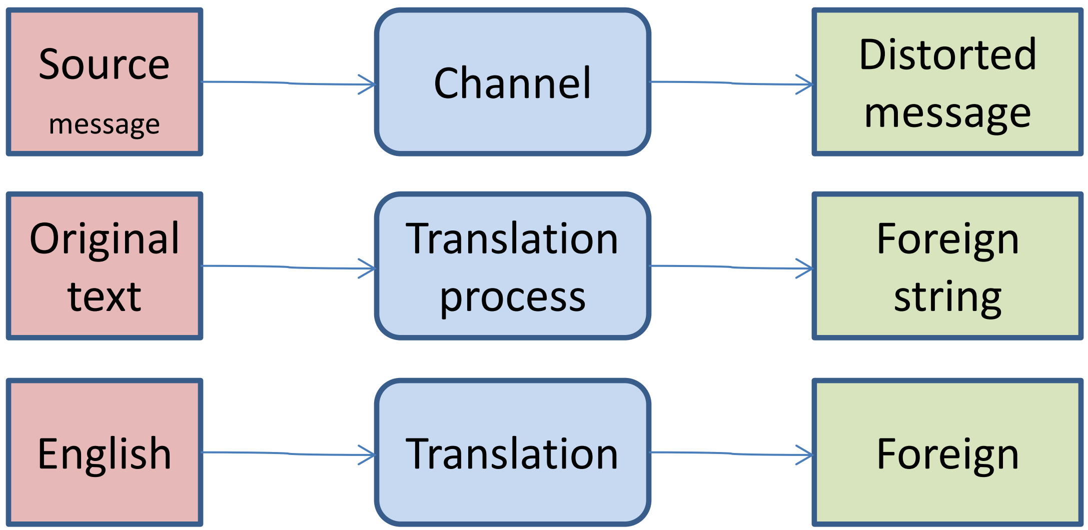
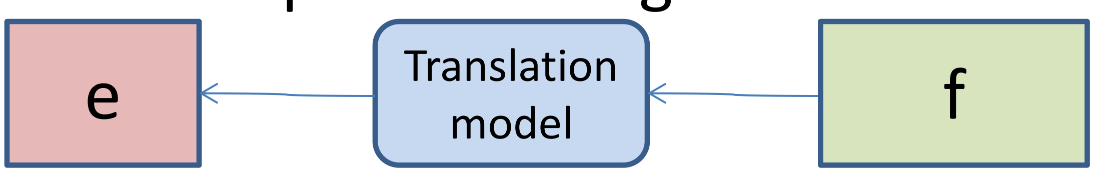
Lecture 5.1: phrase-based MT

Background: Noisy channel model

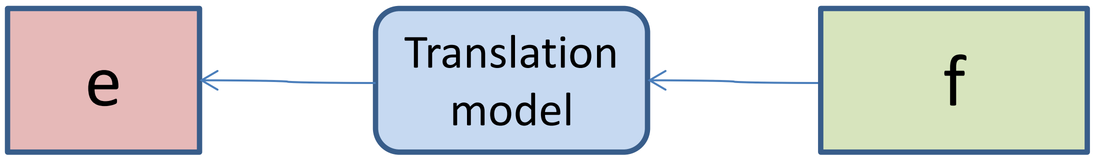
- An original message went through a distortion model (noisy channel) we get only the “distorted” message.



- Our goal: given a translation and a translation model, find the most probable original



Noisy channel model



We observe *f* (a foreign string). What would be the best *e* English sentence to generate *f* ?

$$\tilde{e} = \arg\max_{e \in e^*} p(e|f)$$

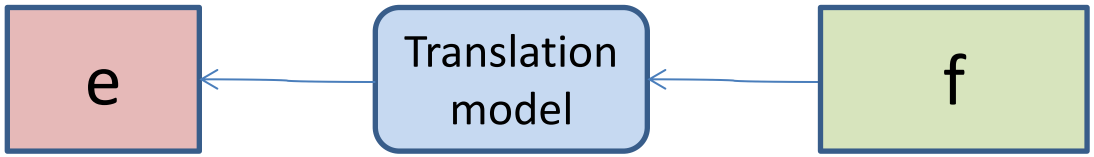
Knowing Bayes rule:

$$p(e|f) = \frac{p(f|e)p(e)}{p(f)}$$

- The best *e* would be:

$$\tilde{e} = \arg\max_{e \in e^*} p(f|e)p(e)$$

Noisy channel model



$$\tilde{e} = \arg\max_{e \in e^*} p(f|e)p(e)$$

- $p(f|e)$ ➔ Phrase table (phrase translation probabilities)

  This will be "learned" out of parallel sentences

- $p(e)$ ➔ Language model (probability of being "good" English)

  This will be "learned" out of English texts

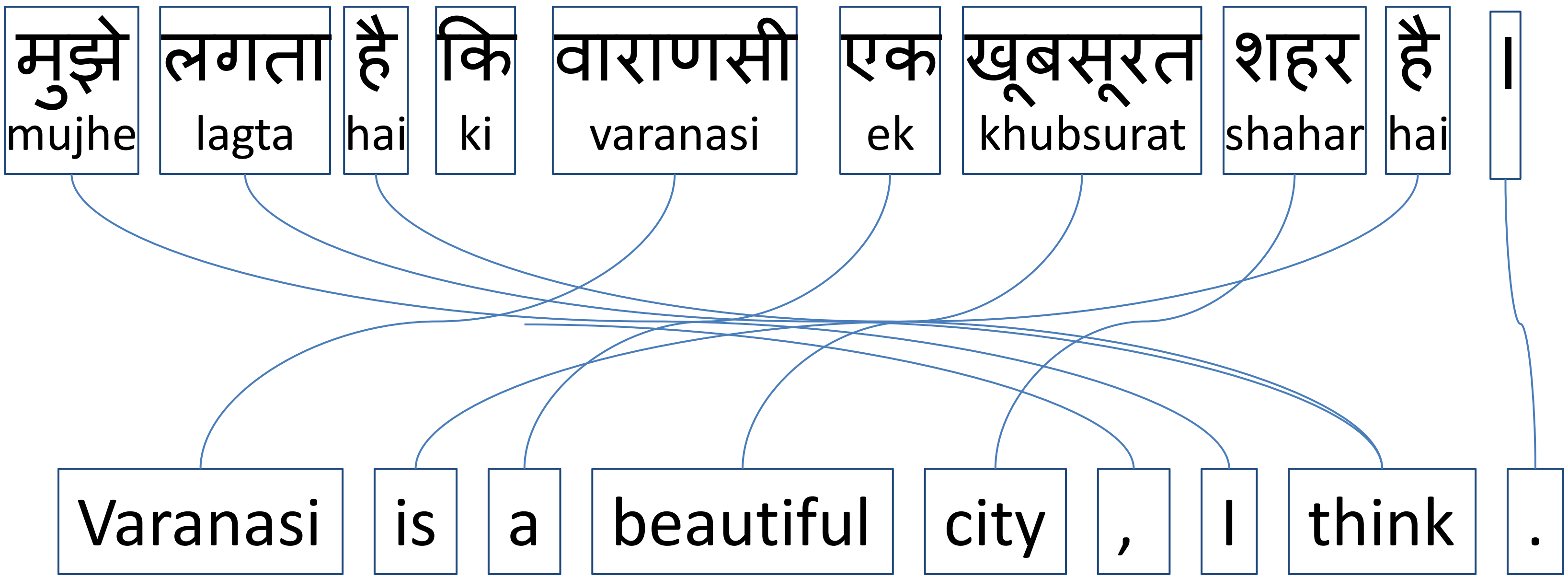
Linear model

$$\tilde{e} = \arg\max_{e \in e^*}(w_t T . w_l L . w_d D . w_s S)$$

- T ➔ translation probability
- L➔Language model probability
- D➔Distortion probability
- S➔Sentence length probability
- Etc.
- $\tilde{e} = \arg\max_{e \in e^*}(\prod_{i=1}^{n} w_i . H_i)$
- $\tilde{e} = \arg\max_{e \in e^*}(\sum_{i=1}^{n} \log w_i \times H_i)$

$$\arg\max_{e} score(e \mid f) = \sum_{i=1}^{n} w_i \cdot h_{e;i}$$

Word-to-word translation model



- **(word) translation probability**
- (language model probability)
- (sentence length model)
- (distortion probability)

