleumdet* | *defamed*		0.20
$\Rightarrow r_4$:	VP	$\rightarrow$	*mußte* X_1 X_2 *haben* | *must have* VBN_2 NP_1		0.10
$\Rightarrow r_5$:	S	$\rightarrow$	*jemand* X_1 | *someone* VP_1		0.60
r_6:	S	$\rightarrow$	*jemand mußte* X_1 X_2 *haben* | *someone must have* VBN_2 NP_1		0.80
r_7:	S	$\rightarrow$	*jemand mußte* X_1 X_2 *haben* | NP_1 *must have been* VBN_1 *by someone*		0.05

Derivation 2



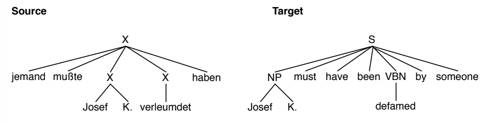
Generating Strings with SCFGs

Input jemand mußte Josef K. verleumdet haben
someone must Josef K. slandered have

Grammar

$\Rightarrow r_1$:	NP	$\rightarrow$	*Josef K.* | *Josef K.*		0.90
r_2:	VBN	$\rightarrow$	*verleumdet* | *slandered*		0.40
$\Rightarrow r_3$:	VBN	$\rightarrow$	*verleumdet* | *defamed*		0.20
r_4:	VP	$\rightarrow$	*mußte* X_1 X_2 *haben* | *must have* VBN_2 NP_1		0.10
r_5:	S	$\rightarrow$	*jemand* X_1 | *someone* VP_1		0.60
r_6:	S	$\rightarrow$	*jemand mußte* X_1 X_2 *haben* | *someone must have* VBN_2 NP_1		0.80
$\Rightarrow r_7$:	S	$\rightarrow$	*jemand mußte* X_1 X_2 *haben* | NP_1 *must have been* VBN_1 *by someone*		0.05

Derivation 6



Translating with SCFGs

Objective Find the highest-scoring synchronous derivation d^*

Solution

1. Project grammar
   Project weighted SCFG to weighted CFG
   $f : G \rightarrow G'$ (many-to-one rule mapping)

2. Parse
   Find Viterbi parse of sentence wrt G'

3. Translate
   Produce synchronous tree pair by applying inverse projection f'

Translating = Parsing

Use synchronous CFG like monolingual CFG

$$G' \quad \begin{array}{rrcl} q_1: & \mathrm{NP} & \rightarrow & \textit{Josef K.} \\ q_2: & \mathrm{VBN} & \rightarrow & \textit{verleumdet} \\ q_3: & \mathrm{VP} & \rightarrow & \textit{mußte}\ \mathrm{NP}\ \mathrm{VBN}\ \textit{haben} \\ q_4: & \mathrm{S} & \rightarrow & \textit{jemand}\ \mathrm{VP} \\ q_5: & \mathrm{S} & \rightarrow & \textit{jemand mußte}\ \mathrm{NP}\ \mathrm{VBN}\ \textit{haben} \end{array}$$

Use standard algorithms for probabilistic parsing

- CKY, CKY+, Earley

Keep track of target side of applied rules
or reconstruct synchronous derivation

MT Evaluation

What Is The Problem?

A typical example from the 2001 NIST evaluation set:

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Evaluation Metrics

Subjective judgements by human evaluators

- translation quality
- grammaticality and style
- inter-annotator agreement

Automatic evaluation metrics

- based on reference translations
- linguistic resources to account for natural variation

Task-based evaluation, e.g.

- estimate post-editing effort
- information preservation for cross-lingual IR

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Adequacy and Fluency

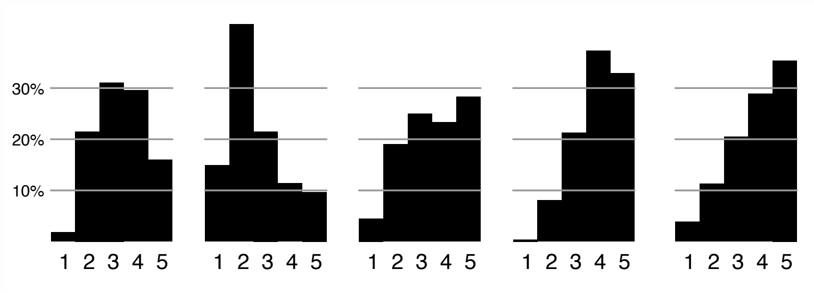
Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

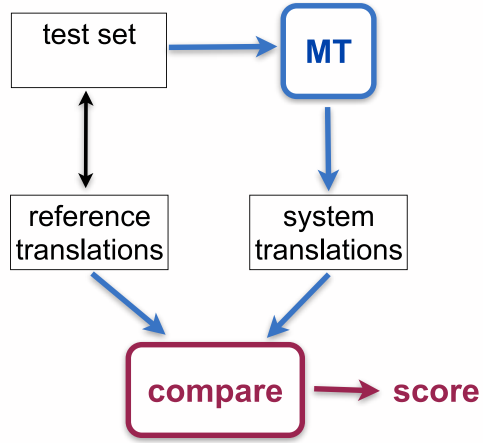
Evaluators Disagree

Histogram of adequacy judgments by different evaluators:



(from WMT 2006 evaluation)

