THE UNIVERSITY *of* EDINBURGH

Machine Translation

12: (Non-neural) Statistical Machine Translation

Rico Sennrich

University of Edinburgh

Today's Lecture

- So far, main focus of lecture was on:
  - neural machine translation
  - research since $\approx$2013
- today, we look at (non-neural) Statistical Machine Translation, and research since $\approx$ 1990

Refresher: A probabilistic model of translation

- Suppose that we have:
  - a source sentence S of length m $(x_1, \ldots, x_m)$
  - a target sentence T of length n $(y_1, \ldots, y_n)$
- We can express translation as a probabilistic model:

$$\begin{aligned} T^* &= \arg\max_T P(T|S) \\ &= \arg\max_T P(S|T)P(T) \end{aligned}$$

Bayes' theorem

- We can model translation via two models:
  - language model to estimate $P(T)$
  - translation model to estimate $P(S|T)$
- Without continuous space representations, how to estimate $P(S|T)$?
  $\rightarrow$ break it up into smaller units

Word Alignment

chicken-and-egg problem

let's break up $P(S|T)$ into small units (words):

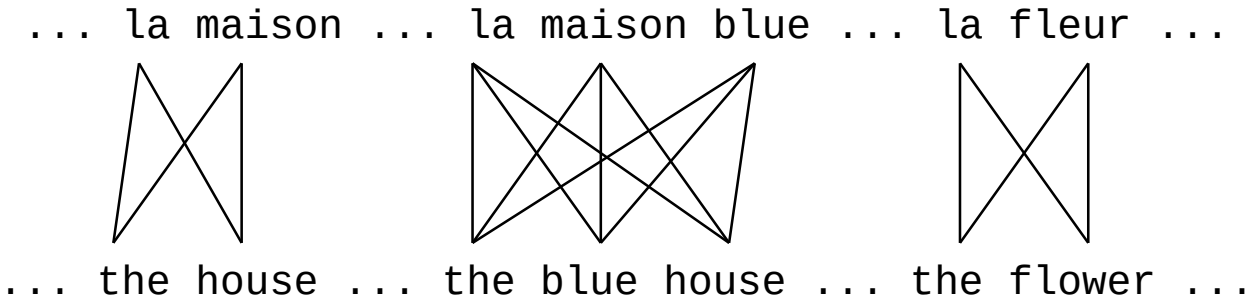
- we can estimate an alignment given a translation model
  expectation step
- we can estimate translation model given a an alignment
  (using relative frequencies)
  maximization step
- what can we do if we have neither?

solution: **Expectation Maximization Algorithm**

- initialize model
- iterate between estimating alignment and translation model

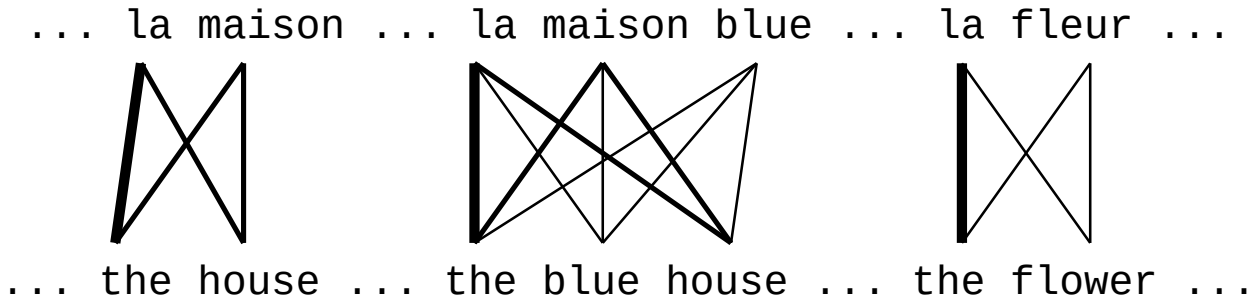
simplest model based on lexical translation; more complex models consider position and fertility

Word Alignment: IBM Models [Brown et al., 1993]



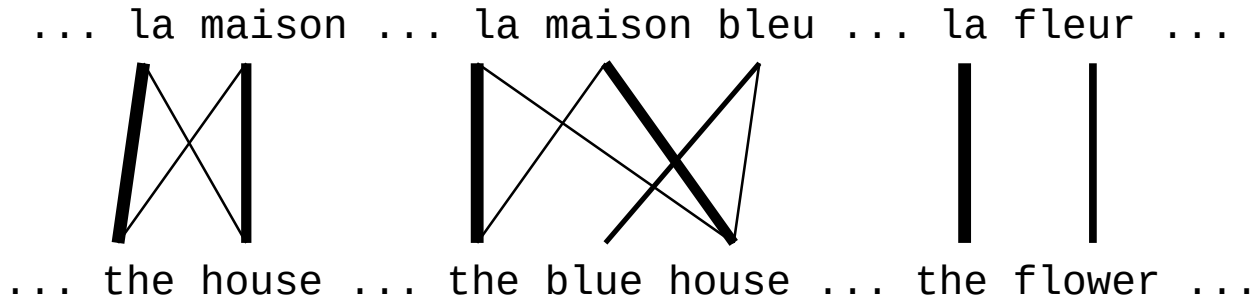
- Initial step: all alignments equally likely
- Model learns that, e.g., la is often aligned with the

Word Alignment: IBM Models [Brown et al., 1993]



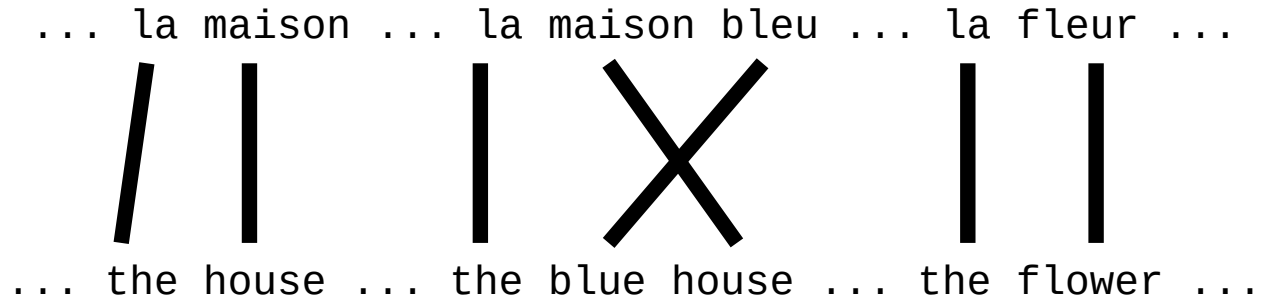
- After one iteration
- Alignments, e.g., between la and the are more likely

Word Alignment: IBM Models [Brown et al., 1993]



- After another iteration
- It becomes apparent that alignments, e.g., between fleur and flower are more likely (pigeon hole principle)

Word Alignment: IBM Models [Brown et al., 1993]



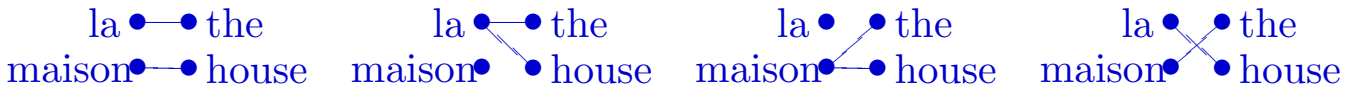
- Convergence
- Inherent hidden structure revealed by EM

