

Machine Translation

12: (Non-neural) Statistical Machine Translation

Rico Sennrich

University of Edinburgh

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

MT – 2018 – 12

Refresher: A probabilistic model of translation

1 Statistical Machine Translation

- Basics
- Phrase-based SMT
- Hierarchical SMT
- Syntax-based SMT

- Suppose that we have:
 - a source sentence S of length m (x_1, \dots, x_m)
 - a target sentence T of length n (y_1, \dots, 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 $P(S|T)$?
→ break it up into smaller units

chicken-and-egg problem

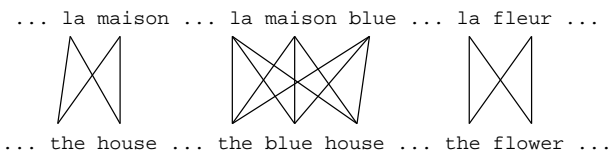
let's break up $P(S|T)$ into small units (words):

- 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?

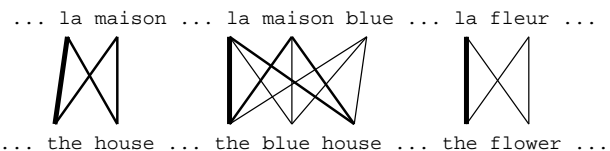
solution: **Expectation Maximization Algorithm**

- initialize model
- iterate between estimating alignment and translation model

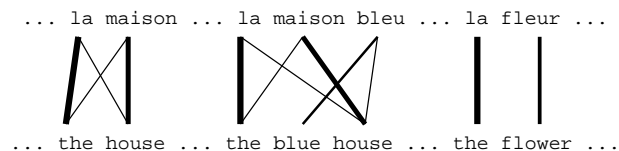
simplest model based on lexical translation; more complex models consider position and fertility



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



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



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



- Convergence
- Inherent hidden structure revealed by EM

- Probabilities

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

$p(e, a f) = 0.56$	$p(e, a f) = 0.035$	$p(e, a f) = 0.08$	$p(e, a f) = 0.005$
$p(a e, f) = 0.824$	$p(a e, f) = 0.052$	$p(a e, f) = 0.118$	$p(a e, f) = 0.007$
- Counts

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

$$T^* = \arg \max_T P(S|T)P(T)$$

Bayes' theorem

$$T^* \approx \arg \max_T \sum_{m=1}^M \lambda_m h_m(S, T)$$

[Och, 2003]

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

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

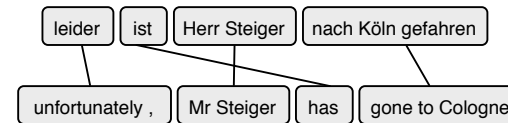
- 1 Statistical Machine Translation
 - Basics
 - **Phrase-based SMT**
 - Hierarchical SMT
 - Syntax-based SMT

