- Jurafsky, D. and Martin, J. H. (2009): Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Pearson: New Jersey: Chapter 25
- Material von Bonnie Dorr's lecture
- Material from Kevin Knight's lecture at Berkeley, 2004

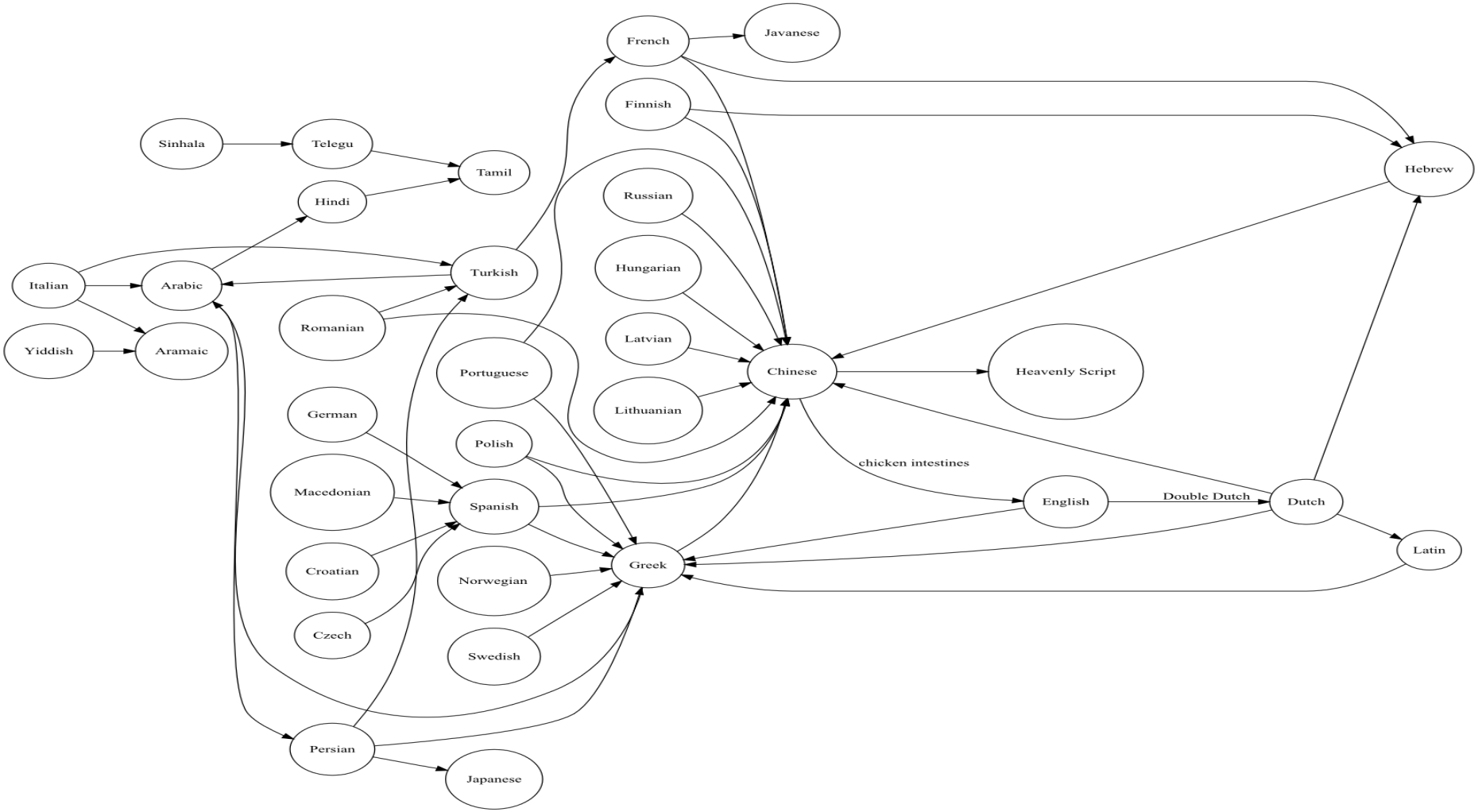
noisy channel model, word alignment, phrase-based translation

STATISTICAL MACHINE TRANSLATION

WHY IS MT CHALLENGING?

- Because languages are hard, even for humans!
- What is the German equivalent of “**It is all Greek to me”**?

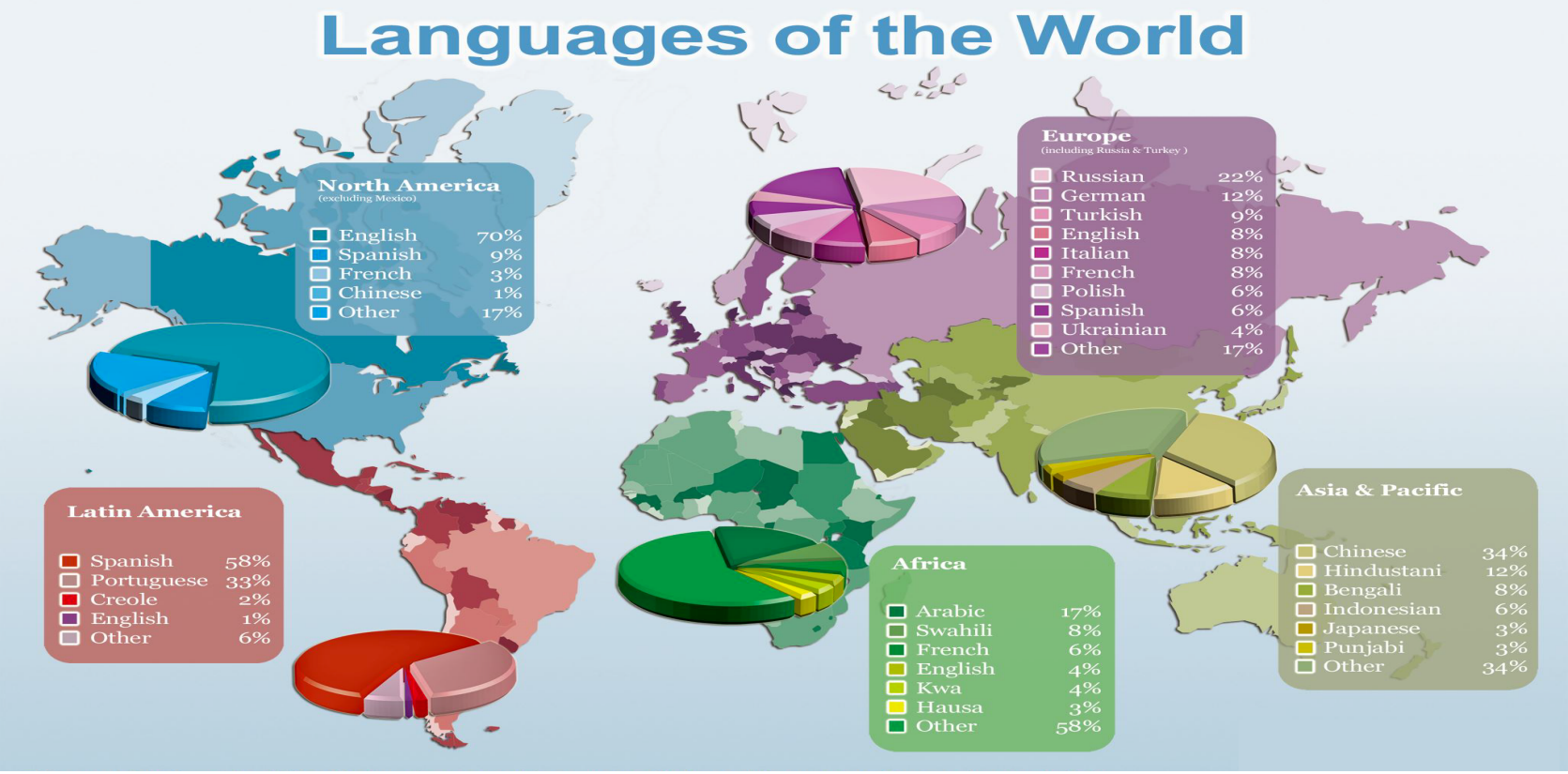
WHY IS MT CHALLENGING?



WHY IS MT CHALLENGING?

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

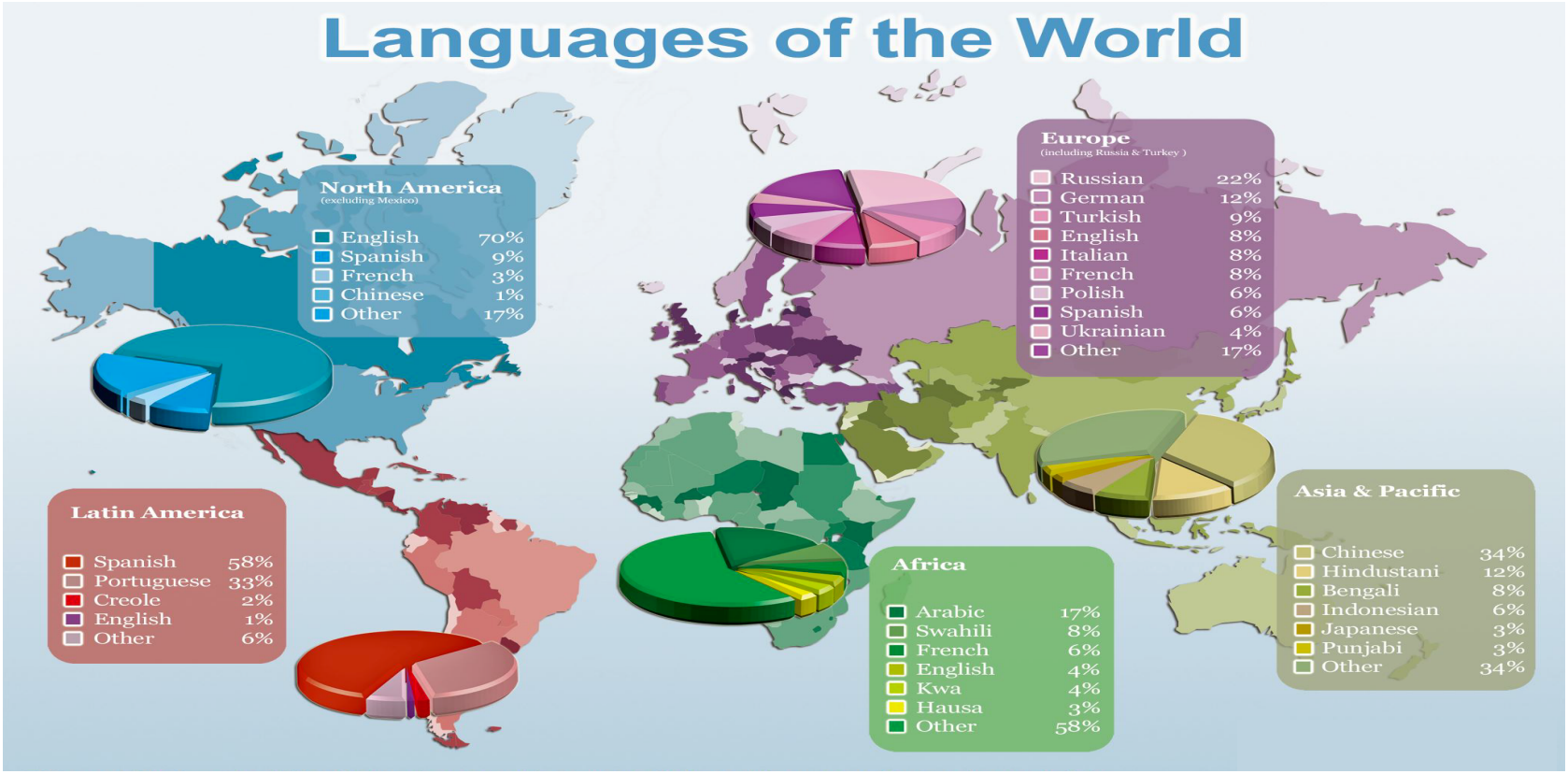
There are around 7,097 languages worldwide but 40% are threatened



WHY IS MT CHALLENGING?