Word Alignment: IBM Models [Brown et al., 1993]

- **Probabilities**

$$p(\text{the}|\text{la}) = 0.7 \quad p(\text{house}|\text{la}) = 0.05$$
$$p(\text{the}|\text{maison}) = 0.1 \quad p(\text{house}|\text{maison}) = 0.8$$

- **Alignments**



$$p(\mathbf{e}, a|\mathbf{f}) = 0.56 \quad p(\mathbf{e}, a|\mathbf{f}) = 0.035 \quad p(\mathbf{e}, a|\mathbf{f}) = 0.08 \quad p(\mathbf{e}, a|\mathbf{f}) = 0.005$$

$$p(a|\mathbf{e}, \mathbf{f}) = 0.824 \quad p(a|\mathbf{e}, \mathbf{f}) = 0.052 \quad p(a|\mathbf{e}, \mathbf{f}) = 0.118 \quad p(a|\mathbf{e}, \mathbf{f}) = 0.007$$

- **Counts**

$$c(\text{the}|\text{la}) = 0.824 + 0.052 \quad c(\text{house}|\text{la}) = 0.052 + 0.007$$
$$c(\text{the}|\text{maison}) = 0.118 + 0.007 \quad c(\text{house}|\text{maison}) = 0.824 + 0.118$$

Linear Models

$$T^* = \arg\max_T P(S|T)P(T) \quad \text{Bayes' theorem}$$

$$T^* \approx \arg\max_T \sum_{m=1}^{M} \lambda_m h_m(S,T) \quad \text{[Och, 2003]}$$

- linear combination of arbitrary features
- Minimum Error Rate Training to optimize feature weights
- big trend in SMT research: engineering new/better features

Word-based SMT

core idea

combine word-based translation model and n-gram language model to compute score of translation

consequences

- + models are easy to compute
- - word translations are assumed to be independent of each other: only LM takes into account context
- - poor at modelling long-distance phenomena: n-gram context is limited

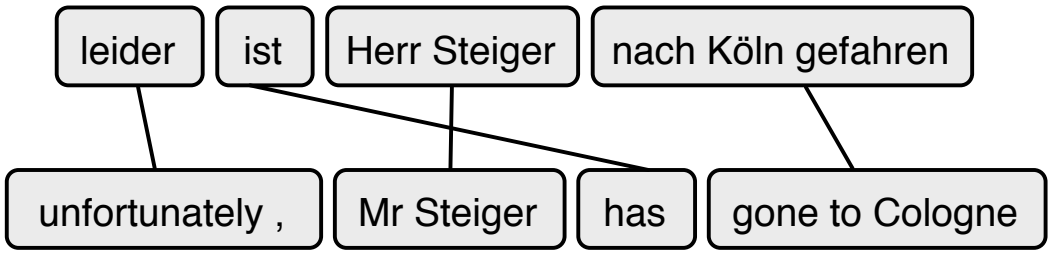
Phrase-based SMT

core idea

Basic translation unit in translation model is not word, but word sequence (phrase)

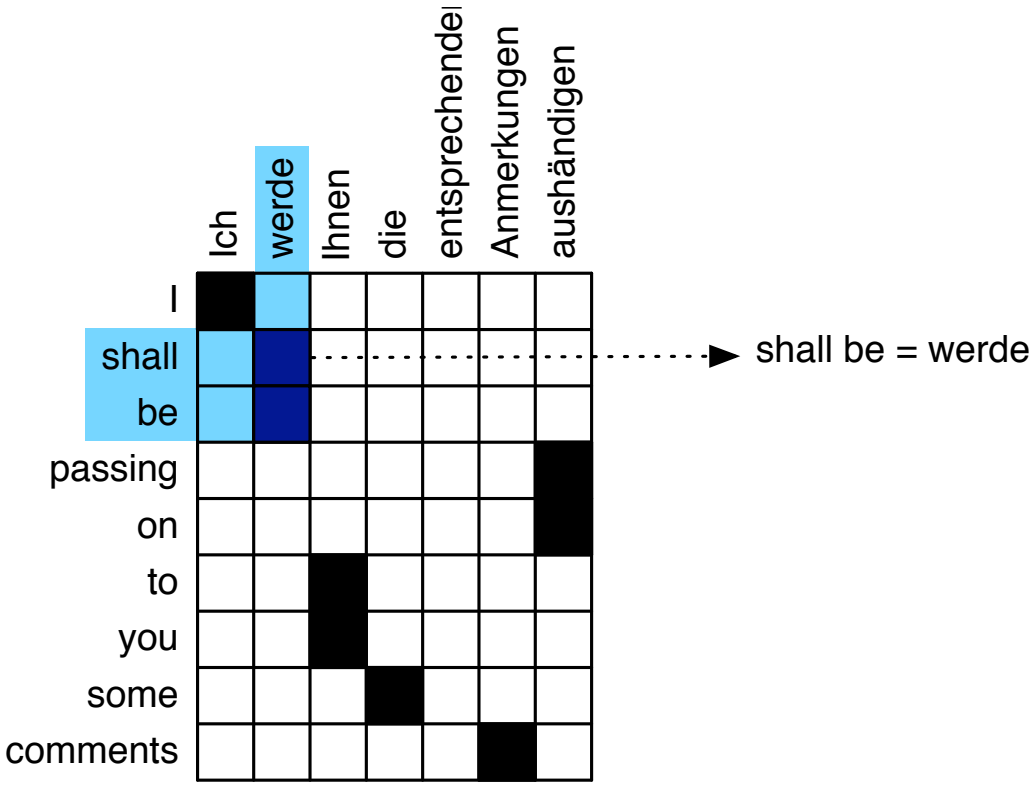
consequences

- + much better memorization of frequent phrase translations
- - large (and noisy) phrase table
- - large search space; requires sophisticated pruning
- - still poor at modelling long-distance phenomena