Model Estimation (for now, only translation probabilities)

- Word translation probabilities are easy to estimate from word alignment links:

$$t(e \mid f) = \frac{count(f \rightarrow e)}{\sum_{\hat{e}} count(f \rightarrow \hat{e})}$$

Credits: U. Germann U. Edinburgh

Model Estimation (for now, only translation probabilities)

- Word translation probabilities are easy to estimate from word alignment links:

$$t(e \| f) = \sum_{\vec{e},\vec{f} \in Corpus} \sum_{i=1;e_i=e}^{|\vec{e}|} \sum_{k=0;f_k=f}^{|\vec{k}|} p(a_i = k \mid \vec{e}, \vec{f})$$

- Word translation probabilities can be inferred from word translation probabilities:

$$p(a_i = k) = \frac{t(e_i \mid f_k)}{\sum_{\hat{k}} t(e_i \mid f_{\hat{k}})}$$

Credits: U. Germann U. Edinburgh

IBM model 1

Model 1

- uniform sentence length probability
- uniform distortion probability

$$p(\vec{e} \mid \vec{f}) = \epsilon \sum_{\vec{a}} \prod_{i=1}^{|\vec{e}|} t(e_i \mid f_{a_i})$$

Credits: U. Germann U. Edinburgh

IBM model 2

Adds distortion probability in the model

- uniform sentence length probability
- distortion probability based on absolute positions within the sentence *d(k | i)*.
- word translation probabilities as in Model 1

Credits: U. Germann U. Edinburgh

IBM models 3 to 5

New generative story

- for each source word f_k pick a fertility n_k with probability $p(n_k \mid f)$.
- copy f_k n_k times
- translate each copy according to $t(e_{k:j} \mid f_k)$
- place translations into target sentence
- **Model 3**: distortion probabilities based on absolute positions
- **Model 4**: distortion probabilities based on positions relative to the target positions of previously placed word(s)
- **Model 5**: eliminates a deficiency of Models 3 and 4; not used in practice.

Credits: U. Germann U. Edinburgh

Nota bene:

- From Model 3, on, individual word translations are not independent of one another any more (because of fertility, relative distortions)!
- full marginalization $\sum_{\vec{a}} p(\vec{e}, \vec{a} \mid \vec{f})$ is too expensive
- initialize *Viterbi Alignment* from lower Model, consider only neighboring alignments during training

Credits: U. Germann U. Edinburgh

Hidden Markov Models for Alignment

- source words f are hidden states
- emit target words according to $t(e \mid f)$
- distortion modeled via transition probabilities between states of Hidden Markov Model
- replaces Model 2 in the standard Giza++ setup

Credits: U. Germann U. Edinburgh

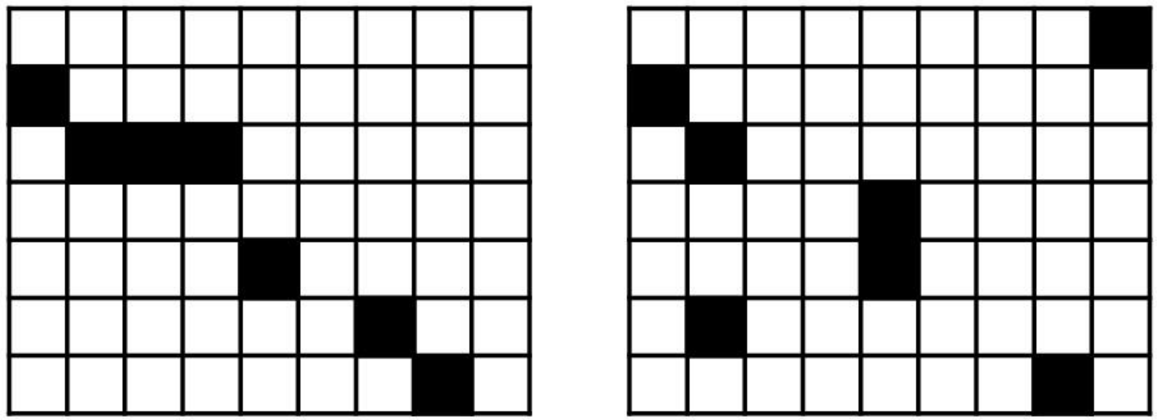
Examples of lexical table generated

On Hindi→English

cat lex.e2f| grep ‘ world '|sort -k 3n|tail

Hindi	***Probability***
विश्वभर	0.0084
विश्वविख्यात	0.0084
विश्व	0.0108
चिह्नाजगत	0.0120
जगत	0.0240
वर्ल्ड	0.0335
विश्वयुद्ध	0.0479
संसार	0.0719
दुनिया	0.0898
विश्व	0.4982

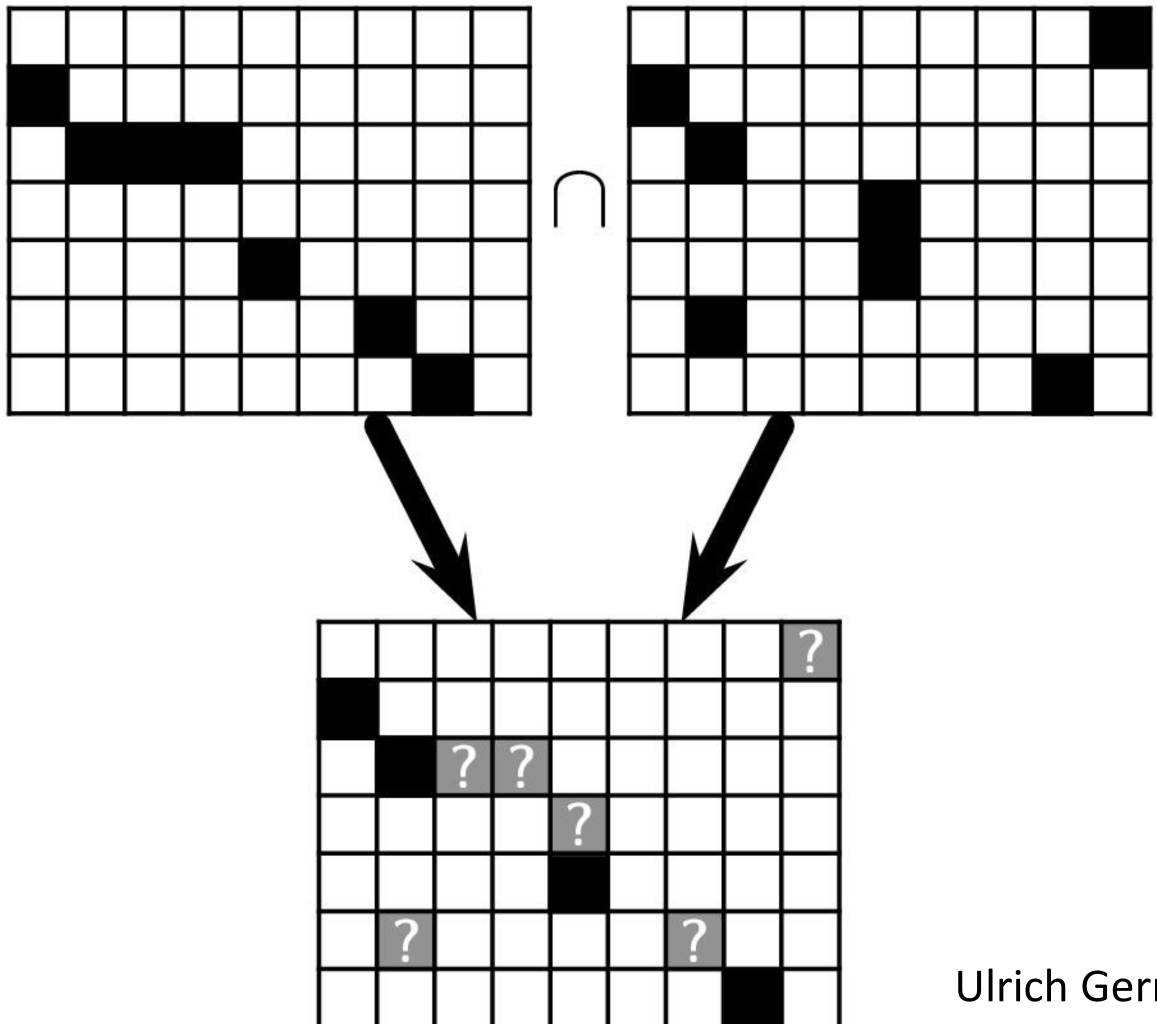
Alignment symmetrisation grow-diag+final-and