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

95% of the population speak only 300 of them



MT EPIC FAILS

MT EPIC FAILS

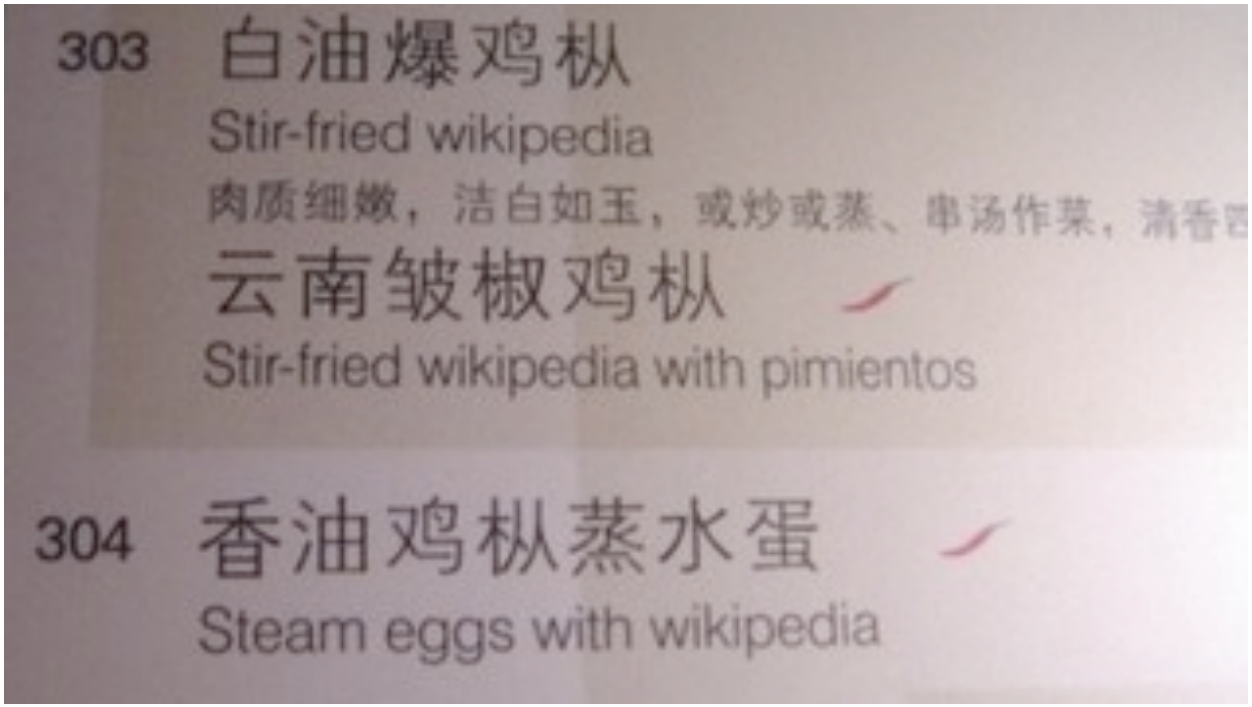
MT EPIC FAILS

MT EPIC FAILS

MT EPIC FAILS

The Arabic is a transliteration of Meat balls in Arabic alphabet, with its English literal translation as “Paul is dead”

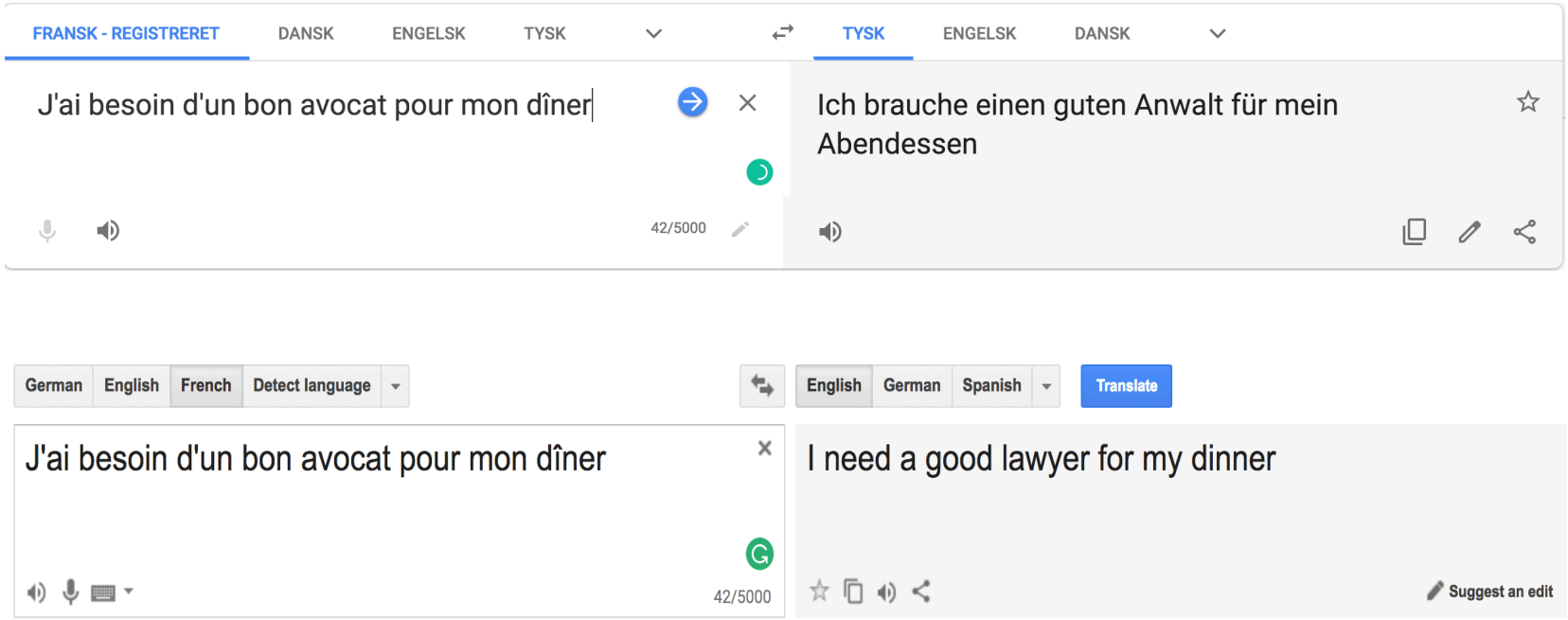
MT EPIC FAILS



MT EPIC FAILS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

And if you think this happens only for complicated languages such as Arabic and Chinese …



MACHINE TRANSLATION PARADIGMS

The are 4 MT Paradigms:

- Rule-based machine translation or RBMT
  - A RBMT system is based on linking the structure of the given input sentence with the structure of the foreign output sentence. To translate a sentence from French to English, one needs:
    - A dictionary that will map each French word to an appropriate English word.
    - Rules representing regular French sentence structure.
    - Rules representing regular English sentence structure.
    - Rules according to which one can relate these two structures together.
- Example-based machine translation or EBMT
  - For example, if we train on the two sentences **"Good postgraduate students in the CSE department are smart"** and **"Programmers are good people"** we would be able to translate **"Programmers in the CSE department are smart"** by just substituting the appropriate parts of the sentences.

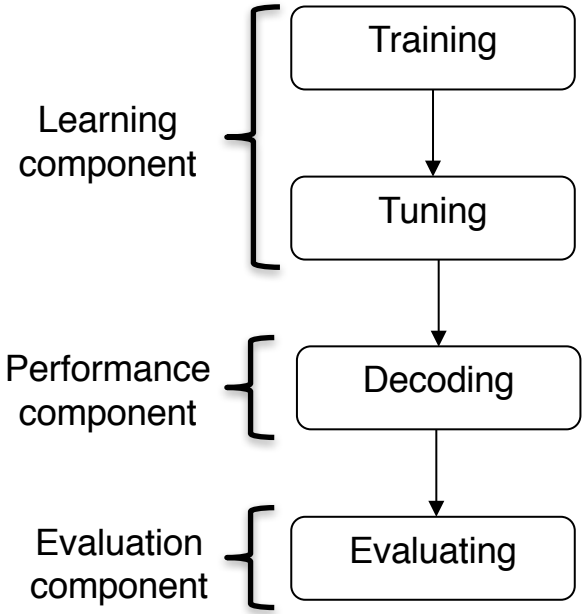
MACHINE TRANSLATION PARADIGMS

The are 4 MT Paradigms:

- Statistical machine translation or SMT
  - Mainly word or phrase-based translations
  - Translation are learned from actual data
  - Pro: Translations are learned automatically
  - Con: Difficult to model complex translation phenomena

- Neural Machine Translation or NMT
  - NMT is a simple new architecture for getting machines to learn to translate by modeling the whole translation process as one big artificial neural network
  - The most common models in NMT are based on the encoder-decoder architecture
  - First, the source sentence is encoded into a vector of embeddings of fixed length from which a decoder (decoder) generates translations.

MT PIPELINE IS SIMILAR TO MANY ML PIPELINES



PARALLEL CORPUS: TRAINING RESOURCE FOR MT

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Most popular:

- EuroParl: European parliament protocols in 11 languages
- Software manuals (KDE, Open Office …)
- Parallel webpages

For the remainder, we assume that we have a **sentence-aligned** parallel corpus.

- there are methods to get to aligned sentences from aligned documents
- there are methods to extract parallel sentences from comparable corpora

PARALLEL CORPUS: TRAINING RESOURCE FOR MT

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

41.	**Chapter 3, Stenmarck (SV)**	**fr**	**nl**
context	That is true **as long as account** is taken of the 20 per cent of the total postal services market where , in practice , there is still a monopoly , that is where the state is the only player .	C' est exact si l' on considère la question en tenant compte des 20 pour cent du marché total des services postaux où le monopole s' est maintenu dans la pratique , c'est-à-dire là où l' État est le seul acteur .	Dat klopt als men alleen kijkt naar 20% van de totale postmarkt , waar de staat in de praktijk nog steeds het monopolie heeft .
42.	**Chapter 3, MacCormick**	**fr**	**nl**
context	The Commission should not , for example , take a stepwise jump from 350 grammes to , as some have suggested , **as low as 50 grammes** .	Par exemple , la Commission devrait éviter de passer de 350g à 50g , comme l' ont suggéré certains .	De Commissie moet bijvoorbeeld niet helemaal van 350 gram naar 50 gram gaan zakken , zoals sommigen hebben geopperd .

COMPUTING TRANSLATION PROBABILITIES

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Imagine that we want to translate from French (f) into English (e).