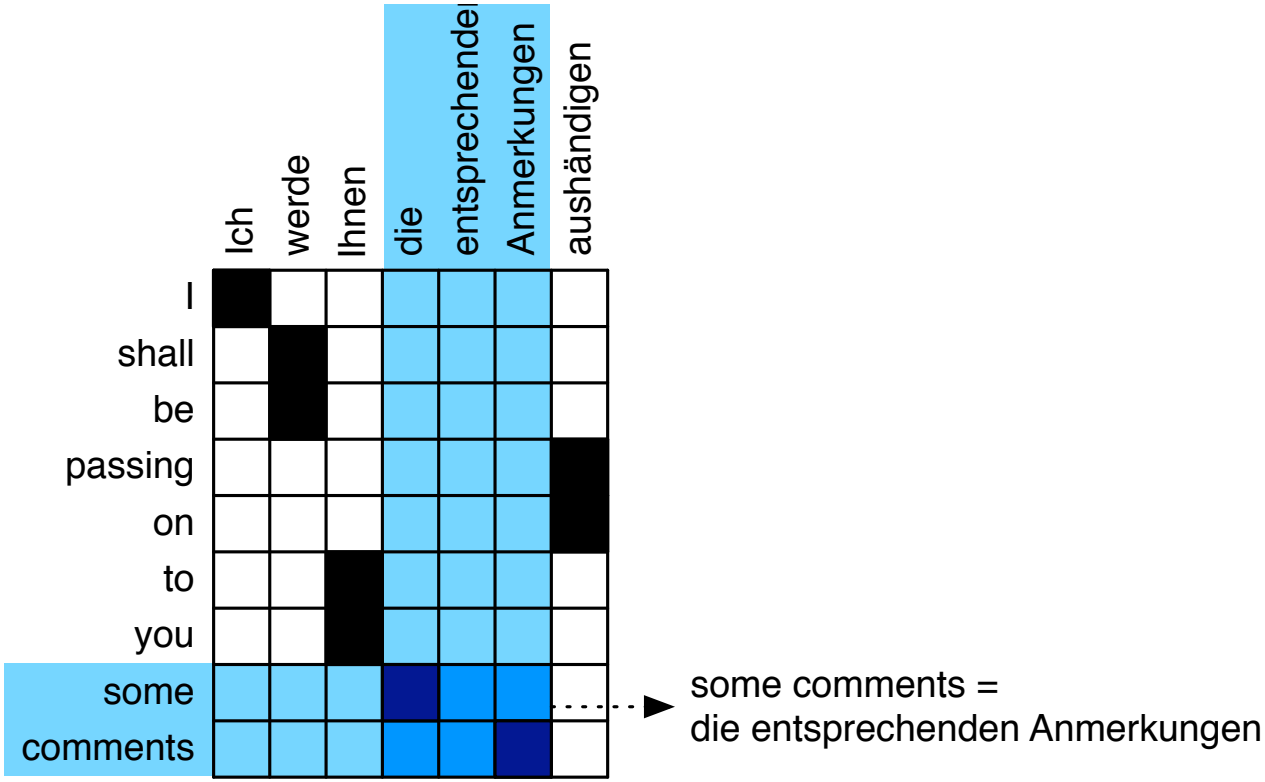
Phrase Extraction

- extraction rules based on word-aligned sentence pair
- phrase pair must be compatible with alignment...
- ...but unaligned words are ok
- phrases are contiguous sequences



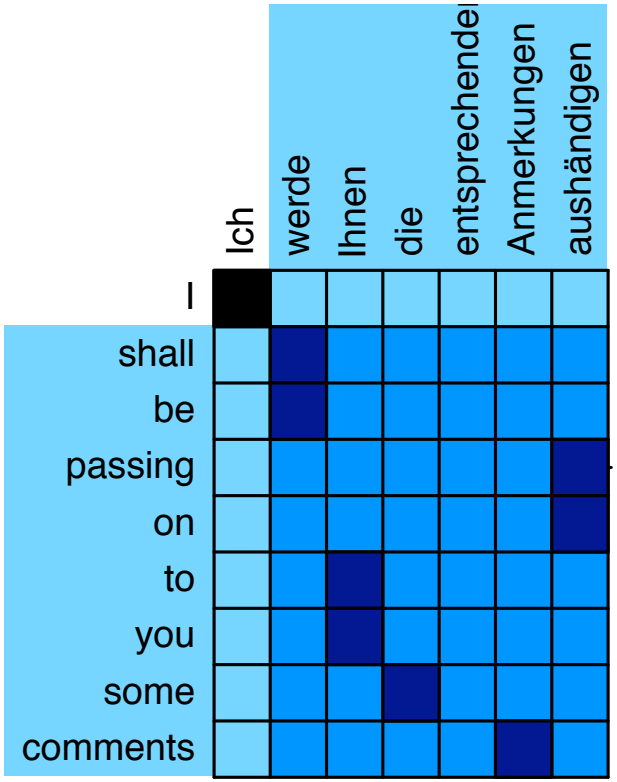
Phrase Extraction

- extraction rules based on word-aligned sentence pair
- phrase pair must be compatible with alignment...
- ...but unaligned words are ok
- phrases are contiguous sequences



Phrase Extraction

- extraction rules based on word-aligned sentence pair
- phrase pair must be compatible with alignment...
- ...but unaligned words are ok
- phrases are contiguous sequences

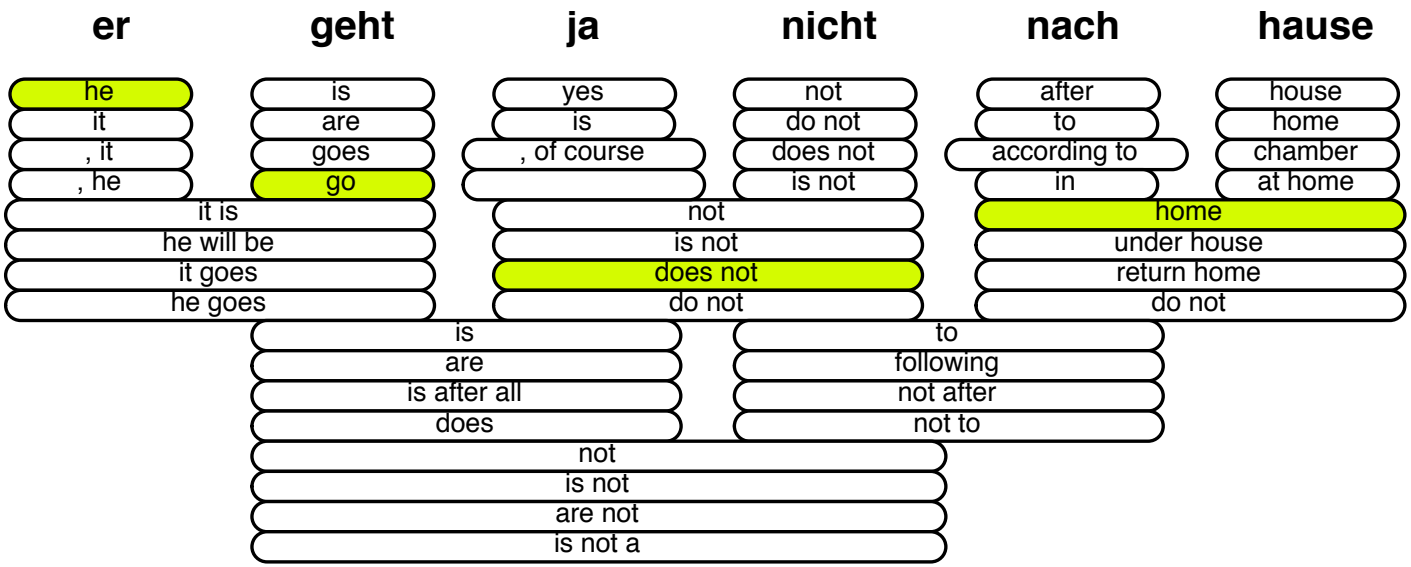


werde Ihnen die entsprechenden Anmerkungen aushändigen
= shall be passing on to you some comments

Common Features in Phrase-based SMT

- phrase translation probabilities (in both directions)
- word translation probabilities (in both directions)
- language model
- reordering model
- constant penalty for each phrase used
- sparse features with learned cost for some (classes of) phrase pairs
- multiple models of each type possible