core idea

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

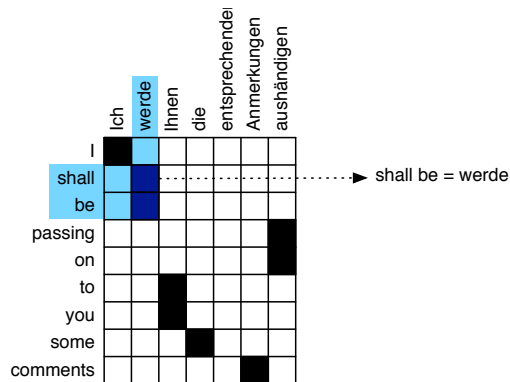
consequences

- + 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



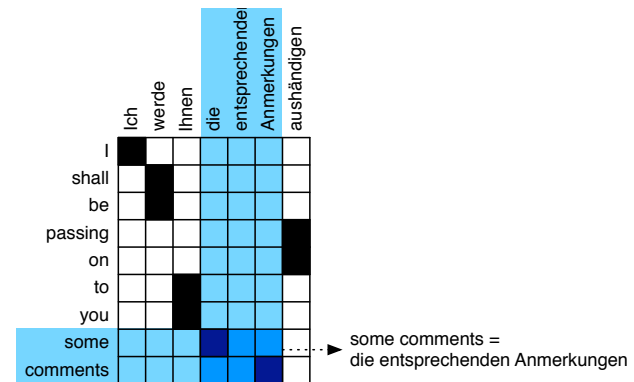
Phrase Extraction

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



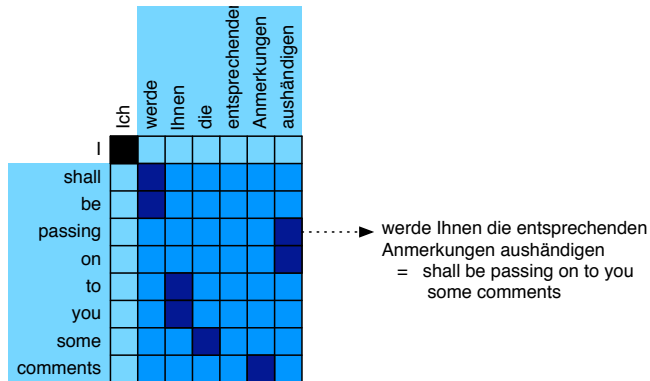
Phrase Extraction

- 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



R. Sennrich

MT – 2018 – 12

10 / 27

Common Features in Phrase-based SMT

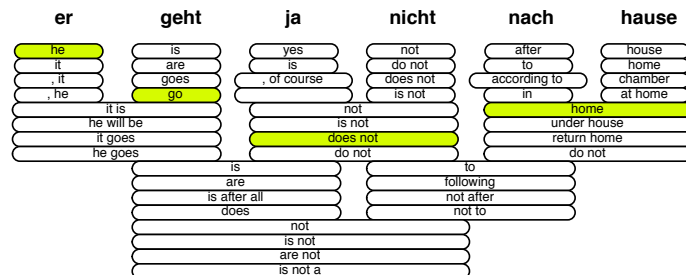
- phrase translation probabilities (in both directions)
- word translation probabilities (in both directions)
- language model
- reordering model
- constant penalty for each phrase used
- sparse features with learned cost for some (classes of) phrase pairs
- multiple models of each type possible

R. Sennrich

MT – 2018 – 12

11 / 27

Decoding



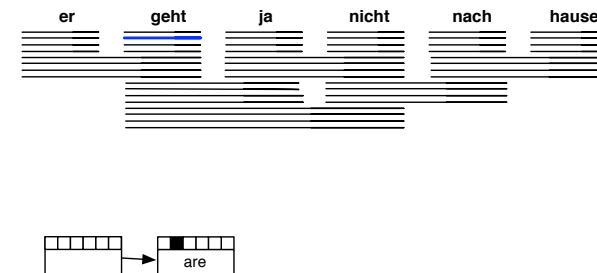
- The machine translation decoder does not know the right answer
 - picking the right translation options
 - arranging them in the right order
- Search problem solved by heuristic beam search

R. Sennrich

MT – 2018 – 12

12 / 27

Decoding



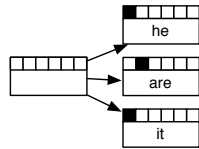
pick any translation option, create new hypothesis

R. Sennrich

MT – 2018 – 12

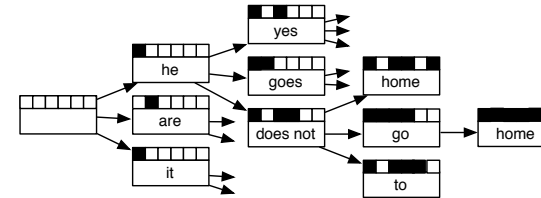
13 / 27

Decoding



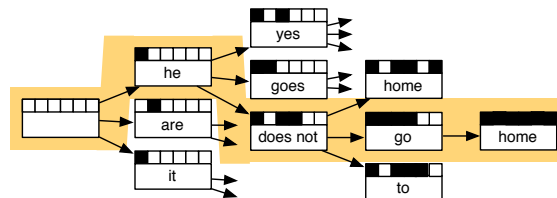
create hypotheses for all other translation options

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

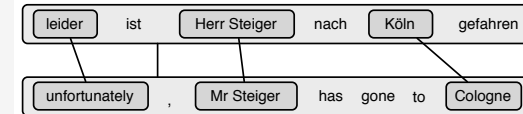
- 1 Statistical Machine Translation
 - Basics
 - Phrase-based SMT
 - Hierarchical SMT
 - Syntax-based SMT

core idea

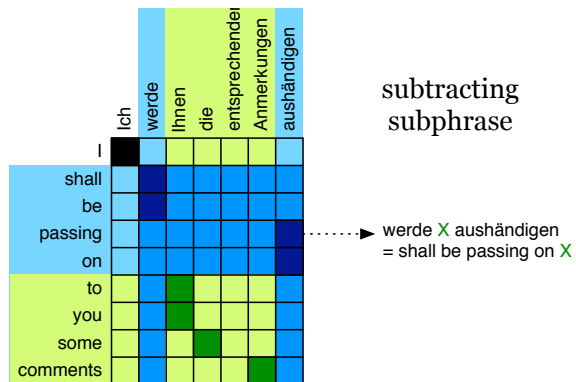
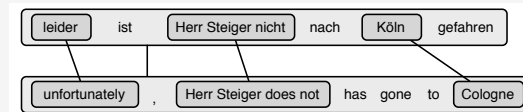
use context-free grammars (CFG) rules as basic translation units
 → allows gaps

consequences

- + better modeling of some reordering patterns



- - overgeneralisation is still possible



Decoding via (S)CFG derivation

- Derivation starts with pair of linked s symbols.

