Hierarchical Machine Translation

Miguel Rios

Universiteit van Amsterdam

May 4, 2018

Content

Recap

- Noisy Channel
  $P(E|F) = \frac{P(E)P(F|E)}{P(F)}$

Recap

- Noisy Channel
  $P(E|F) = \frac{P(E)P(F|E)}{P(F)}$
- Most likely translation
  $\underset{e}{\text{argmax}} P(E|F) = \underset{e}{\text{argmax}} P(E)P(F|E)$
  (1) the chance that someone would say **E** first place
  (2) if say **E**, the chance that someone else would translate it into **F**.
  (3) $P(F|E)$ will ensure that a good **E** will have words that generally translate to words in **F**.
  (4) $P(E)$ language model.

Recap

- Noisy Channel
  $P(E|F) = \frac{P(E)P(F|E)}{P(F)}$
- Most likely translation
  $\underset{e}{\text{argmax}} P(E|F) = \underset{e}{\text{argmax}} P(E)P(F|E)$
  (1) the chance that someone would say **E** first place
  (2) if say **E**, the chance that someone else would translate it into **F**.
  (3) $P(F|E)$ will ensure that a good **E** will have words that generally translate to words in **F**.
  (4) $P(E)$ language model.
- Linear Model

$$S_\theta(e, d, f) = \theta^T \sum_i^n h_i(d_i|e, f)$$