- Given a parallel corpus we can estimate *P(e | f).* The maximum likelihood estimation of *P(e | f)* is: *freq(e,f)/freq(f)*
- Way too specific to get any reasonable frequencies when done on the basis of sentences, vast majority of unseen data will have zero counts
- *P(e | f)* could be re-defined as:

$$P(e \mid f) = \prod_{f^j} \max_{e^i} P(e^i \mid f^j)$$

- Problem: The English words maximizing *P(e | f)* might not result in a readable sentence

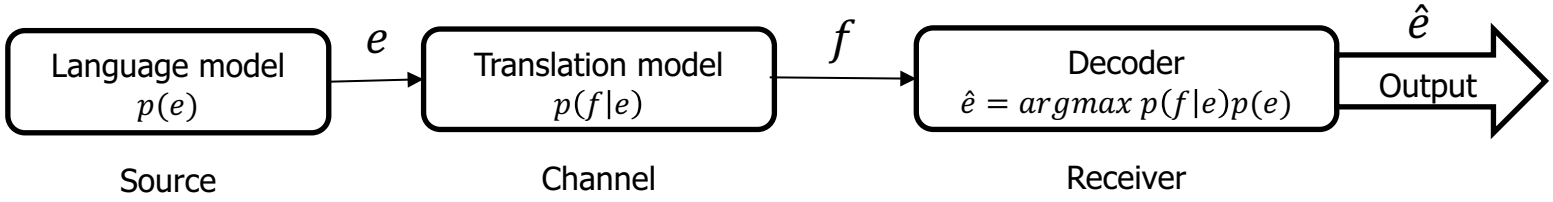
TRAINING PROBLEMS FOR STATISTICAL MT

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- **Translation model or $p(f|e)$**
  - assigns a conditional probability $p(f|e)$ to any foreign/English bisentences.
  - based on lexical translation and the notion of alignment
  - $p(avocat|lawyer)$ *or* $p(avocat|avocado)$
- **Language model or $p(e)$**
  - helps to ensure fluency
  - uses *n-gram* approximation: the next word can be predicted using a short history (one or two words)
    - $e = I\ ate\ an\ apple \longrightarrow p(e) = p(I) \times p(ate|I) \times p(an|I, ate) \times p(apple|ate, an)$

COMBINE THE TRANSLATION MODEL AND LM VIA THE NOISY CHANNEL APPROACH

- Noisy channel model using Bayes' theorem:



$$\hat{e} = \textbf{the optimal translation} = \underset{e}{argmax}\ p(e|f)$$

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

Bayes' inference

$$\hat{e} = \underset{e}{argmax}\ p(f|e)p(e)$$

the point estimate of the posterior — **likelihood** — **prior**

COMPUTING TRANSLATION PROBABILITIES

- We can account for adequacy: each foreign word translates into its most likely English word
- We cannot guarantee that this will result in a fluent English sentence
- Solution: transform *P(e | f)* with Bayes' rule:

$$P(e \mid f) = \frac{P(f \mid e) \cdot P(e)}{P(f)}$$

- *P(f | e)* accounts for adequacy
- *P(e)* accounts for fluency

LANGUAGE MODELING: P(E)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Determine the probability of an English sequence P(e)
- Can use n-gram models, PCFG-based models etc.: anything that assigns a probability for a sequence
- Standard: n-gram model

$$P(e) = P(e^1)P(e^2 \mid e^1)\prod_{i=3}^{I} P(e^i \mid e^{i-1}..e^{i-n+1})$$

- Language model picks the most fluent translation of many possible translations
- Language model can be estimated from a large monolingual corpus

TRANSLATION MODELING: P(F|E)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Determines the probability that the foreign word f_j is a translation of the English word e_i
- How to compute $P(f_j \mid e_i)$ from a parallel corpus? Need to **align** their translations
- Statistical approaches rely on the co-occurrence of e_i and f_j in the parallel data: If e_i and f_j tend to co-occur in parallel sentence pairs, they are likely to be translations of one another
- Commonly, four factors are used:
  - **translation:** How often do e_i and f_j co-occur?
  - **distortion:** How likely is a word occurring at position x to translate into a word occurring at position y? For example: English is a verb-second language, whereas German is a verb-final language
  - **fertility**: How likely is e_i to translate into more than one word? For example: “defeated” can translate into "eine Niederlage erleiden"
  - **null translation**: How likely is a foreign word to be spuriously generated?

IBM MODELS 1-5

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Model 1: Bag of words
  - Unique local maxima
  - Efficient EM algorithm (Model 1–2)
- Model 2: General alignment:

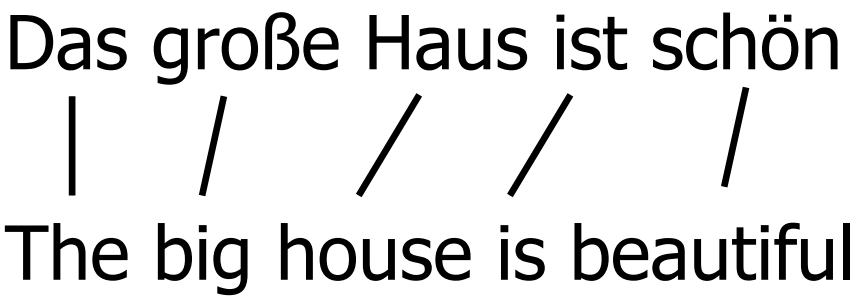
  $a(e^{pos} \mid f^{pos}, e_{length}, f_{length})$
- Model 3: fertility: $n(k \mid e)$
  - No full EM, count only neighbors (Model 3–5)
  - Leaky (Model 3–4)
- Model 4: Relative distortion, word classes
- Model 5: Extra variables to avoid leakiness

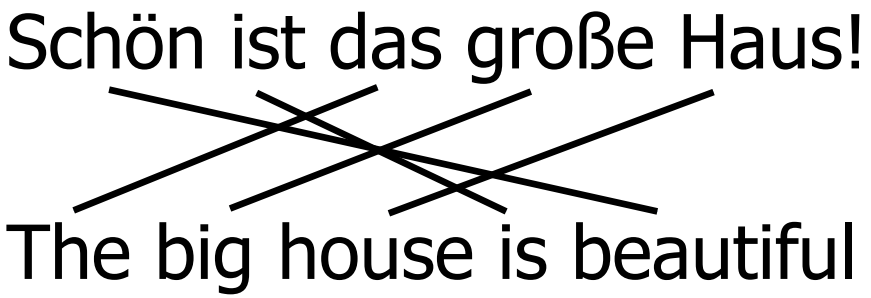
Models of lower orders are used to initialize models of higher orders

WORD BASED ALIGNMENT MODELS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Learn word-to-word correlation between the input/source language and and the output/target language



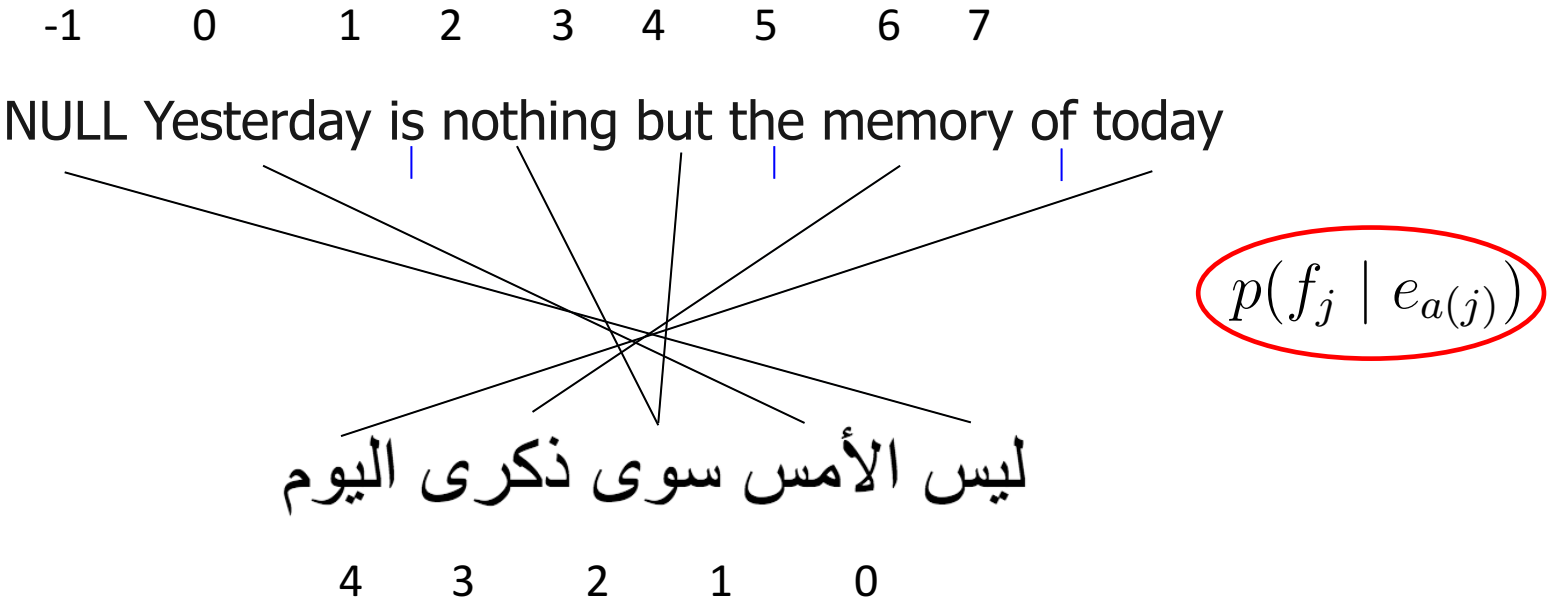


ALIGNMENT FUNCTION: DEFINITION

- Formalizing an alignment with an *alignment function* a
- Mapping a foreign input at position i with an English output at position j denoted i-j
- Example:

$$a: \{0\text{-}0, 1\text{-}1, 2\text{-}2\}$$

IBM MODEL 1


-1 0 1 2 3 4 5 6 7
NULL Yesterday is nothing but the memory of today
ليس الأمس سوى ذكرى اليوم
4 3 2 1 0
$p(f_j \mid e_{a(j)})$

IBM MODEL 1

- Simplest of the IBM models
- Does not consider word order (bag-of-words approach)
- Does not model one-to-many alignments
- Computationally inexpensive
- Useful for parameter estimations that are passed on to more elaborate models
- Translation probability in terms of alignments: $P(f \mid e) = \sum_{a \in A} P(f, a \mid e)$

  where: $P(f, a \mid e) = P(a \mid e) \cdot P(f \mid a, e)$

$$= \frac{1}{(l+1)^m} \prod_{j=1}^{m} P(f^j \mid e^{a_j})$$

and: $$P(f \mid e) = \sum_{a \in A} \frac{1}{(l+1)^m} \prod_{j=1}^{m} P(f^j \mid e^{a_j})$$

IBM MODELS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Given an English sentence $e^1 \ldots e^l$ and a foreign sentence $f^1 \ldots f^m$
- We want to find the 'best' alignment *a*, where *a* is a set of pairs of the form $\{(i, j), \ldots, (i', j')\}$, $0 \leq i, i' \leq l$ and $1 \leq j, j' \leq m$
- Note that if (i , j), (i', j) are in *a*, then i equals i', i.e. no many-to-one alignments are allowed
- We add a spurious NULL word to the English sentence at position 0
- In total there are $(l+1)^m$ different alignments *A*
- Allowing for many-to-many alignments results in $(2^l)^m$ possible alignments *A*