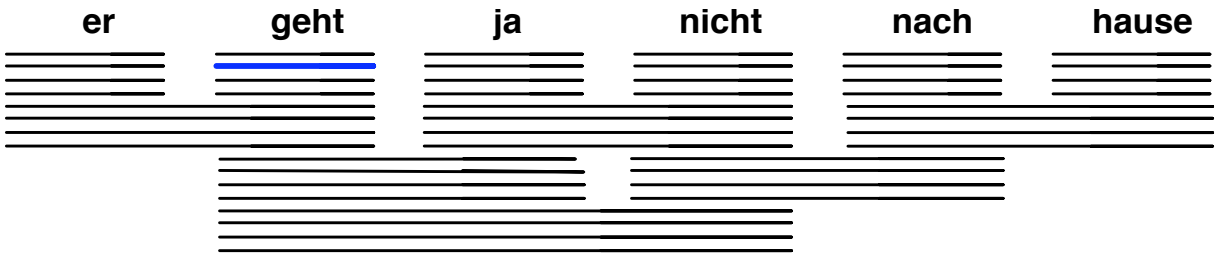
Decoding

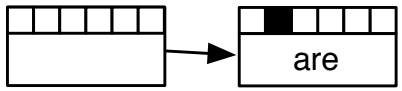


- The machine translation decoder does not know the right answer
  - picking the right translation options
  - arranging them in the right order

→ Search problem solved by heuristic beam search

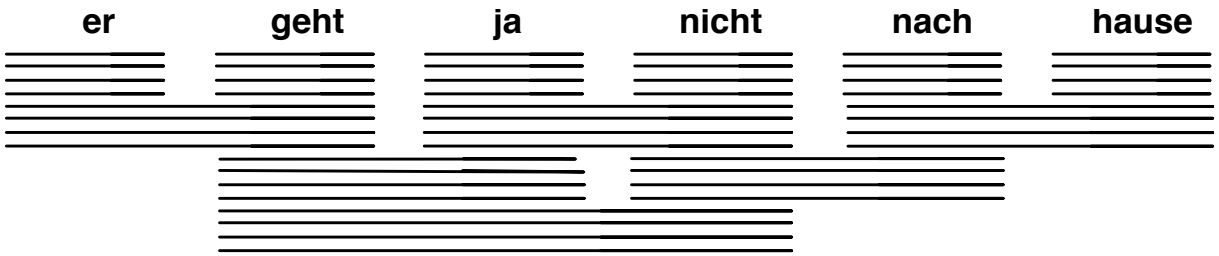
Decoding

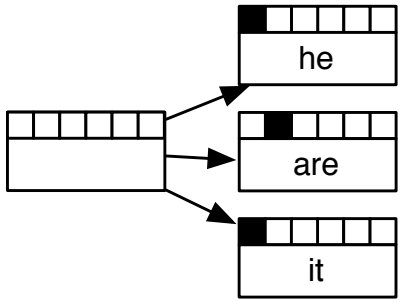




pick any translation option, create new hypothesis

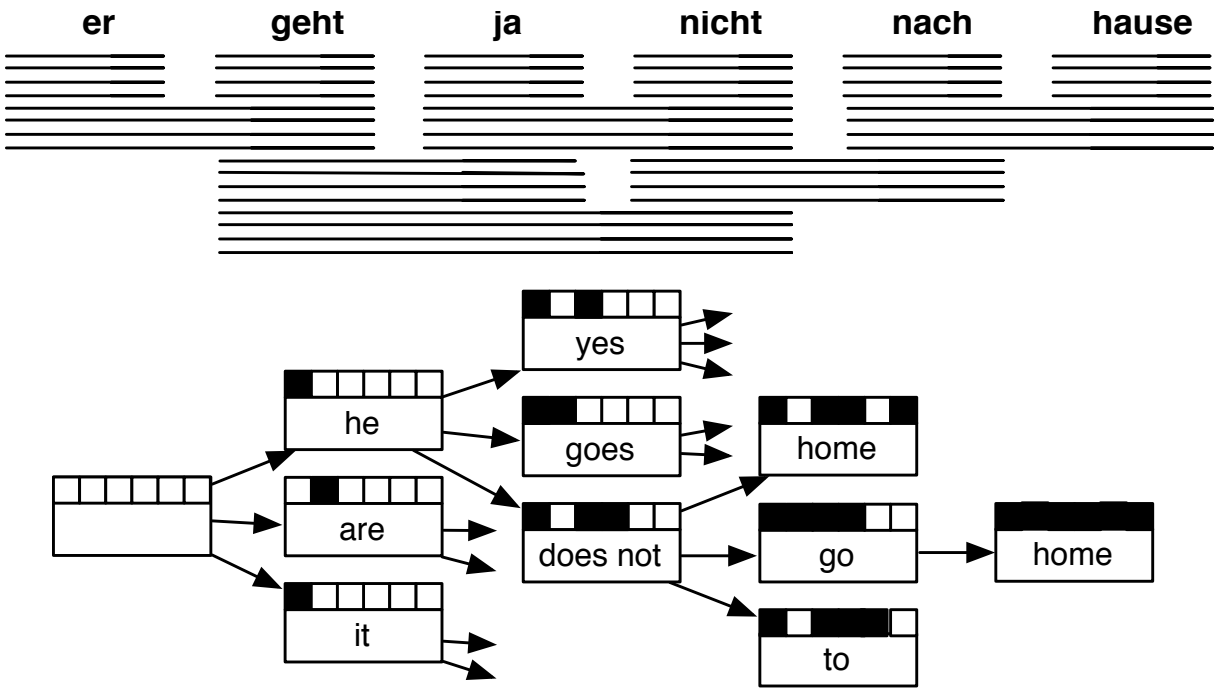
Decoding





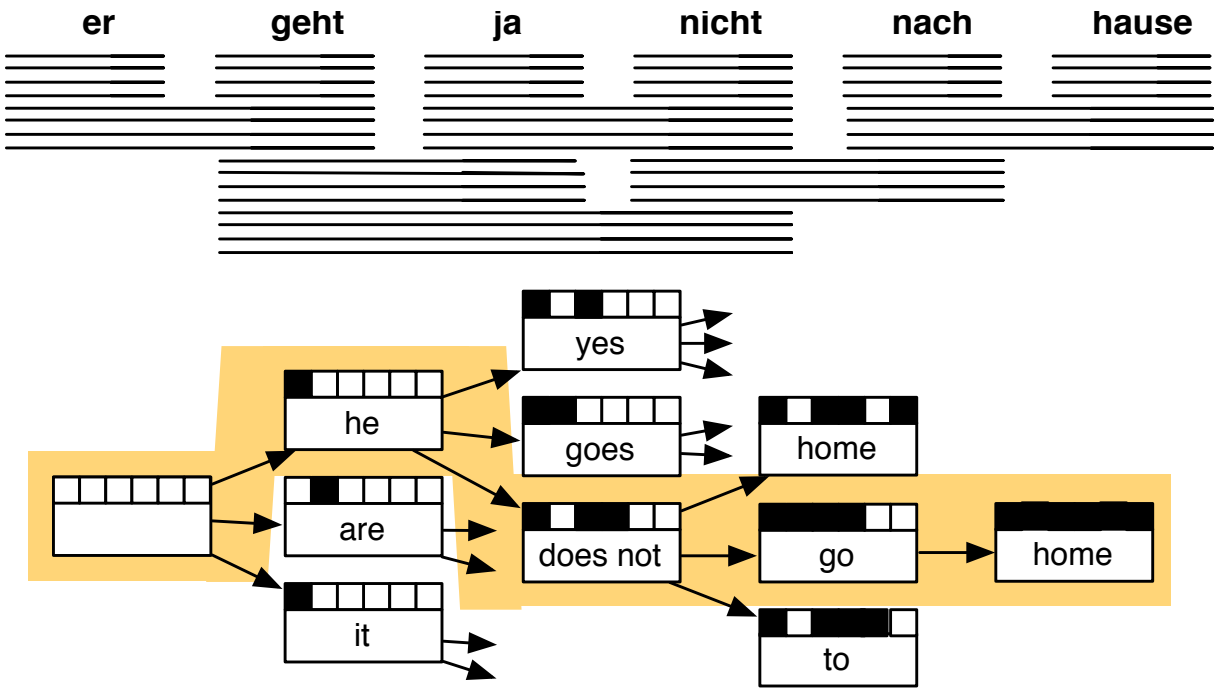
create hypotheses for all other translation options