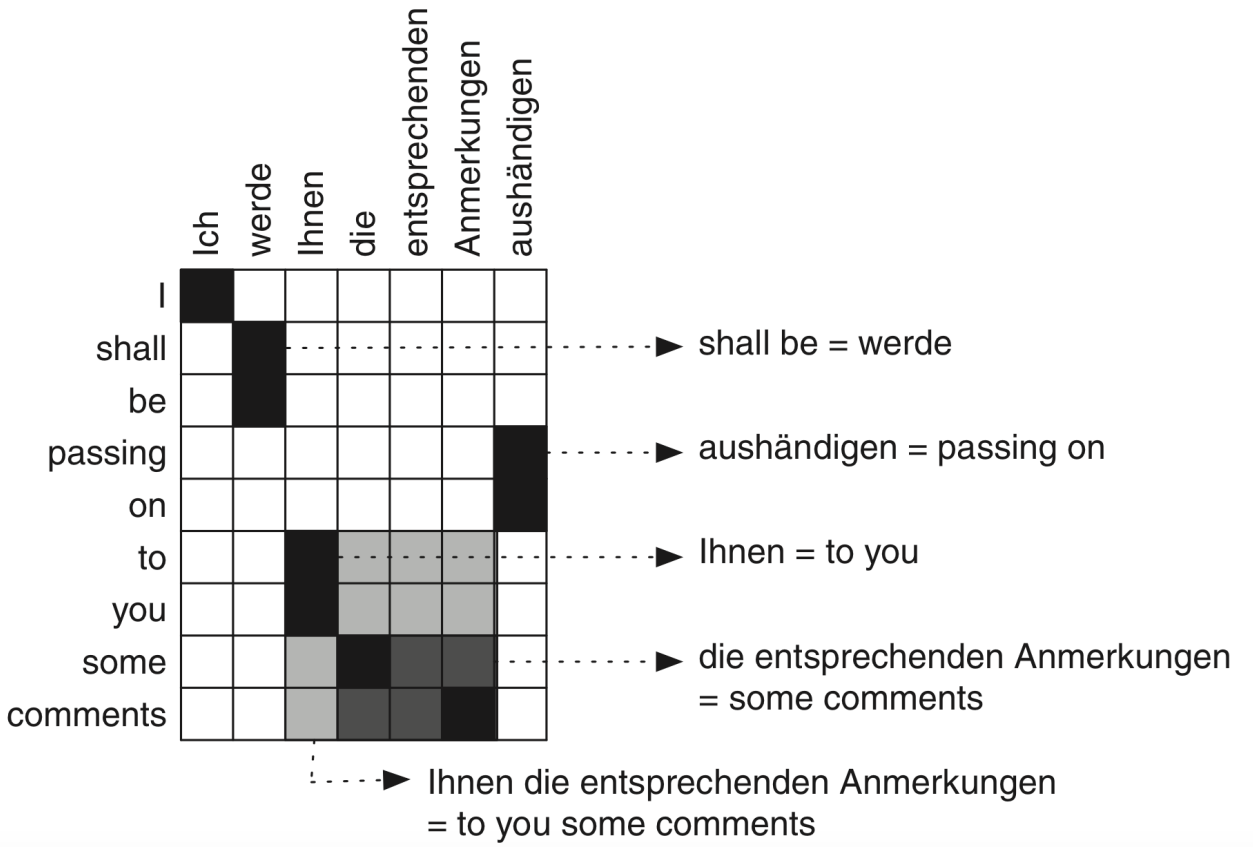


Figure: Koehn [2010]

werde X aushändigen | shall be passing on X

Why hierarchical structure?

Better generalisation

- compositionality
- reordering

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order
  e.g. different syntactic structure

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order
  e.g. different syntactic structure
  e.g. rich morphology

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order
  e.g. different syntactic structure
  e.g. rich morphology

Reordering is arguably one of the hardest problems in MT

Why is reordering important?

Monotone translation is unrealistic

- languages differ wrt word-order
  e.g. different syntactic structure
  e.g. rich morphology

Reordering is arguably one of the hardest problems in MT

- part of the model of translational equivalences
  the part that determines the space of translations

Key aspects

Expressiveness

- how much can two languages differ wrt word order?

Key aspects

Expressiveness

- how much can two languages differ wrt word order?

Modelling

- how many parameters do we have to estimate?