Automatic Evaluation



Word Error Rate

reference translation

System A

		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
securit	7	6	5	5	5	4	4

System B

		airport	security	Israeli	officials	are	responsible
	0	1	2	3	4	5	6
Israeli	1	1	2	2	3	4	5
officials	2	2	2	3	2	3	4
are	3	3	3	3	3	2	3
responsible	4	4	4	4	4	3	2
for	5	5	5	5	5	4	3
airport	6	5	6	6	6	5	4
securit	7	6	5	6	7	6	5

Metric	System A	System B
word error rate (WER)	57%	71%

BLEU (Bilingual Evaluation Understudy)

N-gram overlap between MT output and reference translation
Geometric mean of n-gram precisions (typically 1 to 4)

$$P = \sqrt[n]{precision_1 * precision_2 * \ldots * precision_n}$$
$$= (precision_1 * precision_2 * \ldots * precision_n)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} precision_i\right)^{\frac{1}{n}}$$

Additional brevity penalty for short translation (recall)

$$BP \begin{cases} 1 & \text{if output-length } c > \text{reference-length } r \\ exp(1 - r/c) & \text{if output-length } c \leq \text{reference-length } r \end{cases}$$

Example

SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible
2-GRAM MATCH 4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

Multiple Reference Translations

Account for variability: Use multiple reference translations

- N-grams may match in any of the reference
- closest reference length is used for brevity penalties

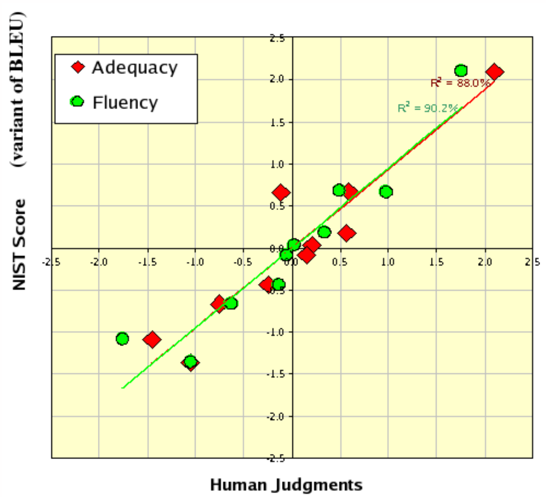
Example:

SYSTEM: Israeli officials responsibility of airport safety
2-GRAM MATCH 2-GRAM MATCH 1-GRAM

REFERENCES:
Israeli officials are responsible for airport security
Israel is in charge of the security at this airport
The security work for this airport is the responsibility of the Israel government
Israeli side was in charge of the security of this airport

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Correlation with Human Judgement



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Typical BLEU Scores

BLEU scores for 110 SMT systems (Koehn 2005)

%	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	-

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Take Home Messages

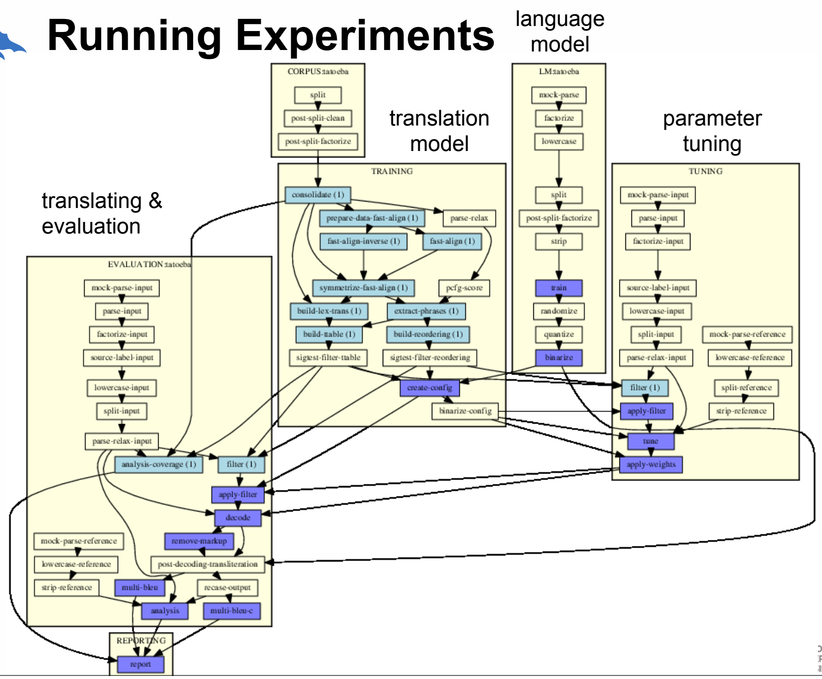
Manual evaluation
- adequacy and fluency
- difficult and expensive

Automatic Evaluation
- comparison to human reference translations
- fast, reusable but not always reliable

Links
- WMT evaluation campaigns: http://www.statmt.org/wmt16/
- IWSLT (spoken MT): http://workshop2016.iwslt.org

Running Experiments

Running Experiments



Next Sessions

MWEs and SMT
- handle MWEs in machine translation
- find MWEs in parallel data sets

Train and use your own SMT model
- language modeling
- word alignment
- translation modeling
- translating test sets and evaluate

Acknowledgments

Slides and images from various people

- Philipp Koehn
- Philip Williams
- Sara Stymne
- Christian Hardmeier
- David Chiang
- Kevin Knight