IBM MODEL 1

- We want to find the most likely alignment:

$$\underset{a \in A}{\operatorname{argmax}} \frac{1}{(l+1)^m} \prod_{j=1}^{m} P(f^j \mid e^{a_j}) = \underset{a \in A}{\operatorname{argmax}} \prod_{j=1}^{m} P(f^j \mid e^{a_j})$$

- We don't need to enumerate all alignments: Since $P(f^j|e^i)$ is independent from $P(f^{j'}|e^{i'})$ we can find the maximum alignment by looking at the individual translation probabilities only
- Let $\underset{a \in A}{\operatorname{argmax}} = (a_1, \ldots, a_m)$, then for each a_j: $a_j = \underset{0 \le i \le l}{\operatorname{argmax}} P(f^j \mid e^i)$

- The best alignment can be computed in a quadratic number of steps: m(l+1)

COMPUTING MODEL 1 PARAMETERS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Step 1: Determine candidates. For each English word e collect all foreign words f that co-occur at least once with e
- Step 2: Initialize P(f|e) uniformly, i.e. P(f|e) = 1/(number of co-occurring foreign words) or P(f|e) = 1/(number of total foreign words)

Step 3: Iteratively refine translation probabilities with EM:

```
for n iterations
     set tc to zero
     for each sentence pair (e,f) of lengths (l,m)
         for j=1 to m
             total=0;
             for i=1 to l: total += P(f^j|e^i);
             for i=1 to l: tc(f^j|e^i) += P(f^j|e^i)/total;
     for each word e
         total=0;
         for each word f s.t. tc(f|e) is defined:  total += tc(f|e);
         for each word f s.t. tc(f|e) is defined:  P(f|e) = tc(f|e)/total;
```

IBM MODEL 1 EXAMPLE

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Parallel ‘corpus’:

  the dog :: der Hund

  the tomcat :: der Kater

- Step 1+2 (collect candidates and initialize uniformly):

  P(der|the) = P(Hund|the) = P(Kater|the) = 1/3

  P(der|dog) = P(Hund|dog) = P(Kater|dog) = 1/3

  P(der|tomcat) = P(Hund|tomcat) = P(Kater|tomcat) = 1/3

  P(der|NULL) = P(Hund|NULL) = P(Kater|NULL) = 1/3

- Step 3: Iterate

NULL the dog :: der Hund

j=1

total = P(der|NULL)+P(der|the)+P(der|dog)= 1

tc(der|NULL) += P(der|NULL)/1 = 0 += .333/1 = 0.333

tc(der|the) += P(der|the)/1 = 0 += .333/1 = 0.333

tc(der|dog) += P(der|dog)/1 = 0 += .333/1 = 0.333

j=2

total = P(Hund|NULL)+P(Hund|the)+P(Hund|dog)=1

tc(Hund|NULL) += P(Hund|NULL)/1 = 0 += .333/1 = 0.333

tc(Hund|the) += P(Hund|the)/1 = 0 += .333/1 = 0.333

tc(Hund|dog) += P(Hund|dog)/1 = 0 += .333/1 = 0.333

IBM MODEL 1 EXAMPLE

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

NULL the tomcat :: der Kater

j=1

total = P(der|NULL)+P(der|the)+P(der|tomcat)=1

tc(der|NULL) += P(der|NULL)/1 = 0.333 += .333/1 = 0.666

tc(der|the) += P(der|the)/1 = 0.333 += .333/1 = 0.666

tc(der|tomcat) += P(der|tomcat)/1 = 0 +=.333/1 = 0.333

j=2

total = P(Kater|NULL)+P(Kater|the)+P(Kater|tomcat)=1

tc(Kater|NULL) += P(Kater|NULL)/1 = 0 += .333/1 = 0.333

tc(Kater|the) += P(Kater|the)/1 = 0 += .333/1 = 0.333

tc(Kater|tomcat) += P(Kater|tomcat)/1 = 0 += .333/1 = 0.333

- Re-compute translation probabilities

total(the) = tc(der|the) + tc(Hund|the) + tc(Kater|the) = 0.666 + 0.333 + 0.333 = 1.333

P(der|the) = tc(der|the)/total(the) = 0.666 / 1.333 = 0.5

P(Hund|the) = tc(Hund|the)/total(the) = 0.333/1.333 = 0.25

P(Kater|the) = tc(Kater|the)/total(the) = 0.333/1.333 = 0.25

total(dog) = tc(der|dog) + tc(Hund|dog) = 0.666

P(der|dog) = tc(der|dog)/total(dog) = 0.333 / 0.666 = 0.5

P(Hund|dog) = tc(Hund|dog)/total(dog) = 0.333 / 0.666 = 0.5

total(tomcat) = tc(der|tomcat) + tc(Kater|tomcat) = 0.333 + 0.333 =0.666

P(der|tomcat) = P(Kater|tomcat) = 0.5

total(NULL) = tc(der|NULL) + tc(Hund|NULL) + tc(KATER|NULL) = 0.666 + 0.333 + 0.333 = 1.333

P(der|NULL) = 0.5 P(Hund|NULL) = 0.25 P(Kater|NULL) = 0.25

IBM MODEL 1 EXAMPLE

- Iteration 2:

NULL the dog :: der Hund

j=1

total = P(der|NULL)+P(der|the)+P(der|dog)
= 0.5 + 0.5 + 0.5 = 1.5

tc(der|NULL) += P(der|NULL)/1.5 = 0 += .5/1.5 = 0.333

tc(der|the) += P(der|the)/1.5 = 0 += .5/1.5 = 0.333

tc(der|dog) += P(der|dog)/1.5= 0 += .5/1.5 = 0.333

j=2

total = P(Hund|NULL)+P(Hund|the)+P(Hund|dog)
= 0.25 + 0.25 + 0.5 = 1

tc(Hund|NULL) += P(Hund|NULL)/1 = 0 += .25/1 = 0.25

tc(Hund|the) += P(Hund|the)/1 = 0 += .25/1 = 0.25

tc(Hund|dog) += P(Hund|dog)/1 = 0 += .5/1 = 0.5

NULL the tomcat :: der Kater

j=1

total = P(der|NULL)+P(der|the)+P(der|tomcat)
= 0.5 + 0.5 + 0.5 = 1.5

tc(der|NULL) += P(der|NULL)/1.5 = 0.333 += .5/1.5 = 0.666

tc(der|the) += P(der|the)/1.5 = 0.333 += .5/1.5 = 0.666

tc(der|tomcat) += P(der|tomcat)/1.5= 0 += .5/1.5 = 0.333

j=2

total = P(Kater|NULL)+P(Kater|the)+P(Kater|tomcat)
= 0.25 + 0.25 + 0.5 = 1

tc(Kater|NULL) += P(Kater|NULL)/1 = 0 += .25/1 = 0.25

tc(Kater|the) += P(Kater|the)/1 = 0 += .25/1 = 0.25

tc(Kater|tomcat) += P(Kater|tomcat)/1 = 0 += .5/1 = 0.5

IBM MODEL 1 EXAMPLE: PROBS AFTER ITERATION 2

- Re-compute translations (iteration 2):

total(the) = tc(der|the) + tc(Hund|the) + tc(Kater|the) = .666 + 0.25 + 0.25 = 1.166

P(der|the) = tc(der|the)/total(the) = .666 / 1.166= **0.57**

P(Hund|the) = tc(Hund|the)/total(the) = 0.25/1.166 = 0.214

P(Kater|the) = tc(Kater|the)/total(the) = 0.25/1.166 = 0.214

total(dog) = tc(der|dog) + tc(Hund|dog) = 0.333 + 0.5 = 0.833

P(der|dog) = tc(der|dog)/total(dog) = 0.333 / 0.833 = 0.4

P(Hund|dog) = tc(Hund|dog)/total(dog) = 0.5 / 0.833 = **0.6**

total(tomcat) = tc(der|tomcat) + tc(Kater|tomcat) = 0.333 + 0.5 = 0.833

P(der|tomcat) = 0.4 **P(Kater|tomcat) = 0.6**

total(NULL) = tc(der|NULL) + tc(Hund|NULL) + tc(KATER|NULL) = .666 + 0.25+ 0.25 = 1.166

P(der|NULL) = 0.57 P(Hund|NULL) = 0.214 P(Kater|NULL) = 0.214

➔Slowly, this example converges to the true translations

FROM MODEL 1 TO MODEL 3

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Model 1

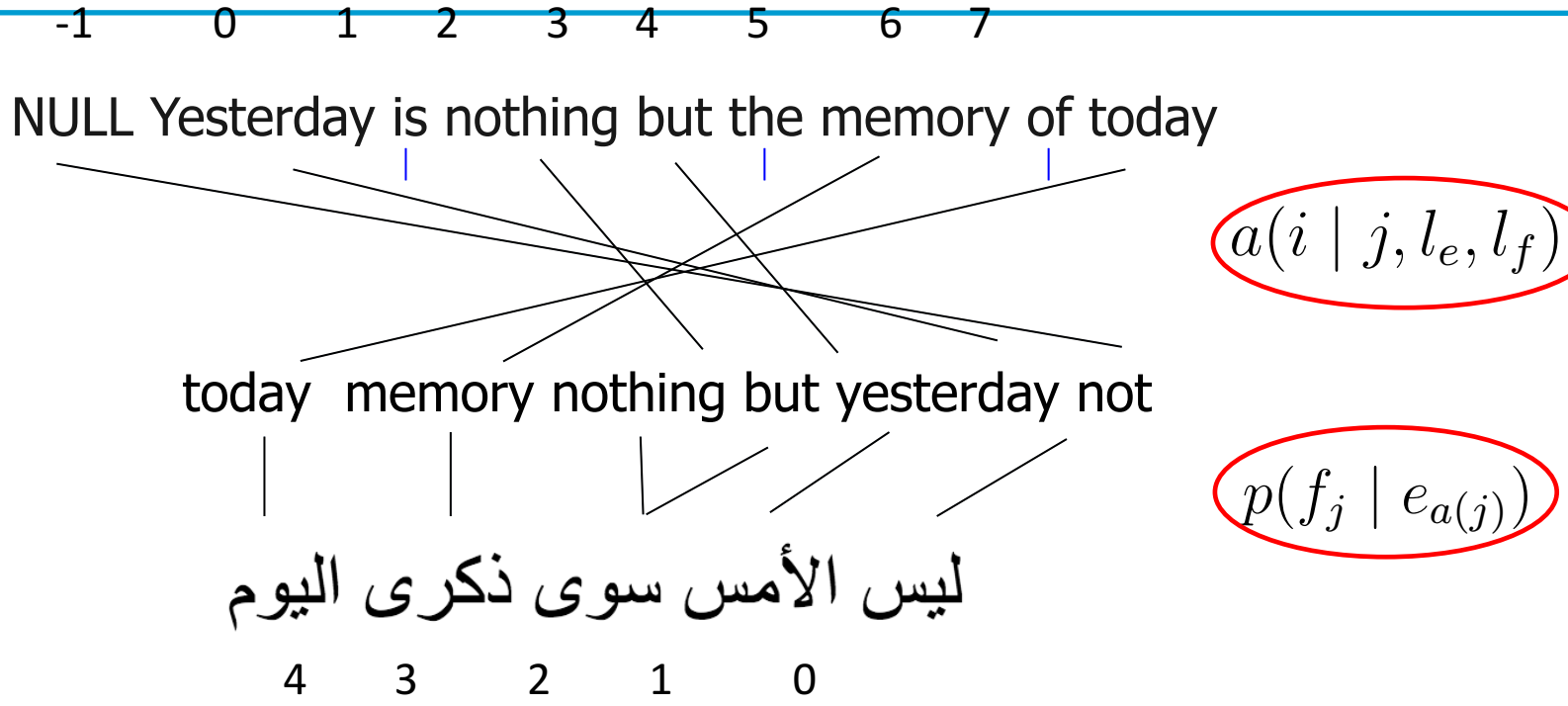
- IBM Model 1 allows for an efficient computation of translation probabilities
- No notion of fertility, i.e., it is possible that the same English word is the best translation for all foreign words
- No positional information, i.e., depending on the language pair, there might be a tendency that words occurring at the beginning of the English sentence are more likely to align to words at the beginning of the foreign sentence

Model 2+3

- (3) Fertility: how likely does a single word align to several words
- (2) Distortion: how likely is the alignment of a word in position i with a word in position j?