Content

Hierarchical phrase-based - Motivation

Local Reordering

	J'	ai	les	yeux	noirs
I	■				
have		■			
black					■
eyes			■	■	

Hierarchical phrase-based - Motivation

Local Reordering

	J'	ai	les	yeux	noirs
I					
have					
black					
eyes					

- Monotone
  J'$_1$ ai$_2$ $\rightarrow$ I$_1$ have$_2$

Hierarchical phrase-based - Motivation

Local Reordering

	j'	ai	les	yeux	noirs
I	■				
have		■			
black					■
eyes			■	■	

- Swap
  les yeux$_4$ noirs$_5$ $\rightarrow$ black$_3$ eyes$_4$

Hierarchical phrase-based - Motivation

Local Reordering

	J'	ai	les	yeux	noirs
I					
have					
black					
eyes					

- Discontinuous
  ai$_2$ X_{3-4} noirs$_5$ $\rightarrow$ have$_2$ black$_3$ X_4

Hierarchical phrase-based - Motivation

Discontiguous Phrases

	Je	ne	vais	pas
I	■			
do		■		■
not		■		■
go			■	

Hierarchical phrase-based - Motivation

Discontiguous Phrases

	Je	ne	vais	pas
I				
do				
not				
go				

- Gappy phrase
  ne vais pas → do not go
  ne X_{vais} pas → do not X_{go}

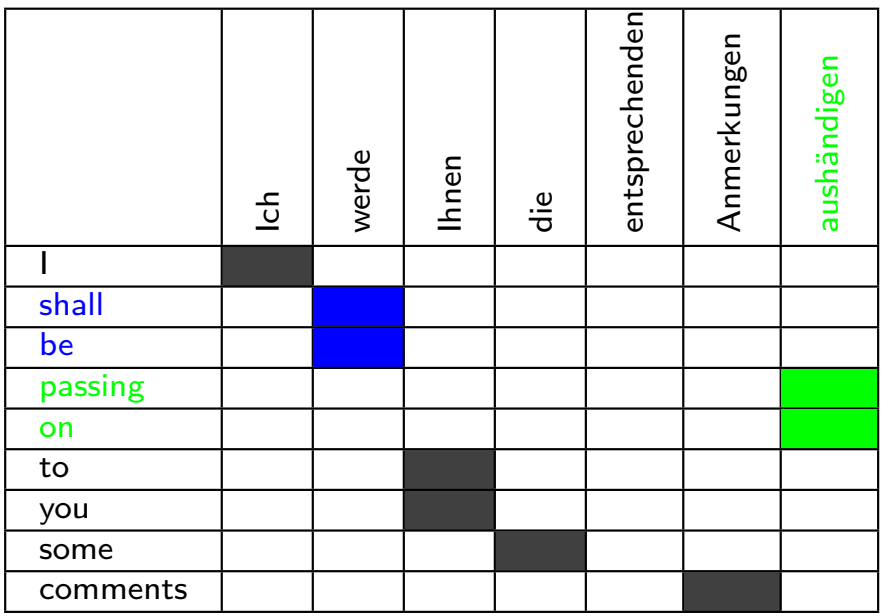
Hierarchical phrase-based - Motivation

Long Distance Reordering

	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I	■						
shall		■					
be		■					
passing							■
on							■
to			■				
you			■				
some				■			
comments						■	

Hierarchical phrase-based - Motivation

Long Distance Reordering



- How can we extract a biphrase for **shall be passing on**?

Hierarchical phrase-based - Motivation

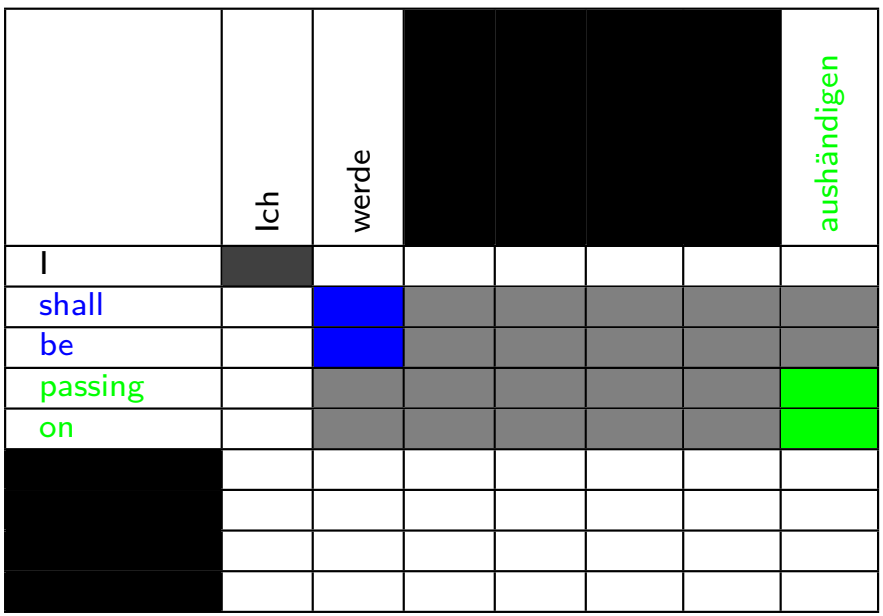
Long Distance Reordering

	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I							
shall							
be							
passing							
on							
to			X				
you			X				
some				X			
comments						X	

- How can we extract a biphrase for **shall be passing on**?
- We cannot, we need to extract **to you some comments** along

Hierarchical phrase-based - Motivation

Long Distance Reordering



- How can we extract a biphrase for **shall be passing on**?
- We cannot, we need to extract **to you some comments** along
- Unless we replace all those words by a variable