Ulrich Germann U. Edinburgh

Alignment symmetrisation grow-diag+final-and

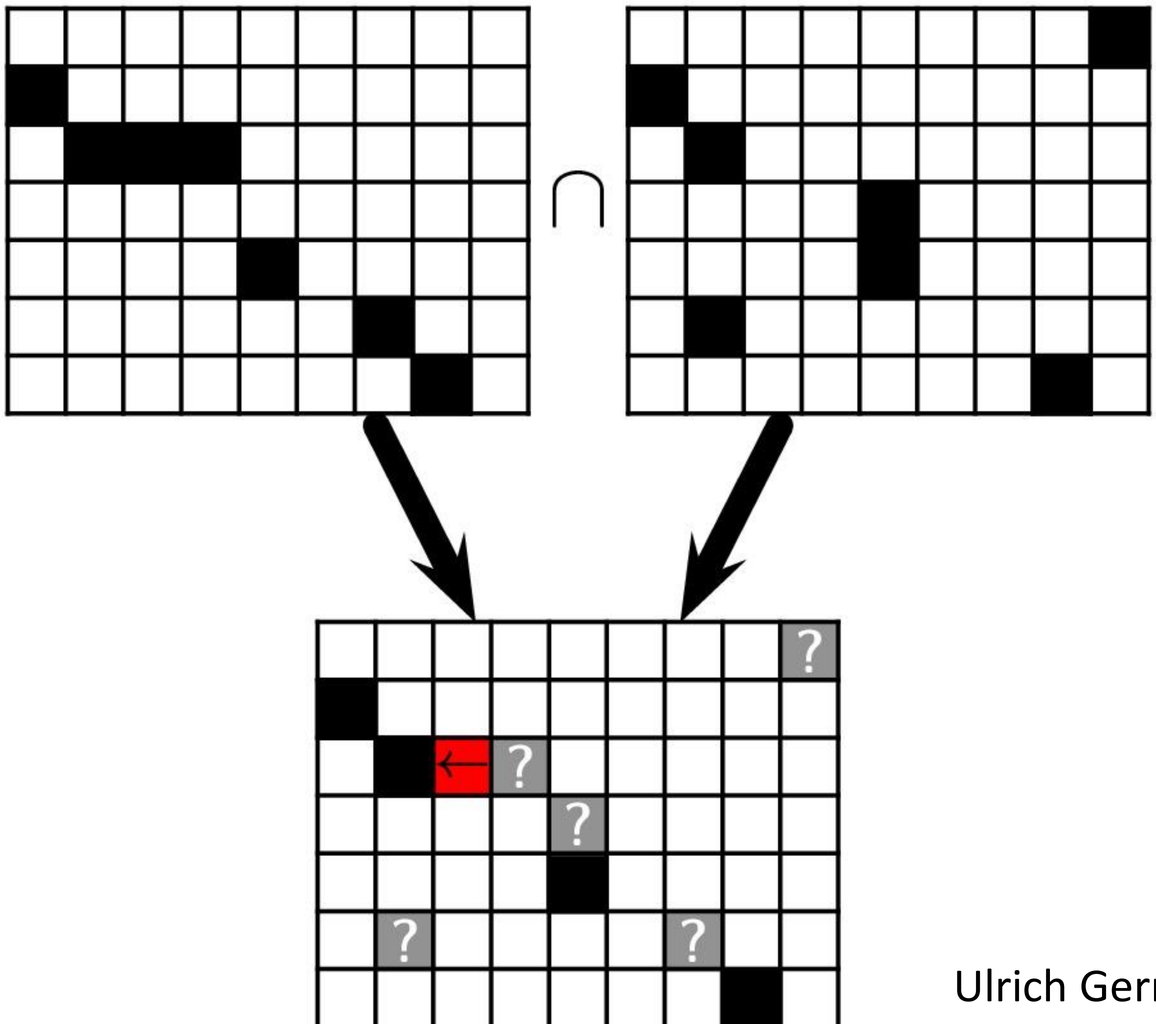
Step 1: Intersect the two alignments:



Ulrich Germann U. Edinburgh

Alignment symmetrisation grow-diag+final-and

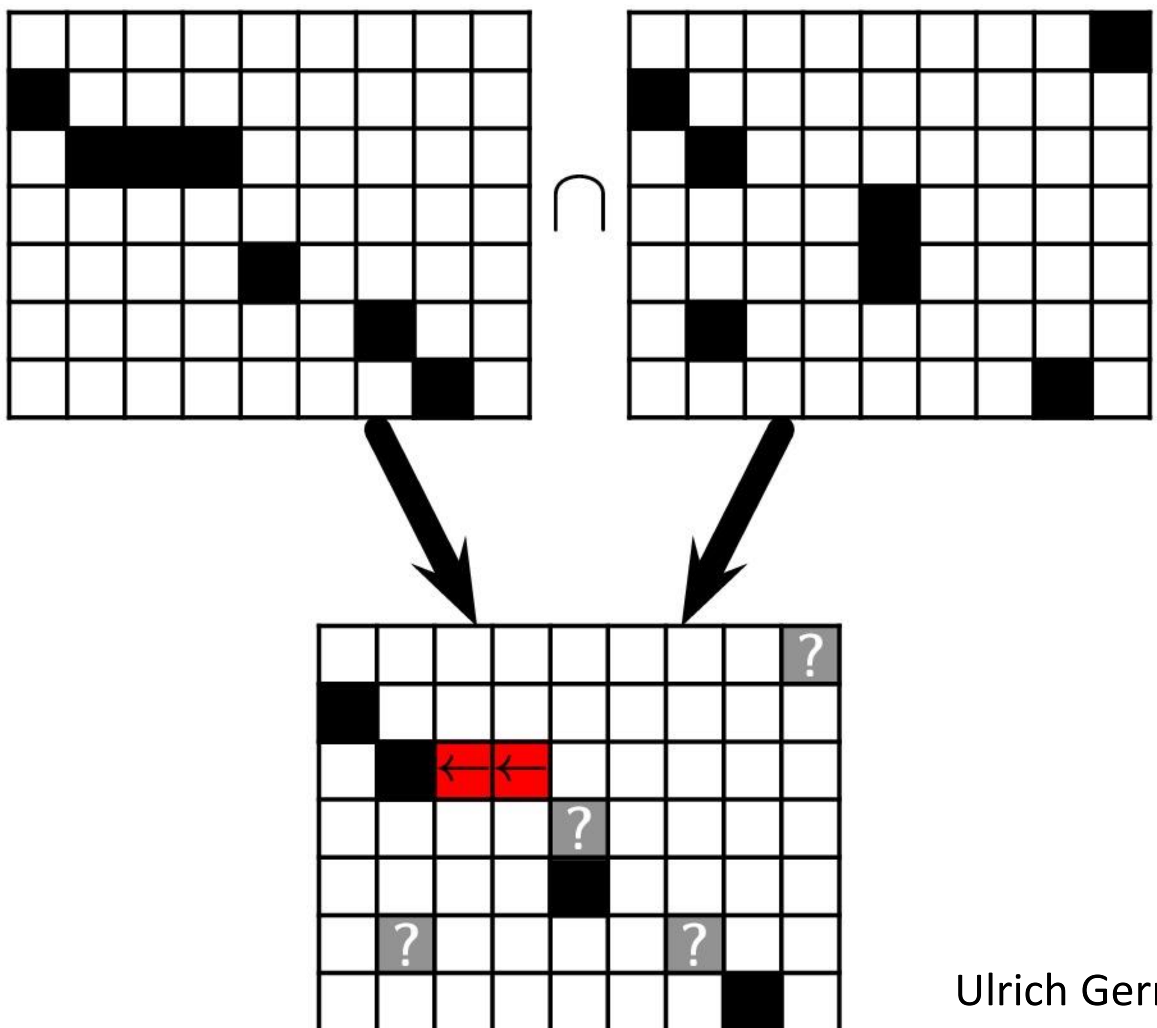
Step 1: Intersect the two alignments:



Ulrich Germann U. Edinburgh

Alignment symmetrisation grow-diag+final-and

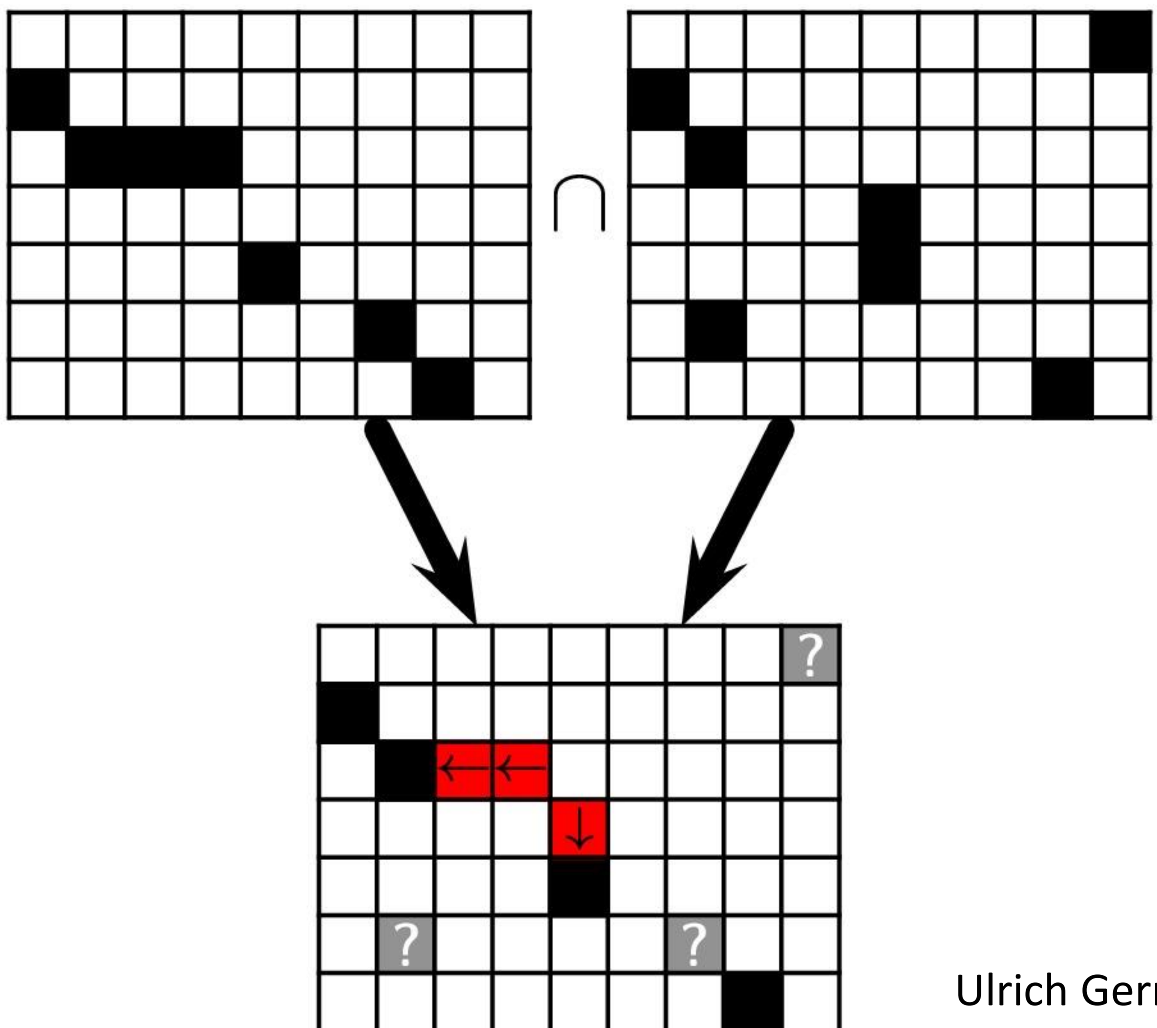
Step 1: Intersect the two alignments:



Ulrich Germann U. Edinburgh

Alignment symmetrisation grow-diag+final-and

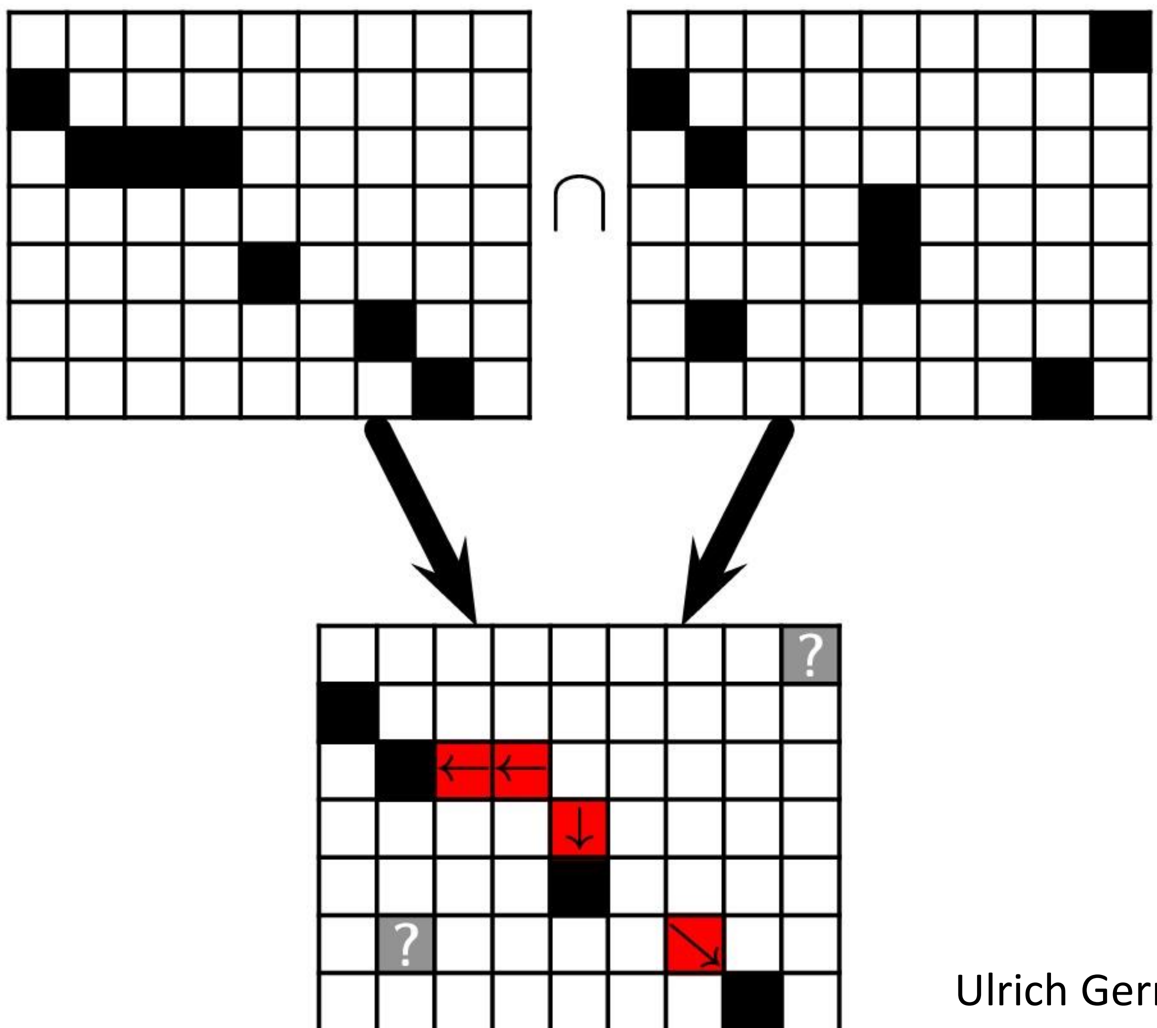
Step 1: Intersect the two alignments:



Ulrich Germann U. Edinburgh

Alignment symmetrisation grow-diag+final-and

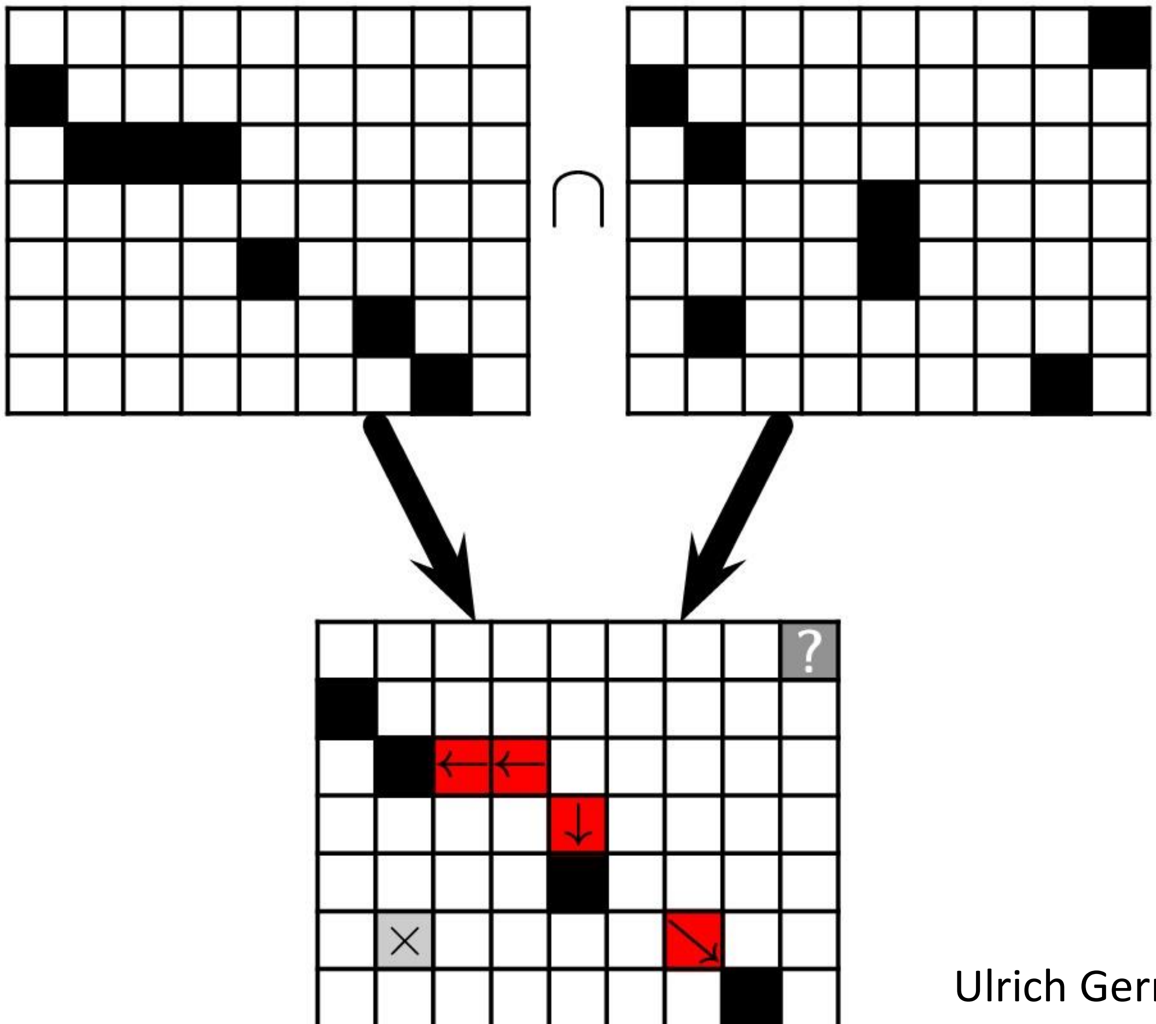
Step 1: Intersect the two alignments:



Ulrich Germann U. Edinburgh

Alignment symmetrisation grow-diag+final-and

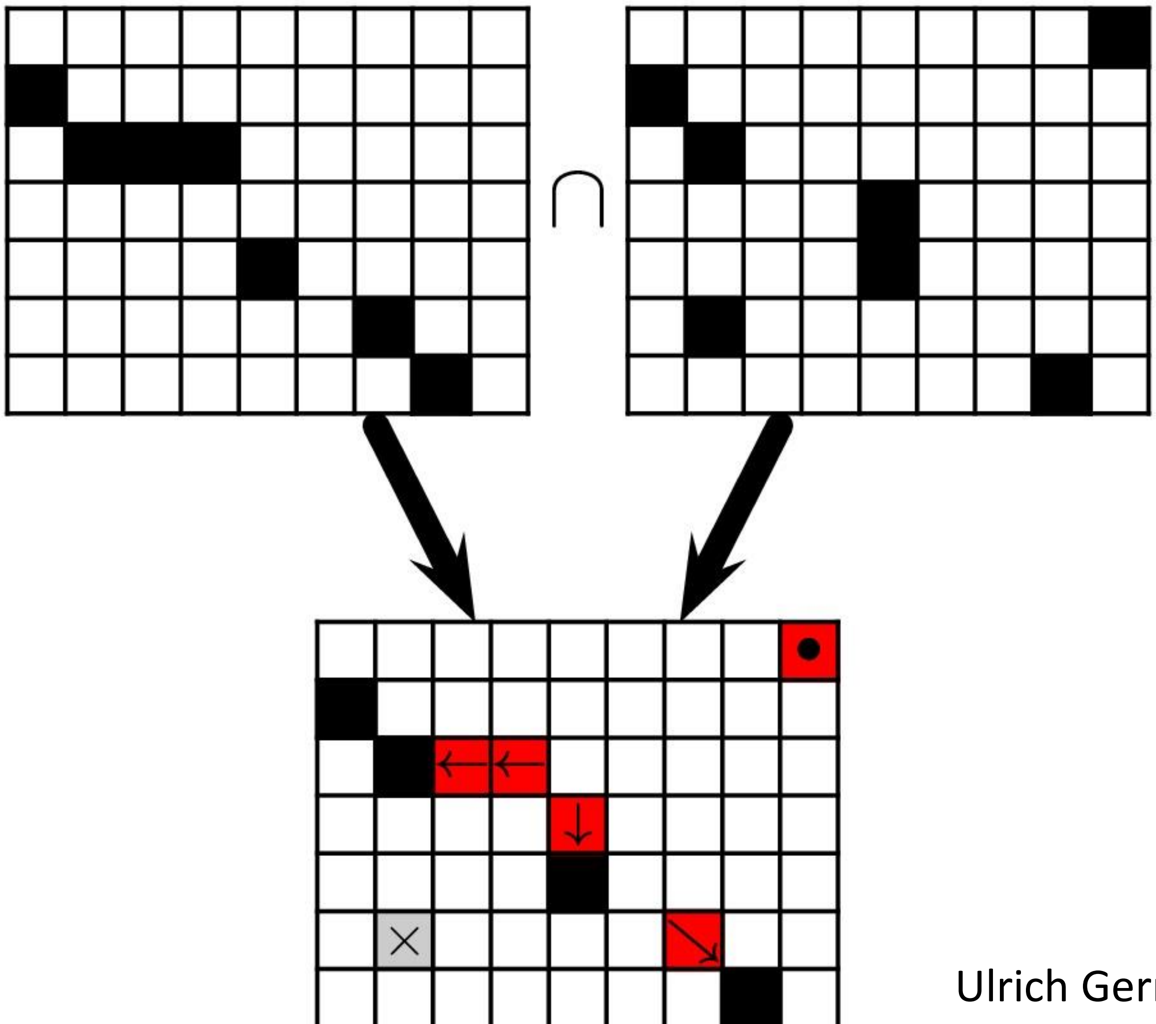
Step 1: Intersect the two alignments:



Ulrich Germann U. Edinburgh

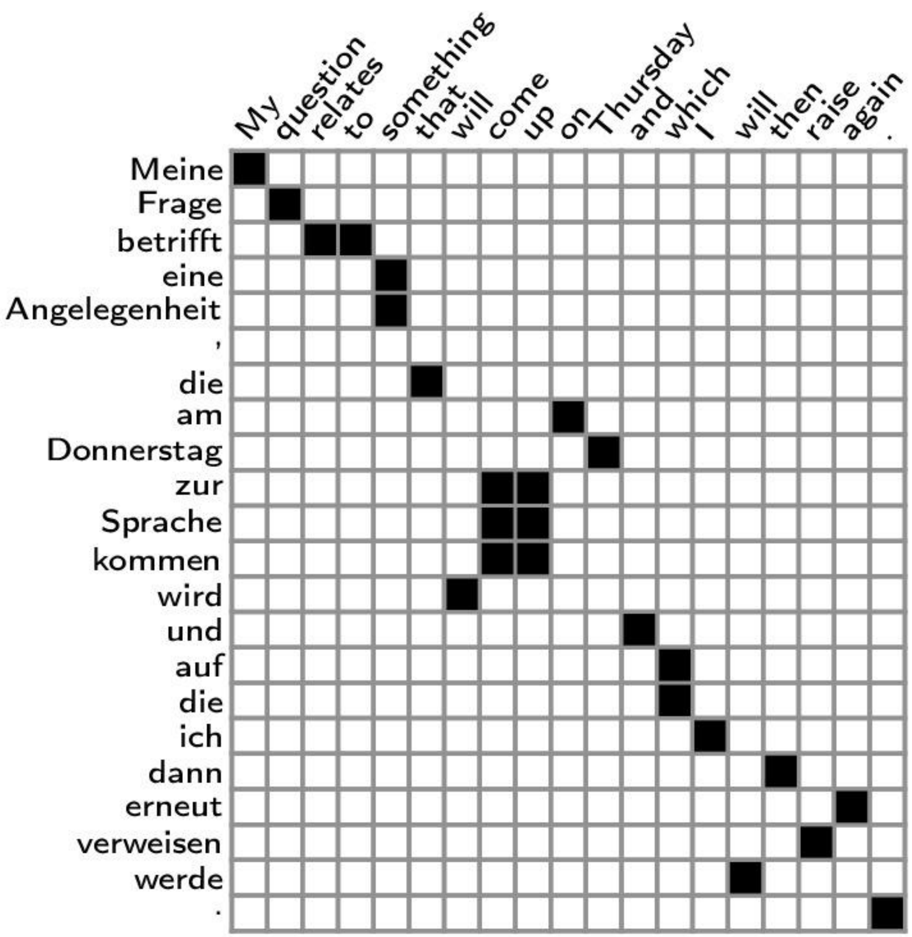
Alignment symmetrisation grow-diag+final-and

Step 1: Intersect the two alignments:



Ulrich Germann U. Edinburgh

Building phrase tables

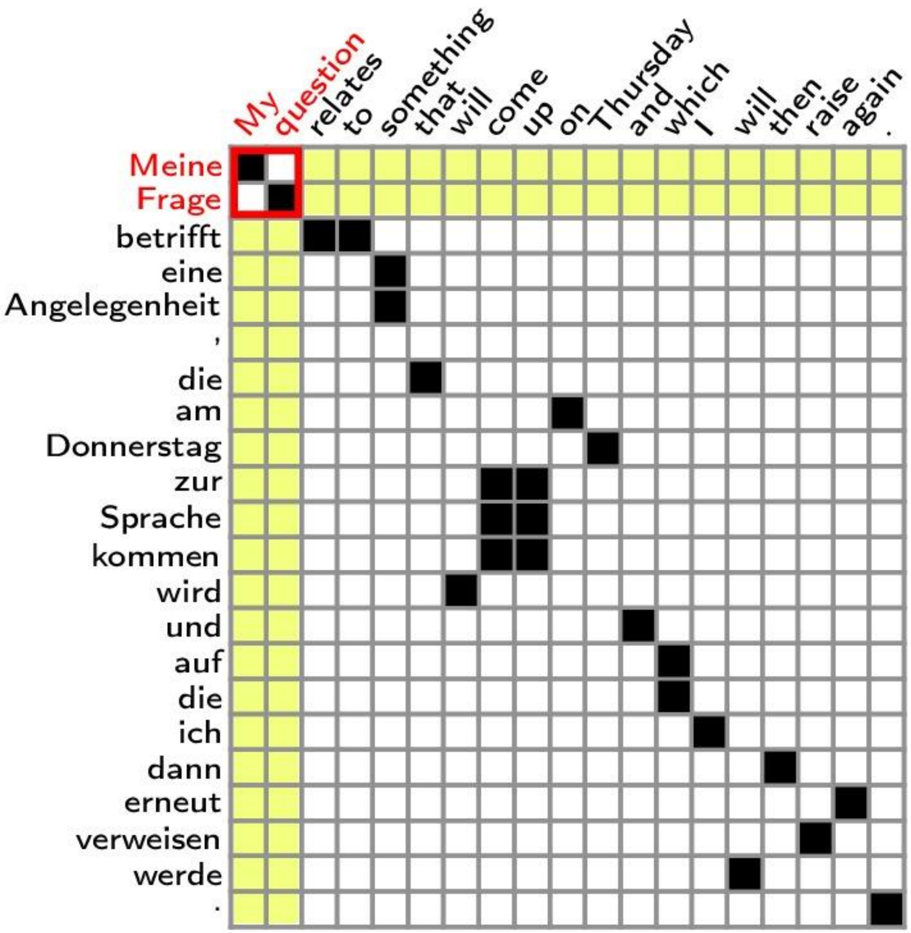


Phrase Table

meine	⇔	my
Frage	⇔	question

Ulrich Germann U. Edinburgh

Building phrase tables

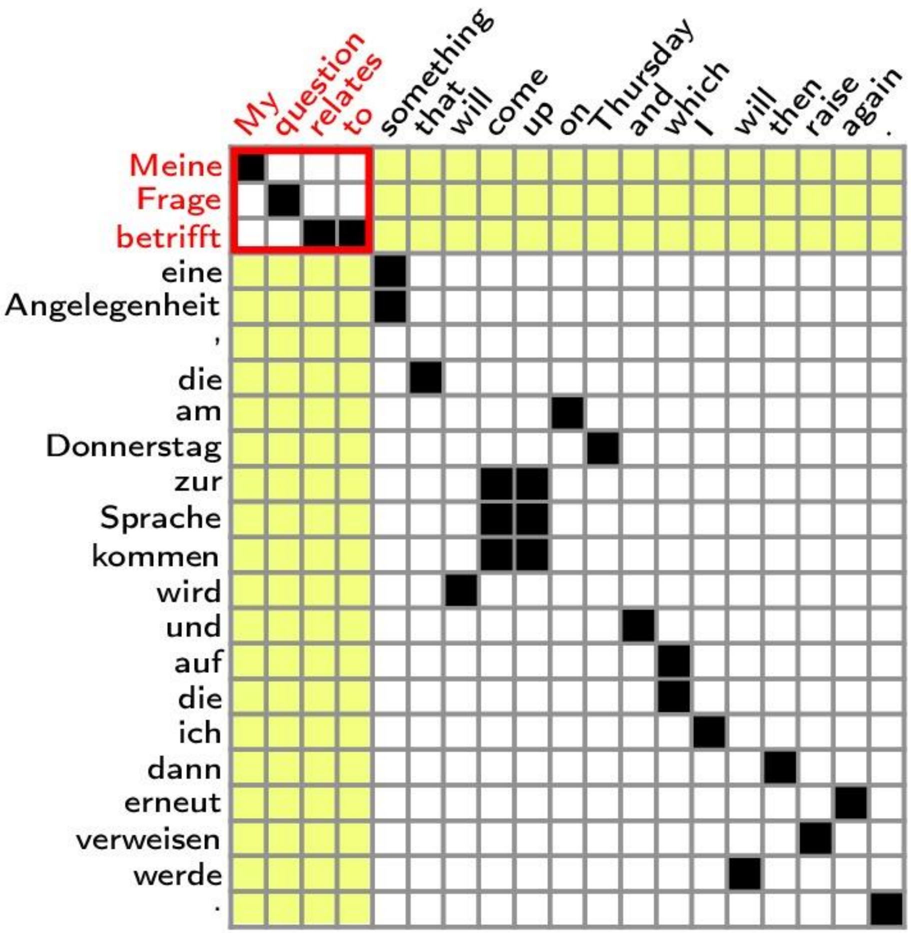


Phrase Table

meine	⇔	my
meine Frage	⇔	my question

Ulrich Germann U. Edinburgh

Building phrase tables

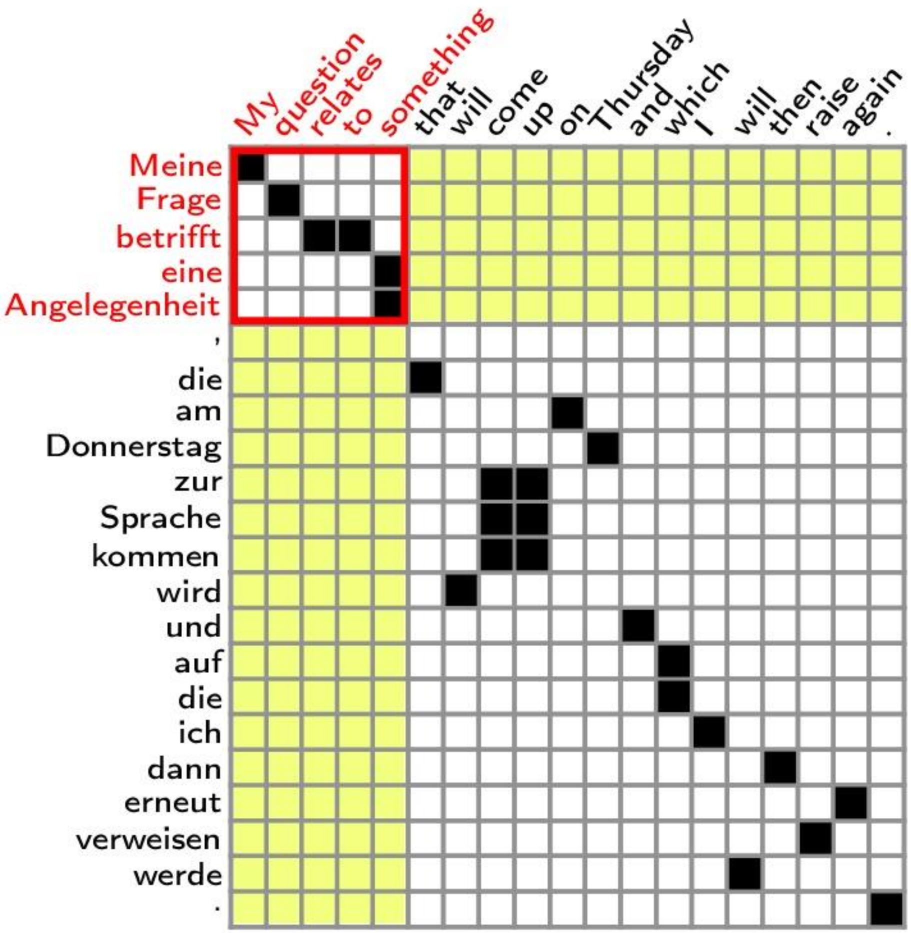


Phrase Table

meine	⇔	my
meine Frage	⇔	my question
meine Frage betrifft	⇔	my question relates to

Ulrich Germann U. Edinburgh

Building phrase tables

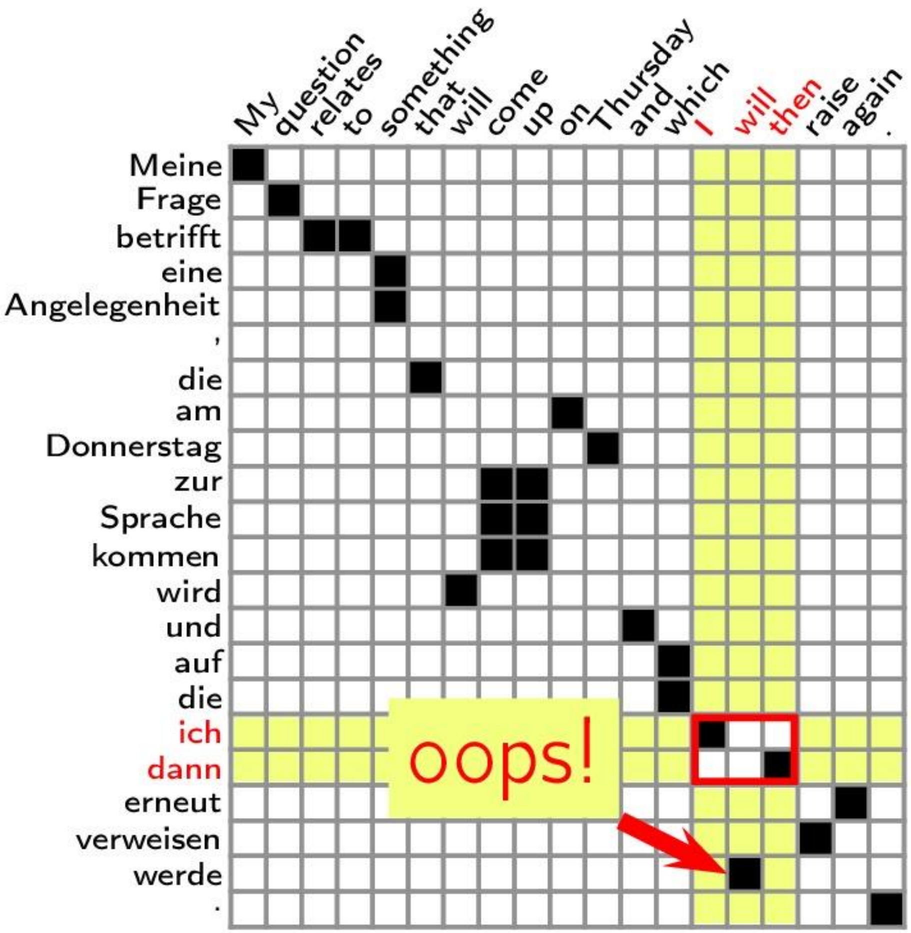


Phrase Table

meine	⇔	my
meine Frage	⇔	my question
meine Frage betrifft	⇔	my question relates to
meine Frage betrifft eine Angelegenheit	⇔	my question relates to something

Ulrich Germann U. Edinburgh

Building phrase tables



Phrase Table

meine	⇔	my
meine Frage	⇔	my question
meine Frage betrifft	⇔	my question relates to
meine Frage betrifft	⇔	my question relates to
eine Angelegenheit		something
⋮		⋮
Frage	⇔	question
Frage betrifft	⇔	question relates to
⋮		⋮
~~ich dann~~	⇔	~~I will then~~
⋮		⋮

Ulrich Germann U. Edinburgh

Scoring phrase table entries

- Weighted linear combination of features

$$P_{TM}(t \mid s) = exp\left(\sum_j \alpha_j f_j(s,t)\right)$$

Ulrich Germann U. Edinburgh

Scoring phrase table entries

- log of smoothed **forward** cond. prob.:

$$smooth\left(\frac{count\,(target\ phrase)}{count\,(source\ phrase)}\right)$$

- log of smoothed backward cond. prob.:

$$smooth\left(\frac{count\,(source\ phrase)}{count\,(target\ phrase)}\right)$$

- “lexically smoothed” (Zens&Ney) **forward** probability

$$\sum\nolimits_t \log P\,(t \mid source\ phrase[, alignment])$$

- “lexically smoothed” **backward** probability

$$\sum\nolimits_s \log P\,(s \mid target\ phrase[, alignment])$$

- length of target phrase (“word penalty”)
- 1 (“phrase penalty”)

Ulrich Germann U. Edinburgh

Example of scores

Look at generated “phrase-table” file

किताब ||| book ||| 0.0590406 0.0469136 0.761905 0.730769 ||| 0-0
पुस्तक ||| book ||| 0.752768 0.528395 0.842975 0.732877 ||| 0-0
पवित्र पुस्तक ||| holy book ||| 1 0.140652 1 0.233762 ||| 0-0 1-1

Scoring translation hypotheses

Log-linear combination of:

Translation Model	assesses the quality of phrase-level translations.
Distortion Model	evaluates jumps between source phrases.
Language Model	evaluates the fluency of the translation hypothesis

$$P\left(\mathit{translation} \mid \mathit{source}\right) = \exp\left(\begin{array}{cl} & \alpha_{TM} \log P_{TM}(\mathit{translation} \mid \mathit{source}) \\ + & \alpha_{DM} \log P_{DM}(\mathit{translation} \mid \mathit{source}) \\ + & \alpha_{LM} \log P_{LM}(\mathit{translation} \mid \mathit{source}) \end{array} \right)$$

Ulrich Germann U. Edinburgh

Scoring translation hypotheses

November inflation rates were higher than expected in the 13 countries of the eurozone .