MODEL 2

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



$$a(i \mid j, l_e, l_f)$$

$$p(f_j \mid e_{a(j)})$$

$$p(f, a \mid e) = \frac{\epsilon}{(l_e+1)^{l_f}} \prod_{j=1}^{l_f} p(f_j \mid e_{a(j)}).a(i \mid j, l_e, l_f)$$

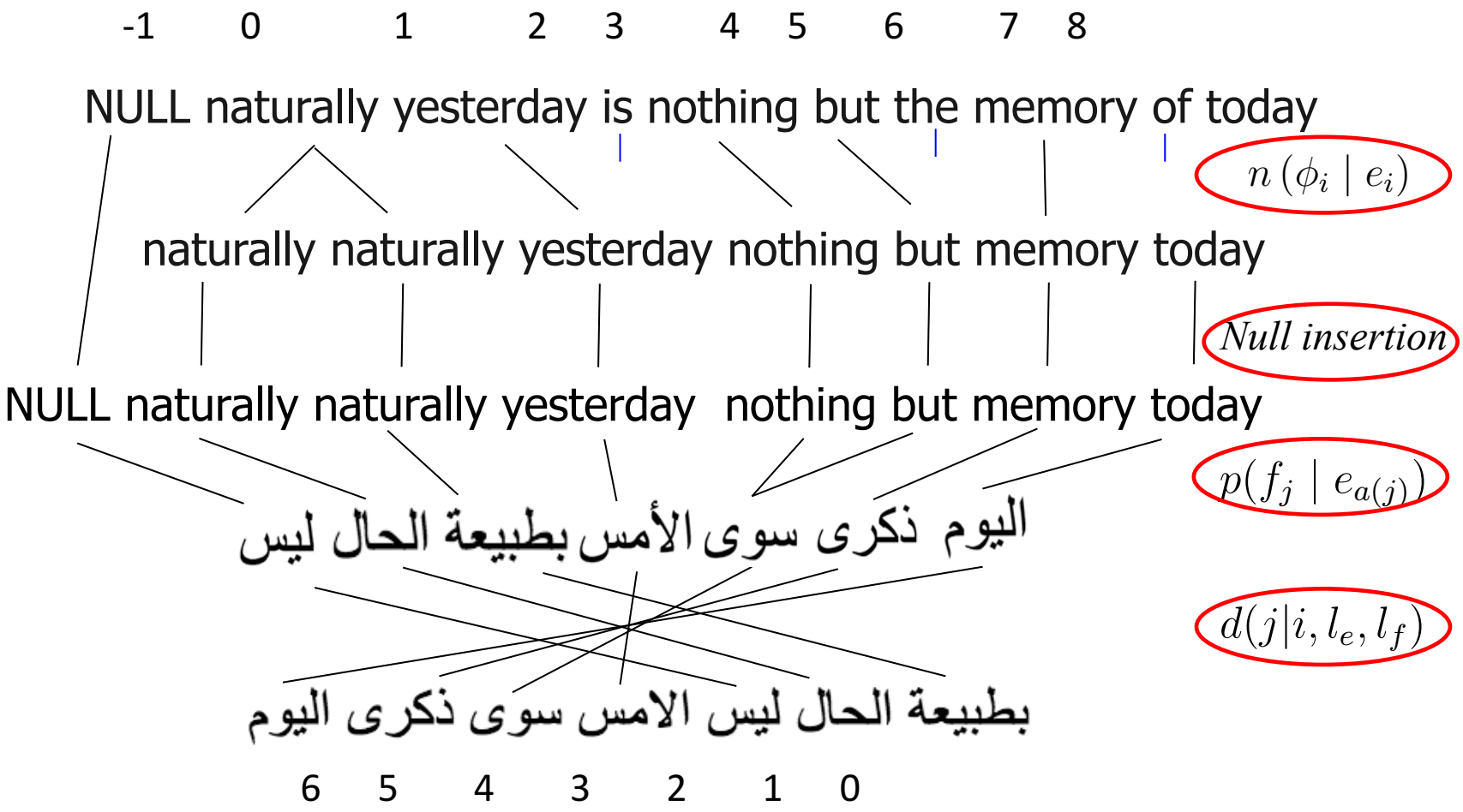
MODEL 2

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

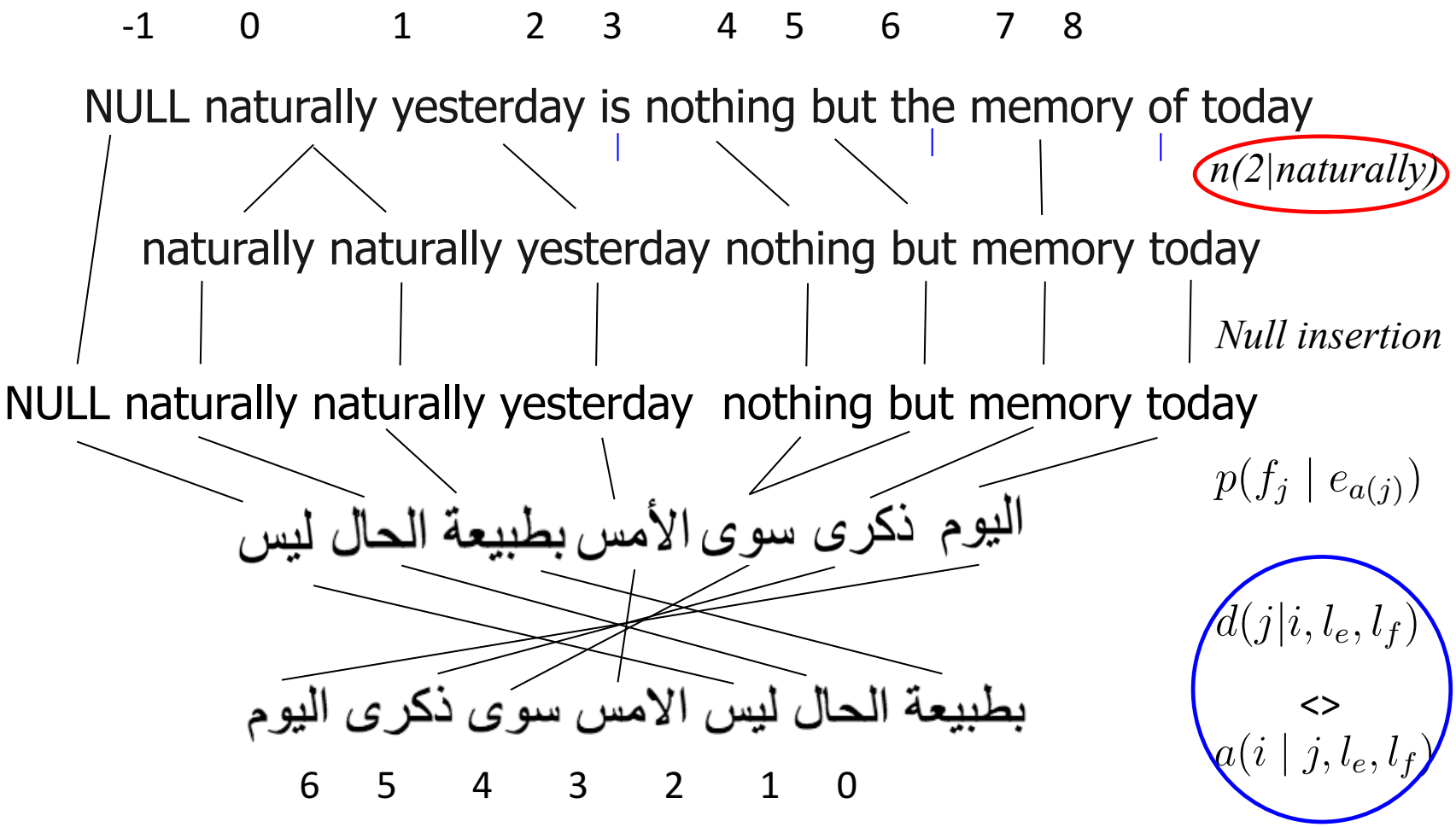
- Adds a new parameter that estimates the alignment
  - The input word alignment affect the alignment probability
- Introduces $a(i \mid j, l_e, l_f)$, the probability that the output word i is mapped to the input word at position j
- Changes the translation probability to:

$$p(f, a \mid e) = \frac{\epsilon}{(l_e+1)^{l_f}} \prod_{j=1}^{l_f} p(f_j \mid e_{a(j)}).a(i \mid j, l_e, l_f)$$

MODEL 3


-1 0 1 2 3 4 5 6 7 8
NULL naturally yesterday is nothing but the memory of today
$n(\phi_i \mid e_i)$
naturally naturally yesterday nothing but memory today
Null insertion
NULL naturally naturally yesterday nothing but memory today
$p(f_j \mid e_{a(j)})$
اليوم ذكرى سوى الأمس بطبيعة الحال ليس
$d(j|i, l_e, l_f)$
بطبيعة الحال ليس الامس سوى ذكرى اليوم
6 5 4 3 2 1 0

MODEL 3


-1 0 1 2 3 4 5 6 7 8
NULL naturally yesterday is nothing but the memory of today
n(2|naturally)
naturally naturally yesterday nothing but memory today
Null insertion
NULL naturally naturally yesterday nothing but memory today
$p(f_j \mid e_{a(j)})$
اليوم ذكرى سوى الأمس بطبيعة الحال ليس
بطبيعة الحال ليس الامس سوى ذكرى اليوم
6 5 4 3 2 1 0
$d(j|i, l_e, l_f)$
<>
$a(i \mid j, l_e, l_f)$

MODEL 3

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Adds the notion of fertility $n(\phi_i \mid e_i)$ to model 1 and 2. Fertility represents how many are generated from each input or the fertility

- Adds a NULL insertion step for all the words that get mapped to epsilon

- Models the distortion $d(j|i, l_e, l_f)$ which produces a foreign output position conditioned on the foreign input (Koehn 2010)

$$p(f \mid e) = \sum_{i=1}^{l_f} \sum_{a_i=1}^{l_e} \begin{pmatrix} l_f - \varphi_0 \\ \varphi_0 \end{pmatrix} p_0^{l_f - 2\varphi_0} p_1^{\varphi_0} * \prod_{i=1}^{l_e} \varphi_i! n(\varphi_i|e_i) * \prod_{j=1}^{l_f} t(f_i \mid e_{a_{(j)}}) d(j \mid a_j, l_f, l_e)$$

MODEL 2

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Fixes the model deficiency of models 1-4
- Includes heuristics that would prohibit the placement of an output word in a position already taken
- Places words in free memory position
- Distributes the probability mass to valid alignments

FERTILITY

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- The best Model 1 alignment could be that a single English word aligns to all foreign words
- This is clearly not desirable and we want to constrain the number of words an English word can align to
- Fertility models a probability distribution that word e aligns to k words: n(k,e)
- Consequence: translation probabilities cannot be computed independently of each other anymore
- IBM Model 3 has to work with full alignments, note there are up to $(l+1)^m$ different alignments: this is infeasible
- Solution: Compute the best alignment with Model 1 and change some of the alignments to generate a set of likely alignments (pegging)
- Model 3 takes this restricted set of alignments as input

NEIGHBORHOOD OF AN ALIGNMENT

Fertility

- Given an alignment a we can derive additional alignments from it by making small changes:
- Changing a link (j,i) to (j,i')
- Swapping a pair of links (j,i) and (j',i') to (j,i') and (j',i)
- The resulting set of alignments is called the neighborhood of a

Distortion

The distortion factor determines how likely it is that an English word in position i aligns to a foreign word in position j, given the lengths of both sentences: d(j | i, l, m). Note: positions are absolute positions

Then, use EM on the neighborhoods to find a better alignment

DEFICIENCY: LEAKY MODEL!