Hierarchical phrase-based - Motivation

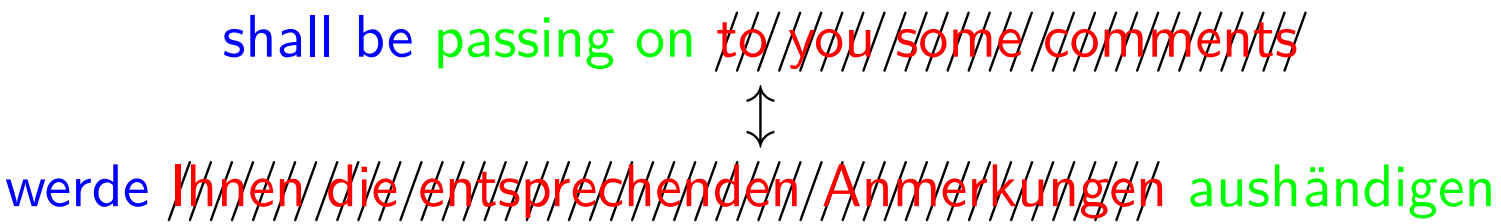
Long Distance Reordering

shall be passing on to you some comments

werde Ihnen die entsprechenden Anmerkungen aushändigen

Hierarchical phrase-based - Motivation

Long Distance Reordering



Hierarchical phrase-based - Motivation

Long Distance Reordering

shall be passing on X

$\updownarrow$

werde X aushändigen

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model
- SCFG decoding

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model
- SCFG decoding

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model
- SCFG decoding

Motivation

- long-distance reordering

Hiero

Extends phrase-based MT with hierarchical rules [Chiang, 2005]

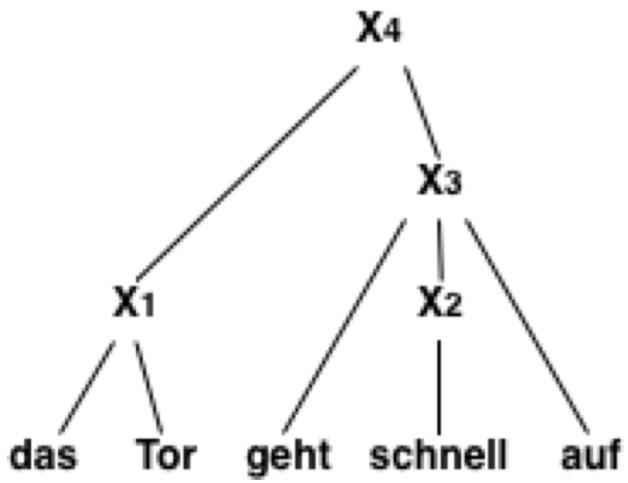
- conditions on word alignment
- heuristic rule extraction
- heuristic scoring by relative frequency counting
- log-linear model
- SCFG decoding

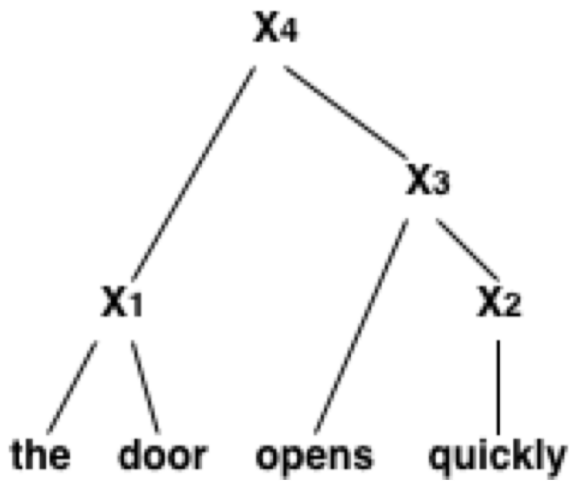
Motivation

- long-distance reordering
- lexicalised reordering

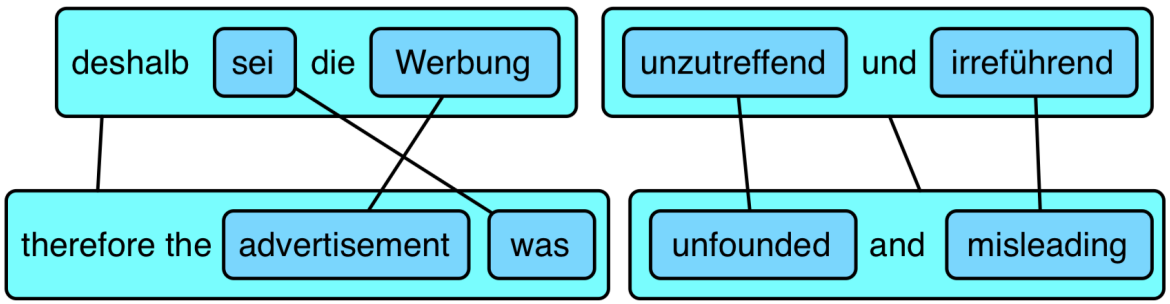
Hiero

PBSMT, one level of hierarchy.
HPBSMT, any kind of tree depth.