Decoding



also create hypotheses from created partial hypothesis

Decoding



backtrack from highest scoring complete hypothesis

Decoding

- large search space (exponential number of hypotheses)
- reduction of search space:
  - recombination of identical hypotheses
  - pruning of hypotheses
- efficient decoding is a lot more complex in SMT than in neural MT

1. Statistical Machine Translation
   - Basics
   - Phrase-based SMT
   - Hierarchical SMT
   - Syntax-based SMT

Hierarchical SMT

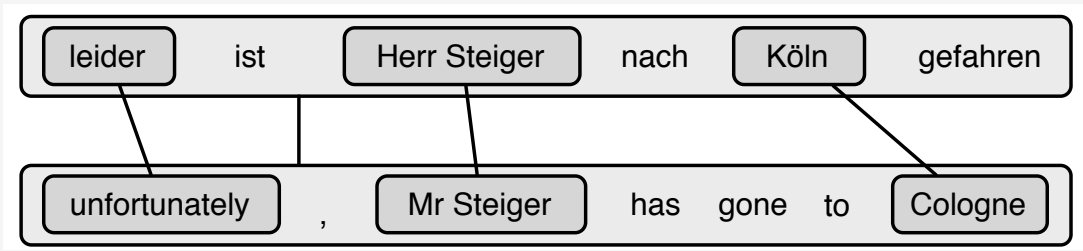
core idea

use context-free grammars (CFG) rules as basic translation units

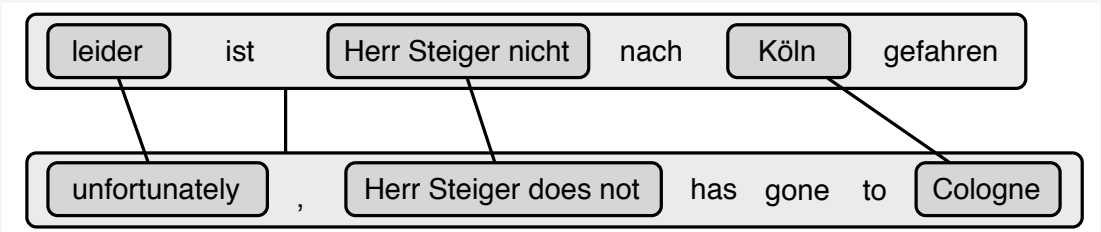
$\rightarrow$ allows gaps

consequences

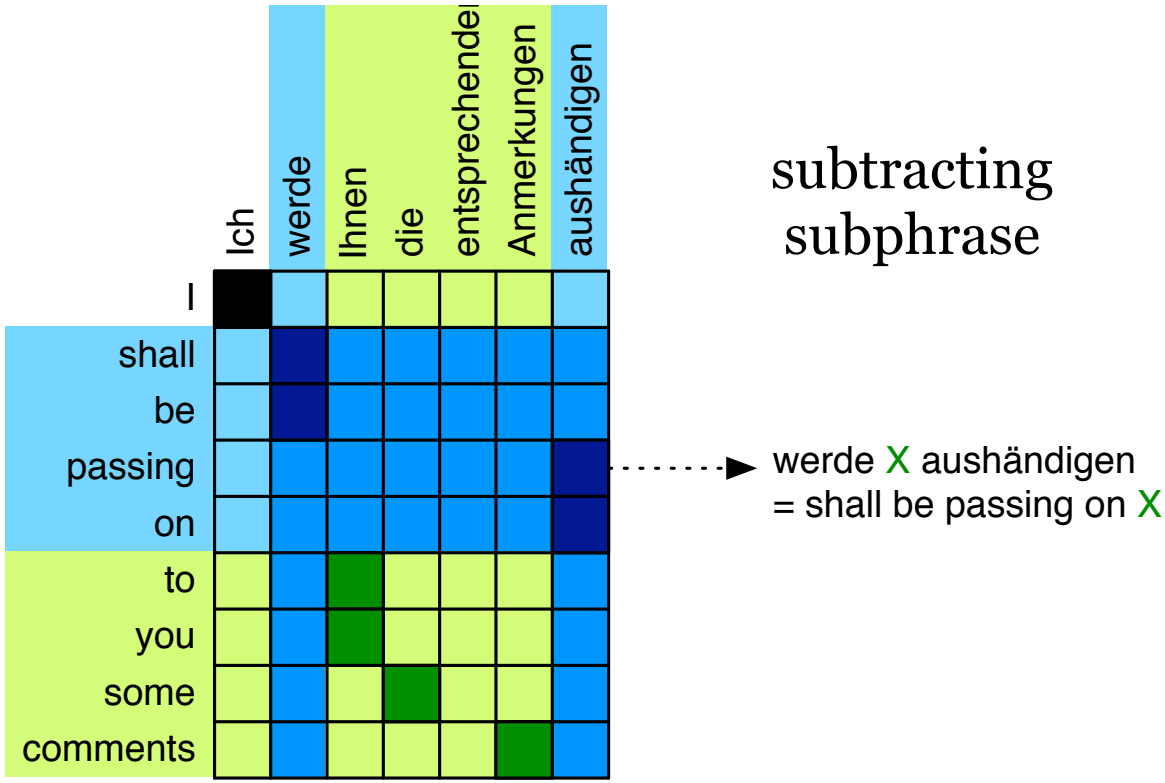
- + better modeling of some reordering patterns



- - overgeneralisation is still possible



Hierarchical Phrase Extraction



Decoding

Decoding via (S)CFG derivation

- Derivation starts with pair of linked S symbols.