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

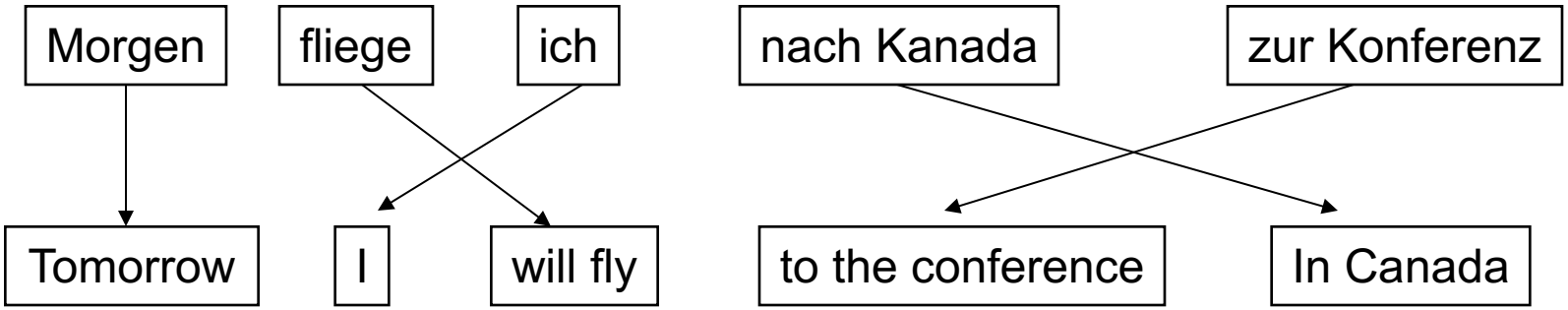
- Problem with IBM Model 3: It assigns probability mass to impossible strings
- Well formed string: “This is possible”
- Ill-formed but possible string: “This possible is”
- Impossible string: “possible This is”
- Impossible strings are due to distortion values that generate different words at the same position
- Impossible strings can still be filtered out in later stages of the translation process

Again, leaky models are functional, but waste probability mass on impossible events.

LIMITATIONS OF IBM MODELS (1-5)

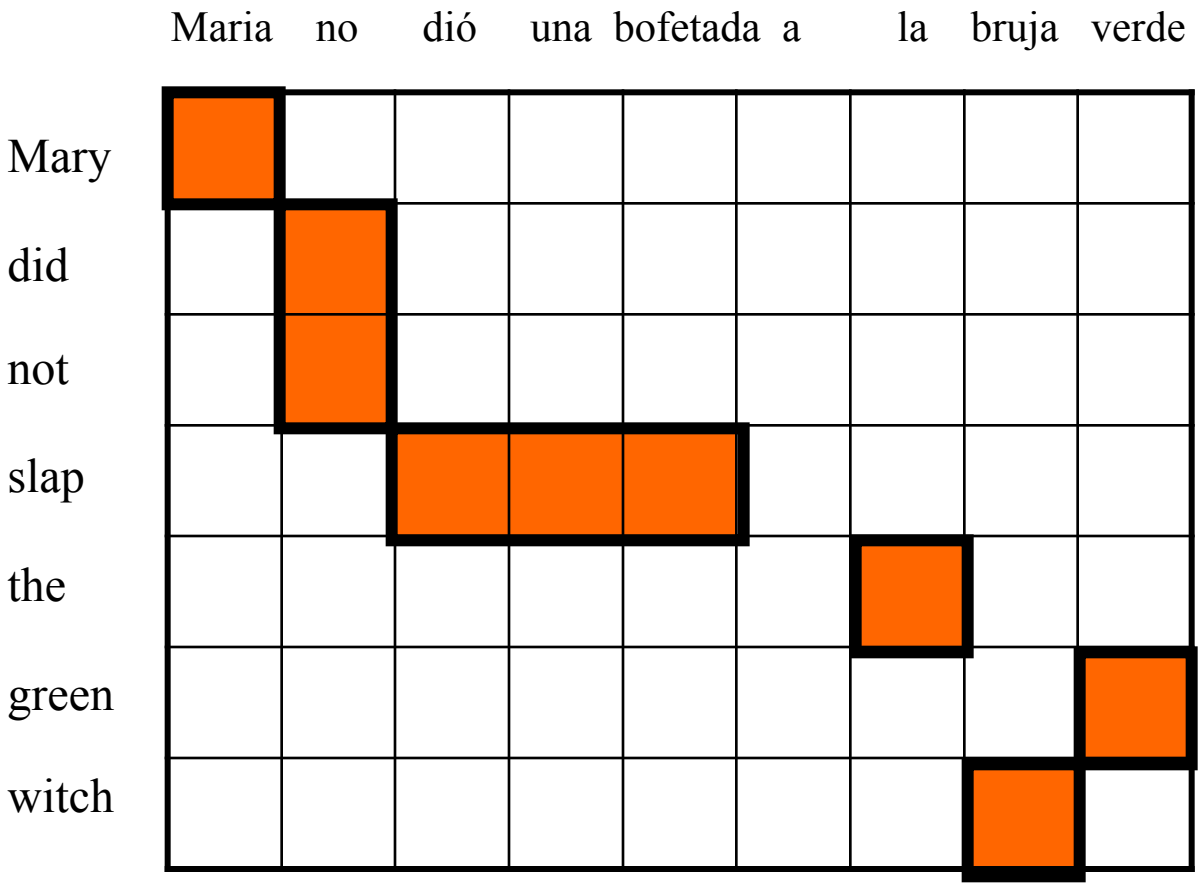
- Only 1-to-N word mapping
- Handling fertility-zero words (difficult for decoding)
- Almost no syntactic information
  - Word classes
  - Relative distortion
  - parse trees
- Long-distance word movement
- Fluency of the output depends entirely on the English language model

PHRASE-BASED STATISTICAL MT



- Foreign input segmented into phrases
- “phrase” is any sequence of words
- Each phrase is probabilistically translated into English
  - P(to the conference | zur Konferenz)
  - P(into the meeting | zur Konferenz)
- Phrases are probabilistically re-ordered

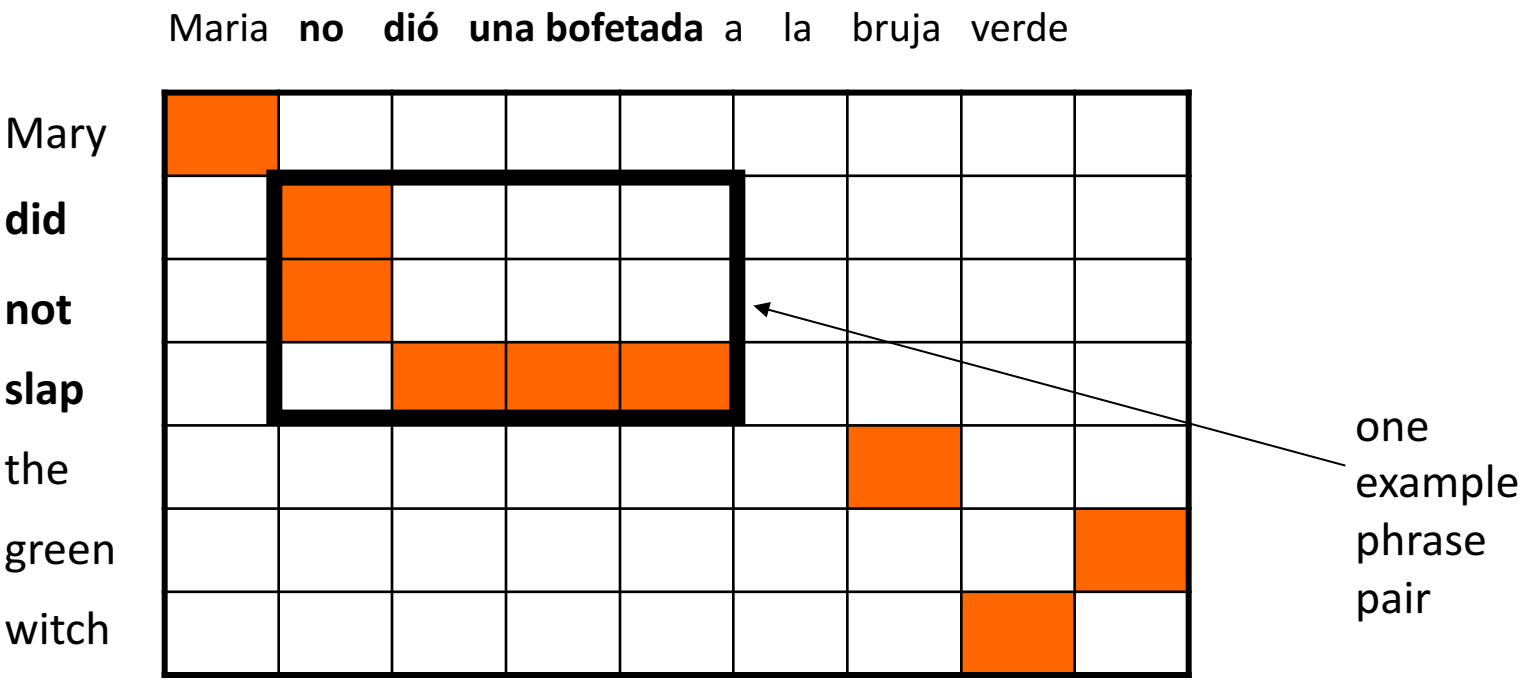
WORD ALIGNMENT INDUCED PHRASES



(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

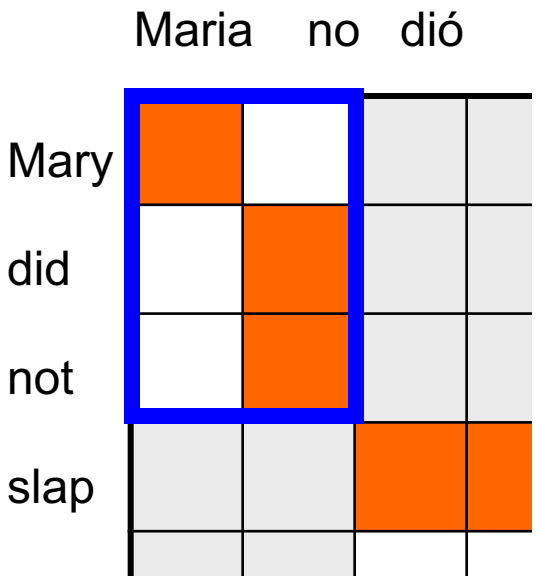
HOW TO LEARN THE PHRASE TRANSLATION TABLE?