Decoding

Decoding via (S)CFG derivation

$\Rightarrow S_2 X_3 \mid S_2 X_3$

- $S \rightarrow S_1 X_2 \mid S_1 X_2$ (glue rule)

Decoding

Decoding via (S)CFG derivation

$\Rightarrow S_2 X_3 \mid S_2 X_3$

$\Rightarrow S_2 X_4 \text{ und } X_5 \mid S_2 X_4 \text{ and } X_5$

- $X \rightarrow X_1 \text{ und } X_2 \mid X_1 \text{ and } X_2$

Decoding

Decoding via (S)CFG derivation

$\Rightarrow S_2 X_3 \mid S_2 X_3$

$\Rightarrow S_2 X_4 \text{ und } X_5 \mid S_2 X_4 \text{ and } X_5$

$\Rightarrow S_2 \text{ unzutreffend und } X_5 \mid S_2 \text{ unfounded and } X_5$

- $X \rightarrow \text{unzutreffend} \mid \text{unfounded}$

Decoding

Decoding via (S)CFG derivation

$\Rightarrow S_2 X_3 \mid S_2 X_3$

$\Rightarrow S_2 X_4 \text{ und } X_5 \mid S_2 X_4 \text{ and } X_5$

$\Rightarrow S_2 \text{ unzutreffend und } X_5 \mid S_2 \text{ unfounded and } X_5$

$\Rightarrow S_2 \text{ unzutreffend und irreführend} \mid S_2 \text{ unfounded and misleading}$

- $X \rightarrow \text{irreführend} \mid \text{misleading}$

Decoding

Decoding via (S)CFG derivation

- ⇒ $S_2 X_3$ | $S_2 X_3$
- ⇒ $S_2 X_4$ und X_5 | $S_2 X_4$ and X_5
- ⇒ S_2 unzutreffend und X_5 | S_2 unfounded and X_5
- ⇒ S_2 unzutreffend und irreführend | S_2 unfounded and misleading
- ⇒ X_6 unzutreffend und irreführend | X_6 unfounded and misleading

- $S \rightarrow X_1$ | X_1 (glue rule)

Decoding

Decoding via (S)CFG derivation

- ⇒ $S_2 X_3$ | $S_2 X_3$
- ⇒ $S_2 X_4$ und X_5 | $S_2 X_4$ and X_5
- ⇒ S_2 unzutreffend und X_5 | S_2 unfounded and X_5
- ⇒ S_2 unzutreffend und irreführend | S_2 unfounded and misleading
- ⇒ X_6 unzutreffend und irreführend | X_6 unfounded and misleading
- ⇒ deshalb X_7 die X_8 unzutreffend und irreführend
| therefore the X_8 X_7 unfounded and misleading

- $X \rightarrow$ deshalb X_1 die X_2 | therefore the X_2 X_1 (non-terminal reordering)

Decoding

Decoding via (S)CFG derivation

- ⇒ $S_2 X_3$ | $S_2 X_3$
- ⇒ $S_2 X_4$ und X_5 | $S_2 X_4$ and X_5
- ⇒ S_2 unzutreffend und X_5 | S_2 unfounded and X_5
- ⇒ S_2 unzutreffend und irreführend | S_2 unfounded and misleading
- ⇒ X_6 unzutreffend und irreführend | X_6 unfounded and misleading
- ⇒ deshalb X_7 die X_8 unzutreffend und irreführend
| therefore the X_8 X_7 unfounded and misleading
- ⇒ deshalb sei die X_8 unzutreffend und irreführend
| therefore the X_8 was unfounded and misleading

- $X \rightarrow$ sei | was

Decoding

Decoding via (S)CFG derivation

- ⇒ $S_2 X_3$ | $S_2 X_3$
- ⇒ $S_2 X_4$ und X_5 | $S_2 X_4$ and X_5
- ⇒ S_2 unzutreffend und X_5 | S_2 unfounded and X_5
- ⇒ S_2 unzutreffend und irreführend | S_2 unfounded and misleading
- ⇒ X_6 unzutreffend und irreführend | X_6 unfounded and misleading
- ⇒ deshalb X_7 die X_8 unzutreffend und irreführend
| therefore the X_8 X_7 unfounded and misleading
- ⇒ deshalb sei die X_8 unzutreffend und irreführend
| therefore the X_8 was unfounded and misleading
- ⇒ deshalb sei die Werbung unzutreffend und irreführend
| therefore the advertisement was unfounded and misleading