Hiero



Rules with two non-terminals:

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

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

Heuristic rule extraction

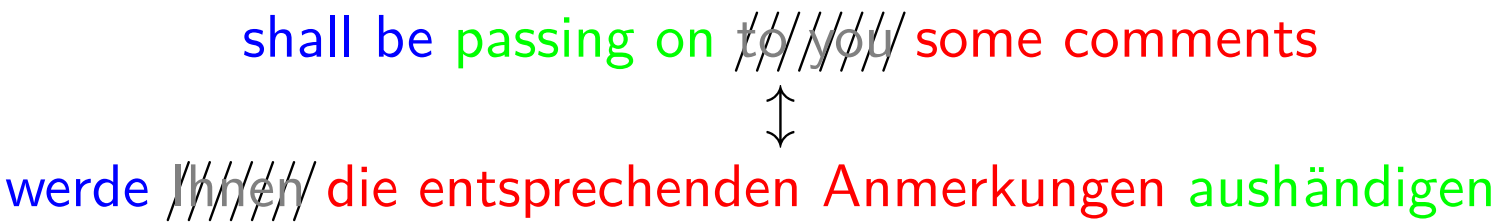
Initial phrase pairs created with same heuristic as PBSMT.

shall be passing on to you some comments

werde Ihnen die entsprechenden Anmerkungen aushändigen

Heuristic rule extraction

Initial phrase pairs created with same heuristic as PBSMT.



Heuristic rule extraction

Initial phrase pairs created with same heuristic as PBSMT.

shall be passing on X_1 some comments

werde X_1 die entsprechenden Anmerkungen aushändigen

Heuristic rule extraction

Initial phrase pairs created with same heuristic as PBSMT.

shall be passing on X_1 some comments

↕

werde X_1 die entsprechenden Anmerkungen aushändigen

Heuristic rule extraction

Initial phrase pairs created with same heuristic as PBSMT.

shall be passing on X_1 X_2

$\updownarrow$

werde X_1 X_2 aushändigen

Heuristic rule extraction

Initial phrase pairs created with same heuristic as PBSMT.

$[X] \rightarrow$ shall be passing on X_1 X_2 | werde X_1 X_2 aushändigen

$[X] \rightarrow$ shall be passing on X_3 | werde X_3 aushändigen

$[X] \rightarrow$ to you | Ihnen

$[X] \rightarrow$ some comments | die entsprechenden Anmerkungen

$[X] \rightarrow$ to you some comments | Ihnen die entsprechenden Anmerkungen

Hiero - Scoring

Relative frequency: assume all fragments have been "observed"
Give a count of one to phrase pair occurrence, then distribute its weight equally among the obtained rules.

- Joint rule probatility: $p(LHS, RHS_{source}, RHS_{target})$

$$p(X, \text{la maison } X_1, \text{the } X_1 \text{ house})$$

- Rule application probability: $p(RHS_{source}, RHS_{target}|LHS)$

$$p(\text{la maison } X_1, \text{the } X_1 \text{ house}|X)$$

- Direct translation probability: $p(RHS_{target}|RHS_{source}, LHS)$

$$p(\text{the } X_1 \text{ house}|\text{la maison } X_1, X)$$

- Noisy-channel translation probability: $p(RHS_{source}|RHS_{target}, LHS)$

$$p(\text{la maison } X_1|\text{the } X_1 \text{ house}, X)$$

- Lexical translation probability

$\prod_{t_i \in RHS_{target}} p(t_i|RHS_{source}, a)$ $\prod_{s_i \in RHS_{source}} p(s_i|RHS_{target}, a)$

Hiero - Model

Log-linear combination of features

Hiero - Model

Log-linear combination of features Linear model

$$S_\theta(e, d, f) = \theta^T \sum_{r_{s,t} \epsilon d} h_i(r_{s,t}|e, f)$$

where s is a span over F,
t is a span over E
and r is a rule.
Weighted synchronous CFG.
LM.