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

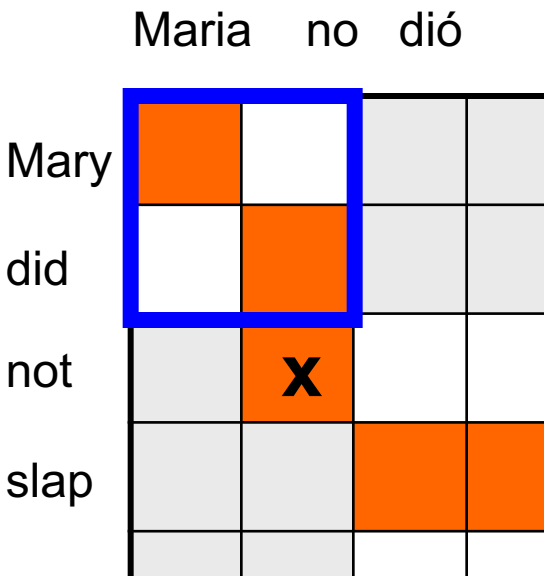


- Collect **all phrase pairs** that are **consistent** with the word alignment

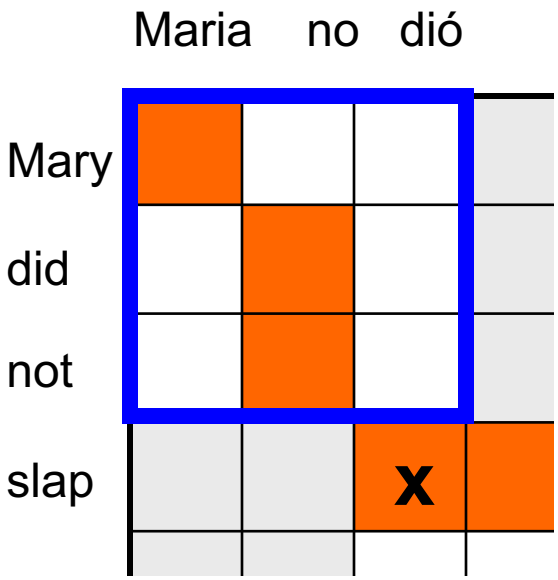
CONSISTENT WITH WORD ALIGNMENT



consistent



inconsistent



inconsistent

- Phrase alignment must contain all alignment points for all the words in both phrases!

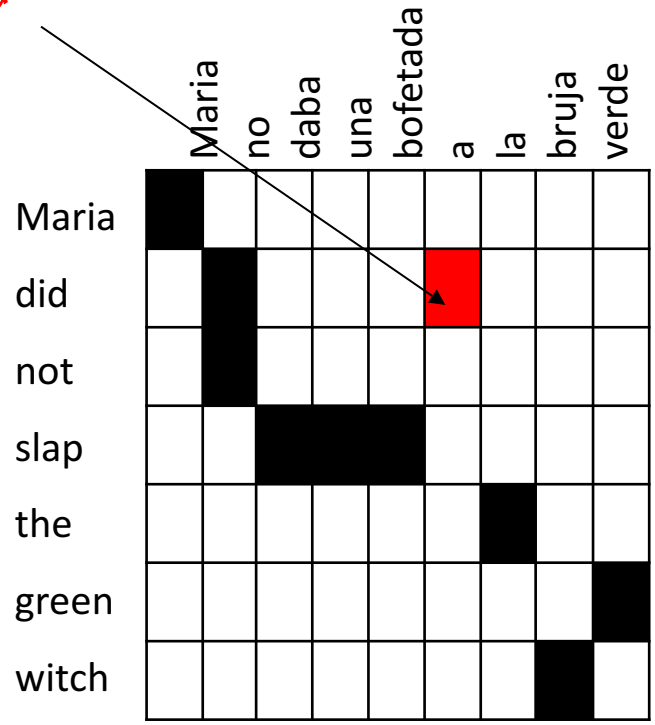
GIZA++

USES HEURISTICS SUCH AS *UNION* TO COMBINE THE ALIGNMENTS

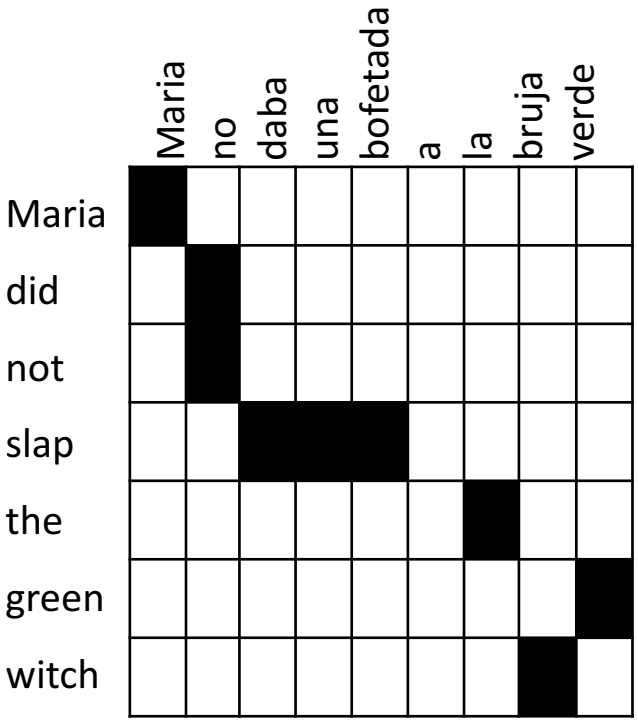
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Union

Over aligning



Correct

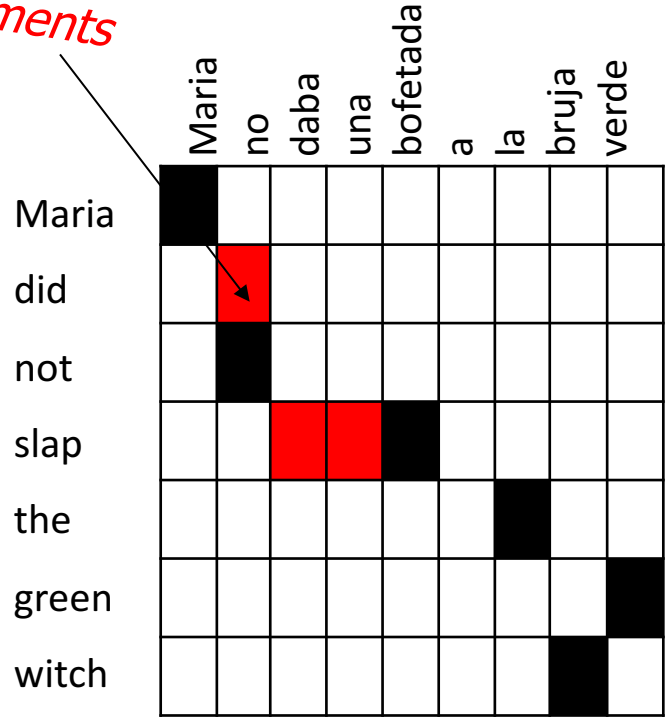


GIZA++

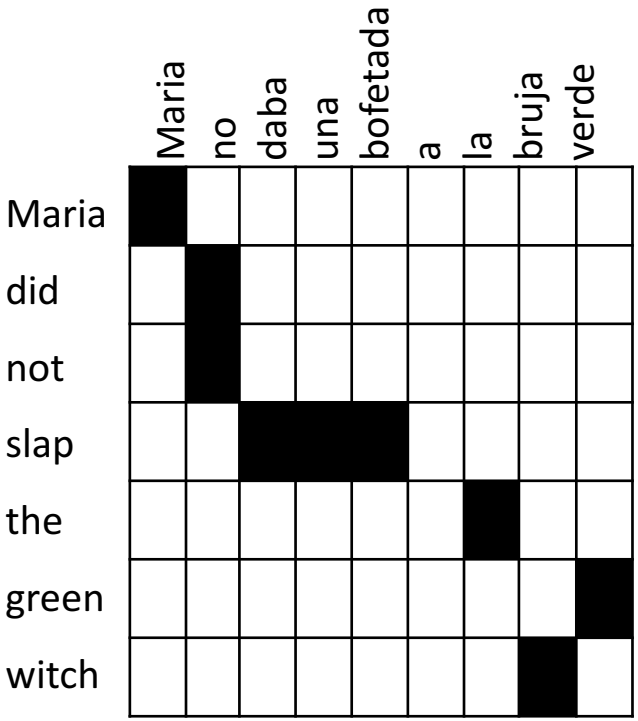
USES HEURISTICS SUCH AS *UNION* TO COMBINE THE ALIGNMENTS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Intersection