Decoding

Decoding via (S)CFG derivation

$$\Rightarrow \quad \mathrm{S}_2\ \mathrm{X}_3 \quad | \quad \mathrm{S}_2\ \mathrm{X}_3$$

- $\mathrm{S} \rightarrow \mathrm{S}_1\ \mathrm{X}_2 \quad | \quad \mathrm{S}_1\ \mathrm{X}_2$ (glue rule)

Decoding

Decoding via (S)CFG derivation

$\Rightarrow \quad \mathrm{S}_2\ \mathrm{X}_3 \quad | \quad \mathrm{S}_2\ \mathrm{X}_3$

$\Rightarrow \quad \mathrm{S}_2\ \mathrm{X}_4\ \mathit{und}\ \mathrm{X}_5 \quad | \quad \mathrm{S}_2\ \mathrm{X}_4\ \mathit{and}\ \mathrm{X}_5$

- $\mathrm{X} \rightarrow \mathrm{X}_1\ \mathit{und}\ \mathrm{X}_2 \quad | \quad \mathrm{X}_1\ \mathit{and}\ \mathrm{X}_2$

Decoding via (S)CFG derivation

$$\Rightarrow \quad \mathrm{S}_2\ \mathrm{X}_3 \quad | \quad \mathrm{S}_2\ \mathrm{X}_3$$

$$\Rightarrow \quad \mathrm{S}_2\ \mathrm{X}_4\ \mathit{und}\ \mathrm{X}_5 \quad | \quad \mathrm{S}_2\ \mathrm{X}_4\ \mathit{and}\ \mathrm{X}_5$$

$$\Rightarrow \quad \mathrm{S}_2\ \mathit{unzutreffend}\ \mathit{und}\ \mathrm{X}_5 \quad | \quad \mathrm{S}_2\ \mathit{unfounded}\ \mathit{and}\ \mathrm{X}_5$$

- $\mathrm{X} \rightarrow \mathit{unzutreffend} \quad | \quad \mathit{unfounded}$

Decoding via (S)CFG derivation

$\Rightarrow$ $\text{S}_2\ \text{X}_3$ | $\text{S}_2\ \text{X}_3$

$\Rightarrow$ $\text{S}_2\ \text{X}_4\ \textit{und}\ \text{X}_5$ | $\text{S}_2\ \text{X}_4\ \textit{and}\ \text{X}_5$

$\Rightarrow$ $\text{S}_2\ \textit{unzutreffend}\ \textit{und}\ \text{X}_5$ | $\text{S}_2\ \textit{unfounded}\ \textit{and}\ \text{X}_5$

$\Rightarrow$ $\text{S}_2\ \textit{unzutreffend}\ \textit{und}\ \textit{irreführend}$ | $\text{S}_2\ \textit{unfounded}\ \textit{and}\ \textit{misleading}$

- $\text{X} \rightarrow \textit{irreführend}$ | $\textit{misleading}$

Decoding via (S)CFG derivation

$\Rightarrow$ $\text{S}_2\ \text{X}_3 \mid \text{S}_2\ \text{X}_3$

$\Rightarrow$ $\text{S}_2\ \text{X}_4\ \textit{und}\ \text{X}_5 \mid \text{S}_2\ \text{X}_4\ \textit{and}\ \text{X}_5$

$\Rightarrow$ $\text{S}_2\ \textit{unzutreffend und}\ \text{X}_5 \mid \text{S}_2\ \textit{unfounded and}\ \text{X}_5$

$\Rightarrow$ $\text{S}_2\ \textit{unzutreffend und irreführend} \mid \text{S}_2\ \textit{unfounded and misleading}$

$\Rightarrow$ $\text{X}_6\ \textit{unzutreffend und irreführend} \mid \text{X}_6\ \textit{unfounded and misleading}$

- $\text{S} \rightarrow \text{X}_1 \mid \text{X}_1$ (glue rule)

Decoding

Decoding via (S)CFG derivation

$\Rightarrow$ $\text{S}_2\ \text{X}_3 \mid \text{S}_2\ \text{X}_3$

$\Rightarrow$ $\text{S}_2\ \text{X}_4\ \textit{und}\ \text{X}_5 \mid \text{S}_2\ \text{X}_4\ \textit{and}\ \text{X}_5$

$\Rightarrow$ $\text{S}_2\ \textit{unzutreffend und}\ \text{X}_5 \mid \text{S}_2\ \textit{unfounded and}\ \text{X}_5$

$\Rightarrow$ $\text{S}_2\ \textit{unzutreffend und irreführend} \mid \text{S}_2\ \textit{unfounded and misleading}$

$\Rightarrow$ $\text{X}_6\ \textit{unzutreffend und irreführend} \mid \text{X}_6\ \textit{unfounded and misleading}$

$\Rightarrow$ $\textit{deshalb}\ \text{X}_7\ \textit{die}\ \text{X}_8\ \textit{unzutreffend und irreführend}$
$\mid\ \textit{therefore the}\ \text{X}_8\ \text{X}_7\ \textit{unfounded and misleading}$

- $\text{X} \rightarrow \textit{deshalb}\ \text{X}_1\ \textit{die}\ \text{X}_2 \mid \textit{therefore the}\ \text{X}_2\ \text{X}_1$ (non-terminal reordering)

Decoding via (S)CFG derivation

$\Rightarrow$ $\text{S}_2\ \text{X}_3 \mid \text{S}_2\ \text{X}_3$

$\Rightarrow$ $\text{S}_2\ \text{X}_4\ \textit{und}\ \text{X}_5 \mid \text{S}_2\ \text{X}_4\ \textit{and}\ \text{X}_5$

$\Rightarrow$ $\text{S}_2\ \textit{unzutreffend und}\ \text{X}_5 \mid \text{S}_2\ \textit{unfounded and}\ \text{X}_5$

$\Rightarrow$ $\text{S}_2\ \textit{unzutreffend und irreführend} \mid \text{S}_2\ \textit{unfounded and misleading}$

$\Rightarrow$ $\text{X}_6\ \textit{unzutreffend und irreführend} \mid \text{X}_6\ \textit{unfounded and misleading}$

$\Rightarrow$ $\textit{deshalb}\ \text{X}_7\ \textit{die}\ \text{X}_8\ \textit{unzutreffend und irreführend}$
$\mid\ \textit{therefore the}\ \text{X}_8\ \text{X}_7\ \textit{unfounded and misleading}$

$\Rightarrow$ $\textit{deshalb sei die}\ \text{X}_8\ \textit{unzutreffend und irreführend}$
$\mid\ \textit{therefore the}\ \text{X}_8\ \textit{was unfounded and misleading}$

- $\text{X} \rightarrow \textit{sei} \mid \textit{was}$

Decoding via (S)CFG derivation

$\Rightarrow$ $\mathrm{S}_2\ \mathrm{X}_3 \mid \mathrm{S}_2\ \mathrm{X}_3$

$\Rightarrow$ $\mathrm{S}_2\ \mathrm{X}_4\ \textit{und}\ \mathrm{X}_5 \mid \mathrm{S}_2\ \mathrm{X}_4\ \textit{and}\ \mathrm{X}_5$

$\Rightarrow$ $\mathrm{S}_2\ \textit{unzutreffend und}\ \mathrm{X}_5 \mid \mathrm{S}_2\ \textit{unfounded and}\ \mathrm{X}_5$

$\Rightarrow$ $\mathrm{S}_2\ \textit{unzutreffend und irreführend} \mid \mathrm{S}_2\ \textit{unfounded and misleading}$

$\Rightarrow$ $\mathrm{X}_6\ \textit{unzutreffend und irreführend} \mid \mathrm{X}_6\ \textit{unfounded and misleading}$