Content

Decoding

Phrase-based

Tree-based

Decoding

Phrase-based

- Left-to-Right

Tree-based

- Bottom-Up

Decoding

Phrase-based

- Left-to-Right
- Beam Search

Tree-based

- Bottom-Up
- Chart Parsing

Decoding

Phrase-based

- Left-to-Right
- Beam Search
- Formally intersection:

Tree-based

- Bottom-Up
- Chart Parsing
- Formally intersection:

Decoding

Phrase-based

- Left-to-Right
- Beam Search
- Formally intersection:
- FST (TM) $\times$ FSA (LM)

Tree-based

- Bottom-Up
- Chart Parsing
- Formally intersection:
- SCFG (TM) $\times$ FSA (LM)

Content

Discriminative Model

- model consists of features.

Discriminative Model

- model consists of features.
- each feature has a weight.

Discriminative Model

- model consists of features.
- each feature has a weight.
- supervised learning: tune feature weights wrt. an evaluation metric on development data

Discriminative Model

- model consists of features.
- each feature has a weight.
- supervised learning: tune feature weights wrt. an evaluation metric on development data
- Which objective?
  Bilingual Evaluation Understudy metric BLEU

Tuning

Task: find weights so that the model ranks best translations first.

- Translate development corpus using model with current feature weights,
  N -best list of translations (N = 100, 1000, . . .)

Tuning

Task: find weights so that the model ranks best translations first.

- Translate development corpus using model with current feature weights,
  N -best list of translations (N = 100, 1000, . . .)
- Evaluate translations with the objective

Tuning

Task: find weights so that the model ranks best translations first.

- Translate development corpus using model with current feature weights,
  N -best list of translations (N = 100, 1000, . . .)
- Evaluate translations with the objective
- Adjust feature weights to increase the gain

Tuning

Task: find weights so that the model ranks best translations first.

- Translate development corpus using model with current feature weights,
  N -best list of translations (N = 100, 1000, . . .)
- Evaluate translations with the objective
- Adjust feature weights to increase the gain
- Iterate translation, evaluation, and adjustment of feature weights

MERT

Minimum error rate training (MERT)

- coordinate ascent, where the search updates a feature weight which appears most likely to offer improvements.

MERT

Minimum error rate training (MERT)

- coordinate ascent, where the search updates a feature weight which appears most likely to offer improvements.
- Highest point in a hilly city with a grid of streets, like San Francisco. [Koehn, 2008]
  We start along a certain street.
  Find its highest point and continue along the cross-street.
  Also in this cross-street we find the highest point.

MERT

- Line search for best feature weights
  given: sentences with n-best lists of translations

MERT

- Line search for best feature weights
  given: sentences with n-best lists of translations
- iterate n times
  randomize starting feature weights

MERT

- Line search for best feature weights
  given: sentences with n-best lists of translations
- iterate n times
  randomize starting feature weights
  for each feature

MERT

- Line search for best feature weights
  given: sentences with n-best lists of translations
- iterate n times
  randomize starting feature weights
  for each feature
  - find best feature weight