Ulrich Germann U. Edinburgh

Scoring translation hypotheses

November $\boxed{\text{inflation rates}}$ were higher than expected in the 13 countries of the euro...

Teuerungsraten

$\langle s \rangle$ Inflationsraten

...

$$p(t \mid i, \mathcal{M}_{tr}, \mathcal{M}_{lm}, \mathcal{M}_d) =$$

$$\exp\left(\alpha_{tr} \cdot \log p_{tr}(\text{Inflationsraten} \mid \text{inflation rates}) \quad + \alpha_{lm} \cdot \log p_{lm}(\text{Inflationsraten} \mid \langle s \rangle) \right)$$

Ulrich Germann U. Edinburgh

Scoring translation hypotheses

$\boxed{\text{November}}$ $\boxed{\text{inflation rates}}$ were higher than expected in the 13 countries of the euro. .

$\langle s \rangle$ Inflationsraten

$$p(t \mid i, \mathcal{M}_{tr}, \mathcal{M}_{lm}, \mathcal{M}_{d}) =$$

$$\exp\left(\begin{array}{l} \alpha_{tr} \cdot \log p_{tr}\,(\text{Inflationsraten} \mid \text{inflation rates}) \quad + \alpha_{lm} \cdot \log p_{lm}\,(\text{Inflationsraten} \mid \langle s \rangle) \\ + \alpha_{d} \cdot \log p_{d}(-2) \end{array} \right)$$

Ulrich Germann U. Edinburgh

Scoring translation hypotheses

$\boxed{\text{November}}$ inflation rates were higher than expected in the 13 countries of the euro...

$\langle s \rangle$ Inflationsraten im November

$$p(t \mid i, \mathcal{M}_{tr}, \mathcal{M}_{lm}, \mathcal{M}_{d}) =$$

$$\exp\left(\begin{array}{ll} \alpha_{tr} \cdot \log p_{tr}(\text{Inflationsraten} \mid \text{inflation rates}) & + \alpha_{lm} \cdot \log p_{lm}(\text{Inflationsraten} \mid \langle s \rangle) \\ + \alpha_{d} \cdot \log p_{d}(-2) & \\ + \alpha_{tr} \cdot \log p_{tr}(\text{im November} \mid \text{November}) & + \alpha_{lm} \cdot \log p_{lm}(\text{im} \mid \ldots\ \text{Inflationsraten}) \\ & + \alpha_{lm} \cdot \log p_{lm}(\text{November} \mid \ldots\ \text{in}) \end{array}\right.$$

Ulrich Germann U. Edinburgh

Scoring translation hypotheses

November inflation rates | were higher than | | expected in the | 13 countries of the euro. .

⟨s⟩ Inflationsraten im November waren höher als erwartet in den

$$p(t \mid i, \mathcal{M}_{tr}, \mathcal{M}_{lm}, \mathcal{M}_{d}) =$$

$$\exp\left(\begin{array}{ll}
\alpha_{tr} \cdot \log p_{tr}\,(\text{Inflationsraten} \mid \text{inflation rates}) & + \alpha_{lm} \cdot \log p_{lm}\,(\text{Inflationsraten} \mid \langle s \rangle) \\
+ \alpha_{d} \cdot \log p_{d}(-2) & \\
+ \alpha_{tr} \cdot \log p_{tr}\,(\text{im November} \mid \text{November}) & + \alpha_{lm} \cdot \log p_{lm}\,(\text{im} \mid \ldots\ \text{Inflationsraten}) \\
 & + \alpha_{lm} \cdot \log p_{lm}\,(\text{November} \mid \ldots\ \text{in}) \\
+ \alpha_{d} \cdot \log p_{d}(+3) & \\
+ \alpha_{tr} \cdot \log p_{tr}\,(\text{waren} \ldots \text{als} \mid \text{were} \ldots \text{than}) & + \alpha_{lm} \cdot \log p_{lm}\,(\text{waren} \mid \ldots\ \text{November}) \\
 & + \alpha_{lm} \cdot \log p_{lm}\,(\text{höher} \mid \ldots\ \text{waren}) \\
 & + \alpha_{lm} \cdot \log p_{lm}\,(\text{als} \mid \ldots\ \text{höher}) \\
\ldots &
\end{array}\right.$$

Ulrich Germann U. Edinburgh

Scoring translation hypotheses

November inflation rates were higher than $\boxed{\text{expected in the}}$ $\boxed{\text{13 countries}}$ of the euro. .

$\langle s \rangle$ Inflationsraten im November waren höher als erwartet in den 13 Ländern

$$p(t \mid i, \mathcal{M}_{tr}, \mathcal{M}_{lm}, \mathcal{M}_{d}) =$$

$$\exp \left(\begin{array}{ll} \alpha_{tr} \cdot \log p_{tr}\,(\text{Inflationsraten} \mid \text{inflation rates}) & + \alpha_{lm} \cdot \log p_{lm}\,(\text{Inflationsraten} \mid \langle s \rangle) \\ + \alpha_{d} \cdot \log p_{d}(-2) & \\ + \alpha_{tr} \cdot \log p_{tr}\,(\text{im November} \mid \text{November}) & + \alpha_{lm} \cdot \log p_{lm}\,(\text{im} \mid \ldots\ \text{Inflationsraten}) \\ & + \alpha_{lm} \cdot \log p_{lm}\,(\text{November} \mid \ldots\ \text{in}) \\ + \alpha_{d} \cdot \log p_{d}(+3) & \\ + \alpha_{tr} \cdot \log p_{tr}\,(\text{waren} \ldots \text{als} \mid \text{were} \ldots \text{than}) & + \alpha_{lm} \cdot \log p_{lm}\,(\text{waren} \mid \ldots\ \text{November}) \\ & + \alpha_{lm} \cdot \log p_{lm}\,(\text{höher} \mid \ldots\ \text{waren}) \\ & + \alpha_{lm} \cdot \log p_{lm}\,(\text{als} \mid \ldots\ \text{höher}) \\ \ldots & \end{array} \right.$$

Ulrich Germann U. Edinburgh

Scoring translation hypotheses

. . . inflation rates were higher than expected in the $\boxed{\text{13 countries}}$ $\boxed{\text{of the eurozone .}}$

⟨s⟩ Inflationsraten . . . waren höher als erwartet in den 13 Ländern der Eurozone .

$$p(t \mid i, \mathcal{M}_{tr}, \mathcal{M}_{lm}, \mathcal{M}_{d}) =$$

$$\exp\left(\begin{array}{ll}
\alpha_{tr} \cdot \log p_{tr}\,(\text{Inflationsraten} \mid \text{inflation rates}) & + \alpha_{lm} \cdot \log p_{lm}\,(\text{Inflationsraten} \mid \langle s \rangle) \\
+ \alpha_{d} \cdot \log p_{d}(-2) & \\
+ \alpha_{tr} \cdot \log p_{tr}\,(\text{im November} \mid \text{November}) & + \alpha_{lm} \cdot \log p_{lm}\,(\text{im} \mid \ldots \text{ Inflationsraten}) \\
 & + \alpha_{lm} \cdot \log p_{lm}\,(\text{November} \mid \ldots \text{ in}) \\
+ \alpha_{d} \cdot \log p_{d}(+3) & \\
+ \alpha_{tr} \cdot \log p_{tr}\,(\text{waren} \ldots \text{ als} \mid \text{were} \ldots \text{ than}) & + \alpha_{lm} \cdot \log p_{lm}\,(\text{waren} \mid \ldots \text{ November}) \\
 & + \alpha_{lm} \cdot \log p_{lm}\,(\text{höher} \mid \ldots \text{ waren}) \\
 & + \alpha_{lm} \cdot \log p_{lm}\,(\text{als} \mid \ldots \text{ höher}) \\
\ldots &
\end{array}\right.$$

Distortion modeling (reordering)

- Exponential probability decay over distance:

$$p_d(x) = \gamma^{\mathrm{abs}(x)}$$

- Lexicalized discrete model (Koehn et al., 2005)
  - Estimated separately for each phrase.
  - Three types of *type(j)* of jumps:
    - **mono** phrase immediately follows the previously translated phrase
    - **swap** phrase swaps positions with the previously translated phrase
    - **other** anything else
- ...

Ulrich Germann U. Edinburgh