$\Rightarrow$ $\textit{deshalb}\ \mathrm{X}_7\ \textit{die}\ \mathrm{X}_8\ \textit{unzutreffend und irreführend}$
$\mid\ \textit{therefore the}\ \mathrm{X}_8\ \mathrm{X}_7\ \textit{unfounded and misleading}$

$\Rightarrow$ $\textit{deshalb sei die}\ \mathrm{X}_8\ \textit{unzutreffend und irreführend}$
$\mid\ \textit{therefore the}\ \mathrm{X}_8\ \textit{was unfounded and misleading}$

$\Rightarrow$ $\textit{deshalb sei die Werbung unzutreffend und irreführend}$
$\mid\ \textit{therefore the advertisement was unfounded and misleading}$

- $\mathrm{X} \rightarrow \textit{Werbung} \mid \textit{advertisement}$

1. Statistical Machine Translation
   - Basics
   - Phrase-based SMT
   - Hierarchical SMT
   - Syntax-based SMT

Syntax-based SMT

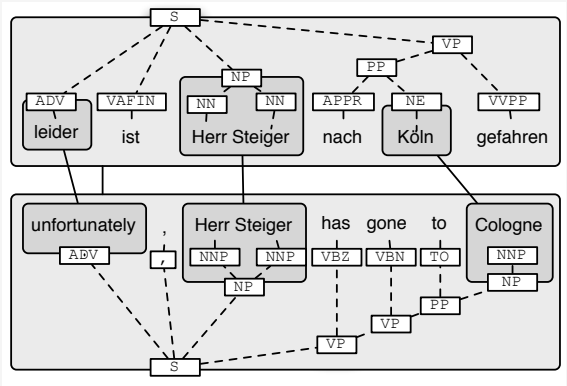
core idea

- use syntax on source, target, or both
- rule extraction constrained by syntax
- potentially use syntactic structures for scoring (syntax-based LMs)

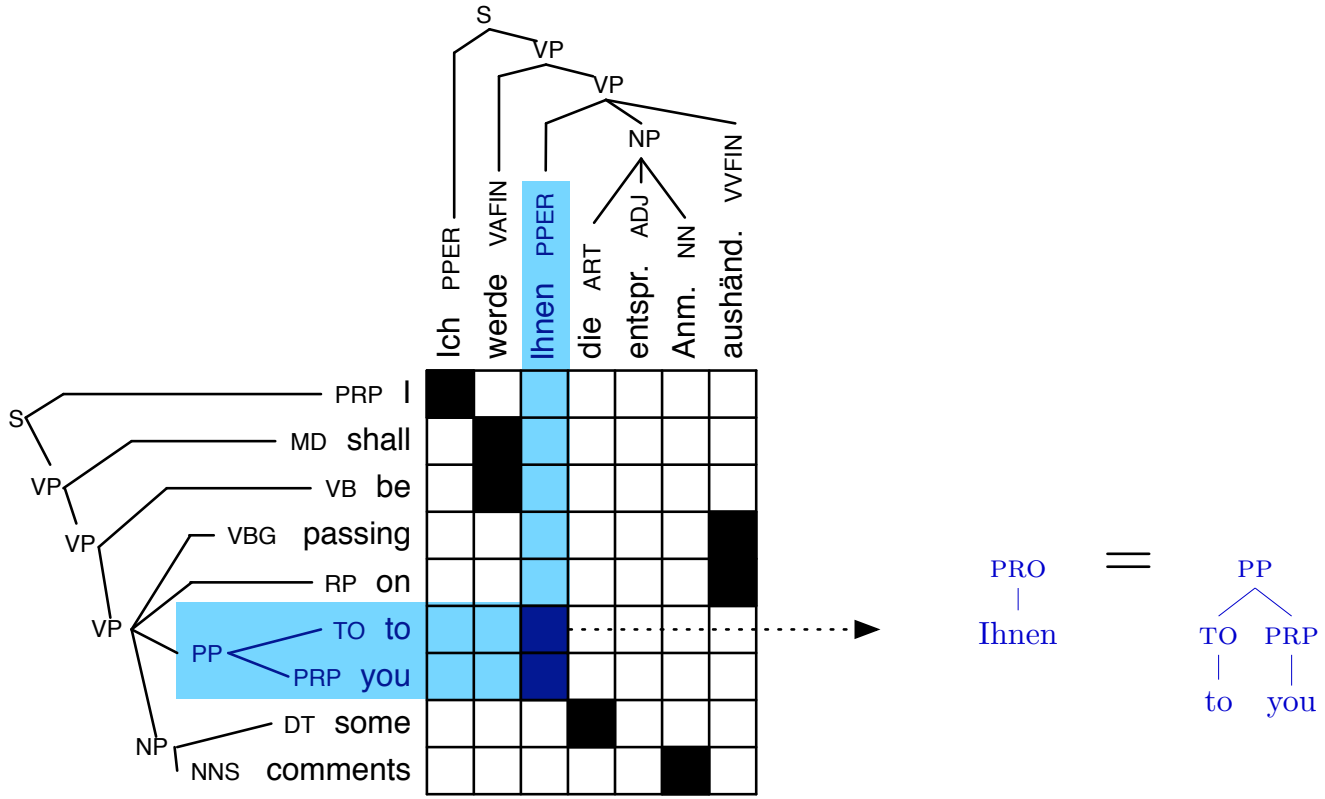
consequences

depend on exact flavor of syntax used; here: string-to-tree SMT

- + less overgeneralisation
- - sparsity in grammar requires relaxation of extraction constraints
- - label matching constraints increase search space during decoding



Syntax-based Phrase Extraction



Decoding

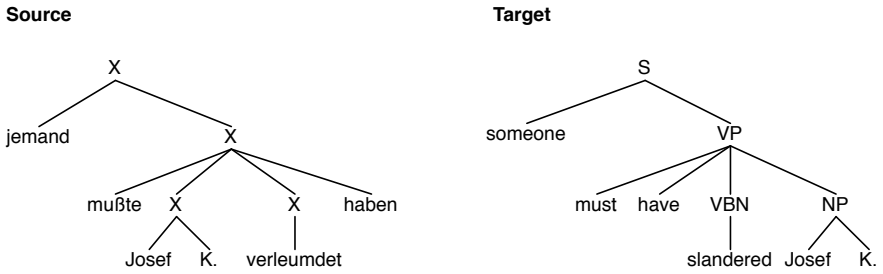
Input jemand mußte Josef K. verleumdet haben

someone must Josef K. slandered have

Grammar

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

Derivation 1



Why Syntax-based SMT?