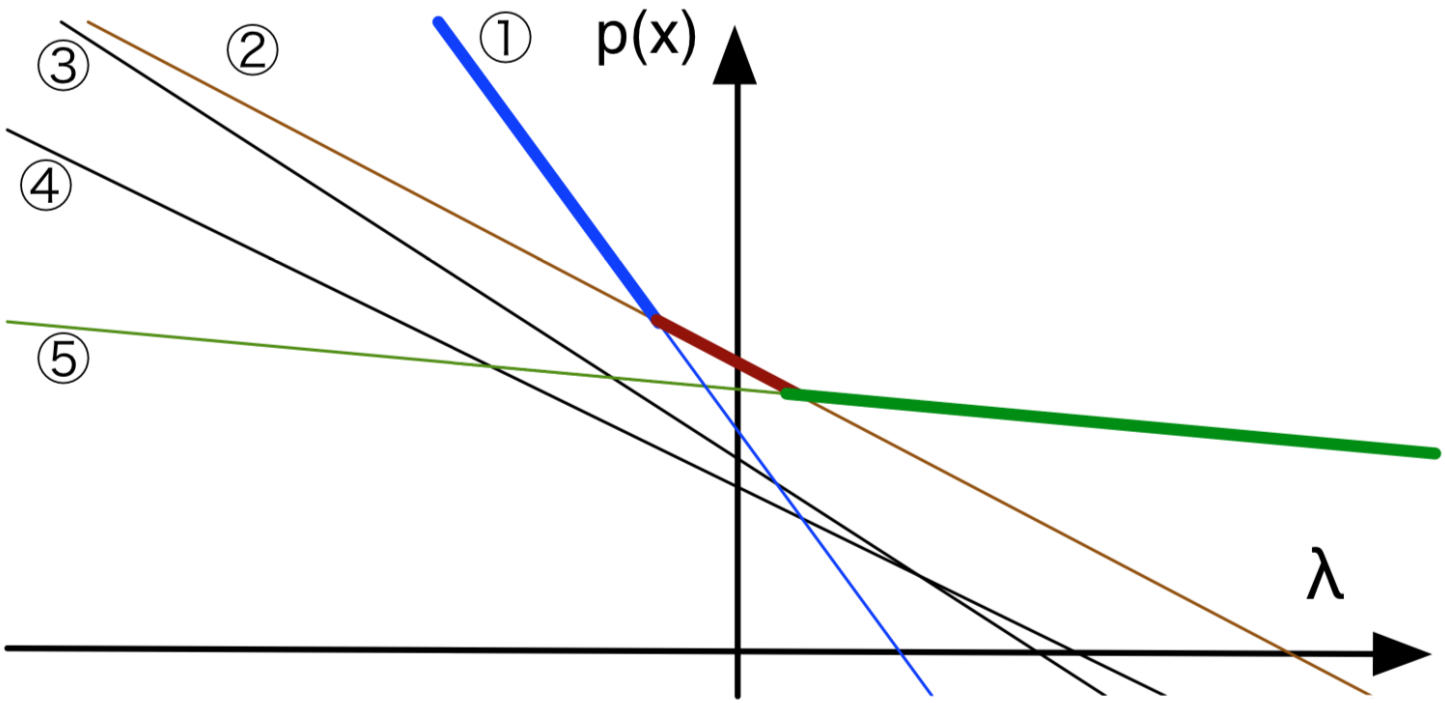
MERT

- Line search for best feature weights
  given: sentences with n-best lists of translations
- iterate n times
  randomize starting feature weights
  for each feature
  - find best feature weight
  - update if different from current

MERT

- Line search for best feature weights
  given: sentences with n-best lists of translations
- iterate n times
  randomize starting feature weights
  for each feature
  - find best feature weight
  - update if different from current
- return best feature weights found in any iteration

MERT



Homework

- Neural Machine Translation and Sequence-to-sequence Models: A Tutorial
  - **Section 5.1 - 5.3** Neural Networks and Feed-forward Language Models
  - **Section 6.1-6.4, 6.5** Recurrent Neural Network Language Models
- Familiarise with preprocessing (Tokenizer, Lowercase, BPE)

Homework

- Deep Learning, NLP, and Representations
  `http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/`
- Understanding LSTM Networks
  `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`

Questions?

References I

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219873. URL `http://www.aclweb.org/anthology/P05-1033`.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010. ISBN 0521874157, 9780521874151.