- $X \rightarrow$ Werbung | advertisement

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

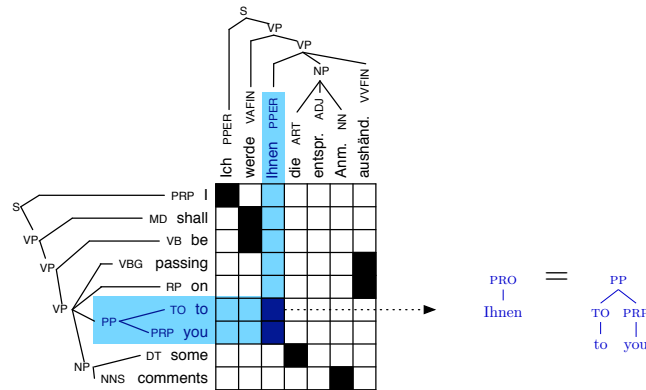
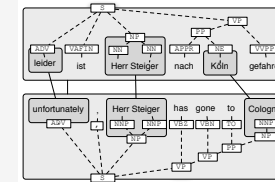
core idea

- 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



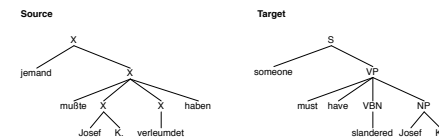
Input

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

Grammar

- ⇒ r₁: NP → Josef K. | Josef K. 0.90
- ⇒ r₂: VBN → verleumdet | slandered 0.40
- r₃: VBN → verleumdet | defamed 0.20
- ⇒ r₄: VP → mußte X₁ X₂ haben | must have VBN₂ NP₁ 0.10
- ⇒ r₅: S → jemand X₁ | someone VP₁ 0.60
- r₆: S → jemand mußte X₁ X₂ haben | someone must have VBN₂ NP₁ 0.80
- r₇: S → jemand mußte X₁ X₂ haben | NP₁ must have been VBN₁ by someone 0.05

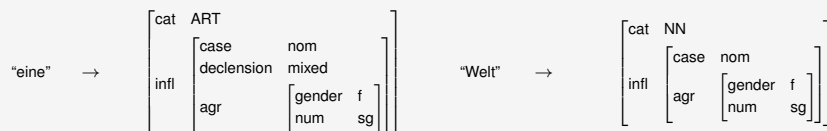
Derivation 1



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]

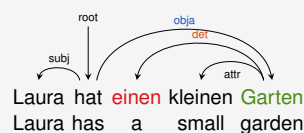


syntax-based neural language model [Sennrich, 2015]

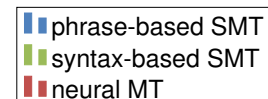
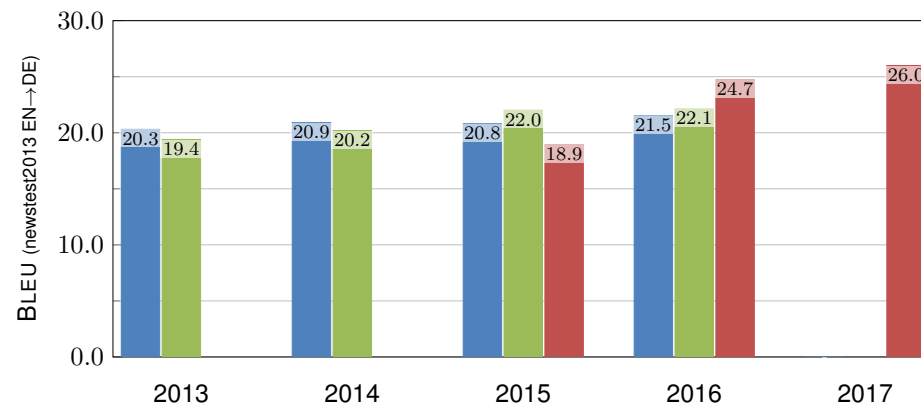
$$P_{\text{SYNTAX}}(T, D) \approx \prod_{i=1}^n P_i(i) \times P_w(i)$$

$$P_i(i) = P(l_i | w_a(i), l_a(i))$$

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



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 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>

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>

-  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.