- many variants (syntax on source/target/both...)
- syntactic constraints for rule extraction and application prevent some over-generalizations
- syntactic structure can be exploited by feature functions:

unification constraints [Williams, 2009]

"eine" $\rightarrow$

$$\begin{bmatrix} \text{cat} & \text{ART} \\ \text{infl} & \begin{bmatrix} \text{case} & \text{nom} \\ \text{declension} & \text{mixed} \\ \text{agr} & \begin{bmatrix} \text{gender} & \text{f} \\ \text{num} & \text{sg} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

"Welt" $\rightarrow$

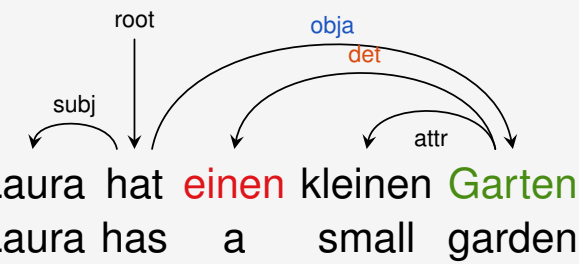
$$\begin{bmatrix} \text{cat} & \text{NN} \\ \text{infl} & \begin{bmatrix} \text{case} & \text{nom} \\ \text{agr} & \begin{bmatrix} \text{gender} & \text{f} \\ \text{num} & \text{sg} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

syntax-based neural language model [Sennrich, 2015]

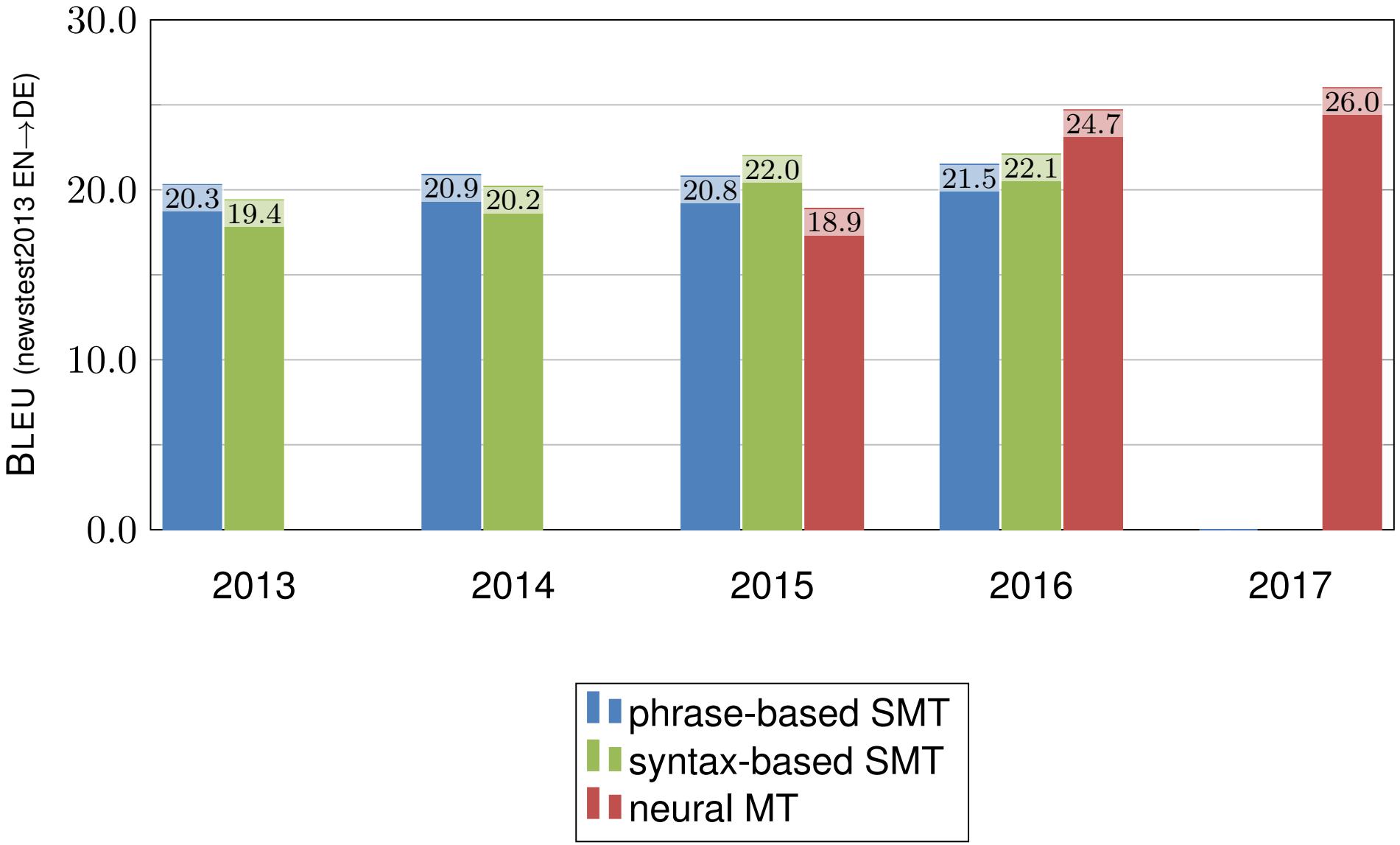
$$P_{\text{SYNTAX}}(T, D) \approx \prod_{i=1}^{n} P_l(i) \times P_w(i)$$

$$P_l(i) = P(l_i \mid w_a(i), l_a(i))$$

$$P_w(i) = P(w_i \mid l_i, w_a(i), l_a(i))$$



Edinburgh's* WMT Results over the Years



*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

What Phrase-based SMT (Still) Does Better than NMT

- better performance in low-data conditions [Koehn and Knowles, 2017]
- clear stopping criterion at decoding time:
  when all source words have been covered by a phrase pair
- good ecosystem of methods for specialized requirements (e.g. inclusion of terminology)
- ability to inspect translation decisions and models:
  - alignment between source and output
  - add/remove phrase table entries


Lifestyle › Tech

Thousands sign petition asking [redacted] to remove homophobic slurs from translation service

Company later obliged and slurs were taken down

Software

Moses SMT Toolkit

- developed in Edinburgh
- many features and extensive documentation:
  `http://www.statmt.org/moses`
- documentation of baseline phrase-based systems:
  `http://www.statmt.org/moses/?n=moses.baseline`
  `http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2017/baseline/baselineSystemPhrase_kj.html`
- config files for SOTA (in 2014/5) syntax-based systems:
  `https://github.com/rsennrich/wmt2014-scripts`

Further Reading

text books

- Philipp Koehn (2009). Statistical Machine Translation.
- Philip Williams; Rico Sennrich; Matt Post; Philipp Koehn (2016). Syntax-based Statistical Machine Translation.

online resources

- syntax-based tutorial by Philip Williams and Philipp Koehn (slide credit to them for some slides shown here):

  `http://homepages.inf.ed.ac.uk/s0898777/syntax-tutorial.pdf`

- slides on word- and phrase-based SMT by Philipp Koehn:

  `http://www.statmt.org/book/slides/04-word-based-models.pdf`

  `http://www.statmt.org/book/slides/05-phrase-based-models.pdf`

  `http://www.statmt.org/book/slides/06-decoding.pdf`

Bibliography I

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., and Mercer, R. L. (1993).
The Mathematics of Statistical Machine Translation: Parameter Estimation.
Computational Linguistics, 19(2):263–311.

Koehn, P. and Knowles, R. (2017).
Six Challenges for Neural Machine Translation.
In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver. Association for Computational Linguistics.

Och, F. J. (2003).
Minimum Error Rate Training in Statistical Machine Translation.
In ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Sennrich, R. (2015).
Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation.
Transactions of the Association for Computational Linguistics, 3:169–182.

Williams, P. (2009).
Towards Statistical Machine Translation with Unification Grammars.
Master's thesis, University of Edinburgh, Edinburgh, UK.