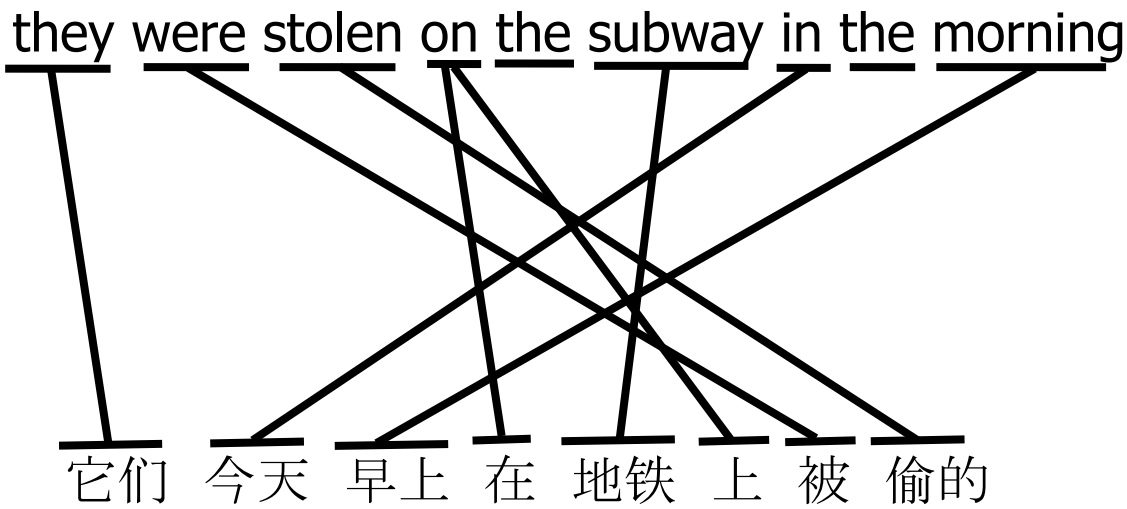
Correct



WORD ALIGNMENT INDUCED PHRASES

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

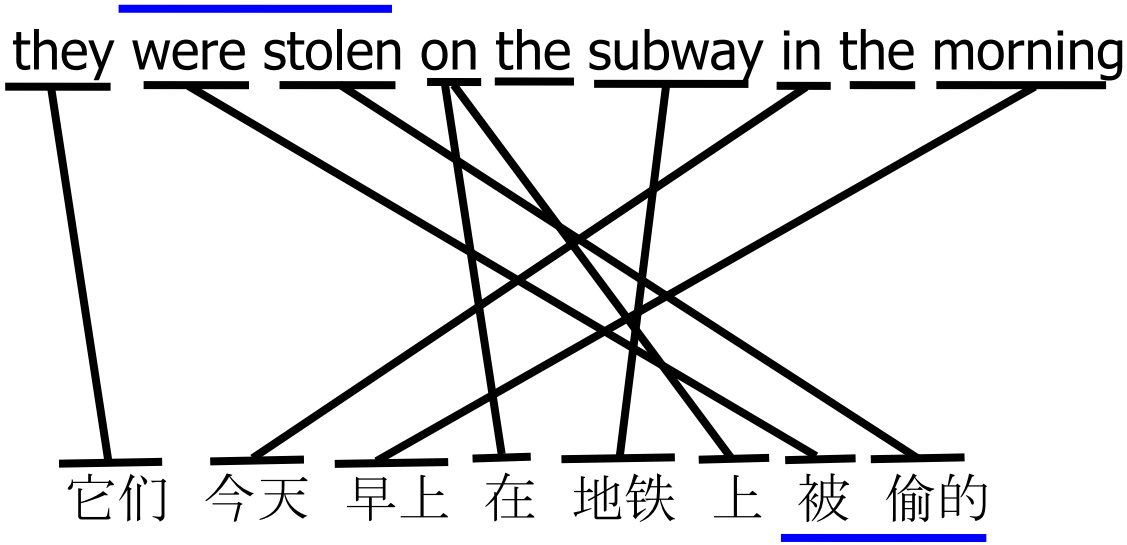
Human alignment



WORD ALIGNMENT INDUCED PHRASES

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

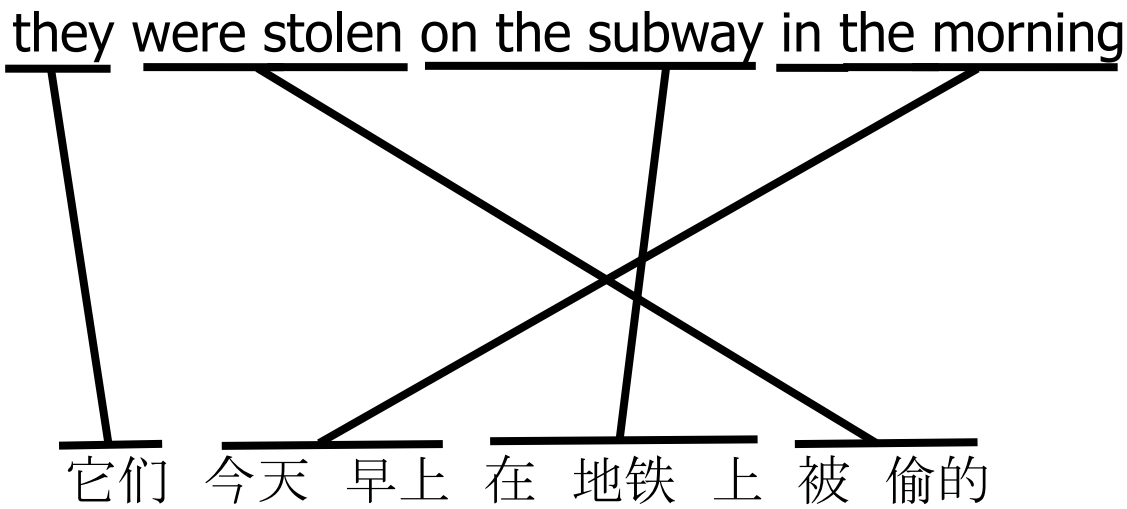
Human alignment



WORD ALIGNMENT INDUCED PHRASES

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Human phrase extraction



ADVANTAGES OF PHRASE-BASED SMT

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Many-to-many mappings can handle non-compositional phrases
- takes both directions into account
- Local context is very useful for disambiguating
  - "Interest rate" → …
  - "Interest in" → …
- The more data, the longer the learned phrases: Sometimes whole sentences

DECODING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Goal is to find a translation that maximizes the product of the translation and language models.

$$\underset{e}{\operatorname{argmax}} P(f \mid e)P(e)$$

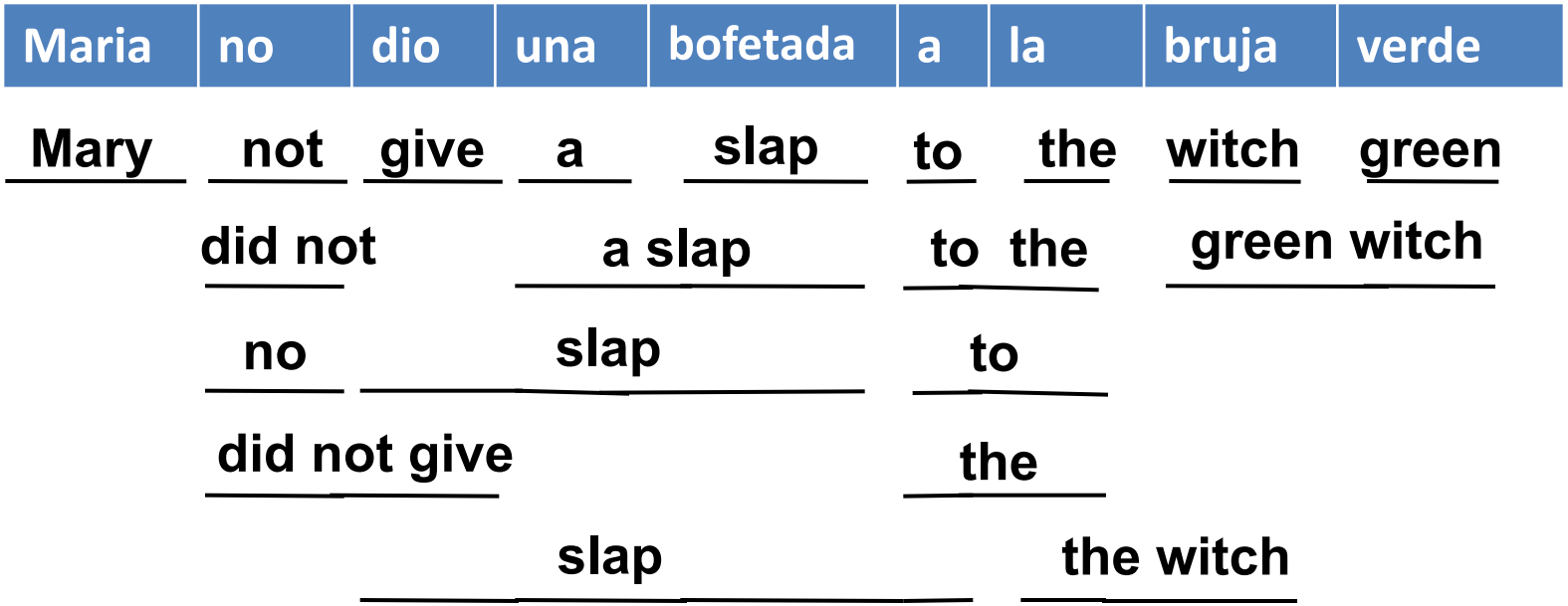
- Cannot explicitly enumerate and test the combinatorial space of all possible translations.
- Must efficiently (heuristically) search the space of translations that approximates the solution to this difficult optimization problem.
- The optimal decoding problem for all reasonable models (e.g. IBM model 1) is NP-complete.

Here:

- phrase-based decoder based on that of Koehn's (2004) Pharaoh system.

Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models, Philipp Koehn, AMTA 2004

SPACE OF TRANSLATIONS



The phrase translation table from the alignment defines the space of possible translations

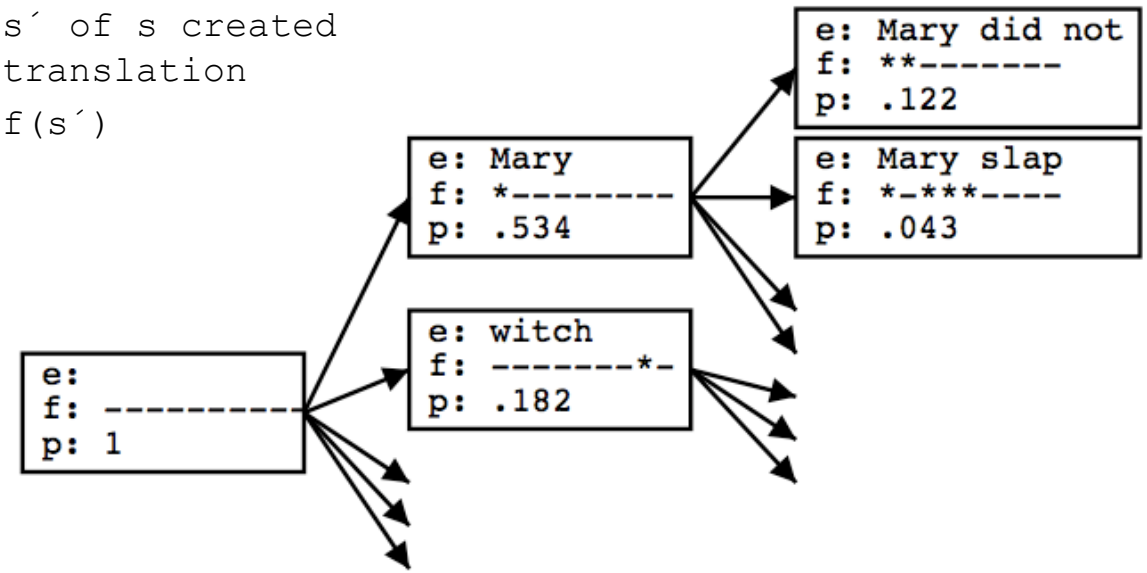
- every word can have multiple translations
- every word can participate in multiple phrases

STACK DECODING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Use a version of heuristic A* search to explore the space of phrase translations to find the best scoring subset that covers the source sentence.

```
Initialize priority queue Q (stack) to empty translation.
Loop:
      s = pop(Q)
      If h is a complete translation, exit loop and return it.
      For each refinement s´ of s created
       by adding a phrase translation
            Compute score f(s´)
            Add s´ to Q
      Sort Q by score f
```



SEARCH HEURISTIC

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- A* is best-first search using the function f to sort the search queue:
  - f(s) = g(s) + h(s)
  - g(s): Cost of existing partial solution
  - h(s): Estimated cost of completion of solution
- If h(s) is an underestimate of the true remaining cost (**admissible heuristic**) then A* is guaranteed to return an optimal solution.
- Known quality of partial translation, E, composed of a set of chosen phrase translations S based on phrase translation and language models:

$$g(s) = \log \frac{1}{\left(\prod_{i \in S} \varphi(\bar{f}_i, \bar{e}_i) d(i) \right) P(E)}$$

ESTIMATING THE FUTURE COST

- True future cost requires knowing the way of translating the remainder of the sentence in a way that maximizes the probability of the final translation.
- However, this is not computationally tractable.
- Therefore under-estimate the cost of remaining translation by ignoring the distortion component and computing the most probable remaining translation ignoring distortion (which is efficiently computable using the Viterbi algorithm)

BEAM SEARCH

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- However, Q grows too large to be efficient and guarantee an optimal result with full A* search.
- Therefore, always cut Q back to only the best (lowest cost) K items to approximate the best translation

```
Initialize priority queue Q (stack) to empty translation.
Loop:
      If top item on Q is a complete translation, exit loop
        and return it.
      For each element s of Q do
          For each refinement s´ of s created
            by adding a phrase translation
                Compute score f(s´)
                Add s´ to Q
      Sort Q by score f
      Prune Q back to only the first (lowest cost) K items
```

MULTISTACK DECODING

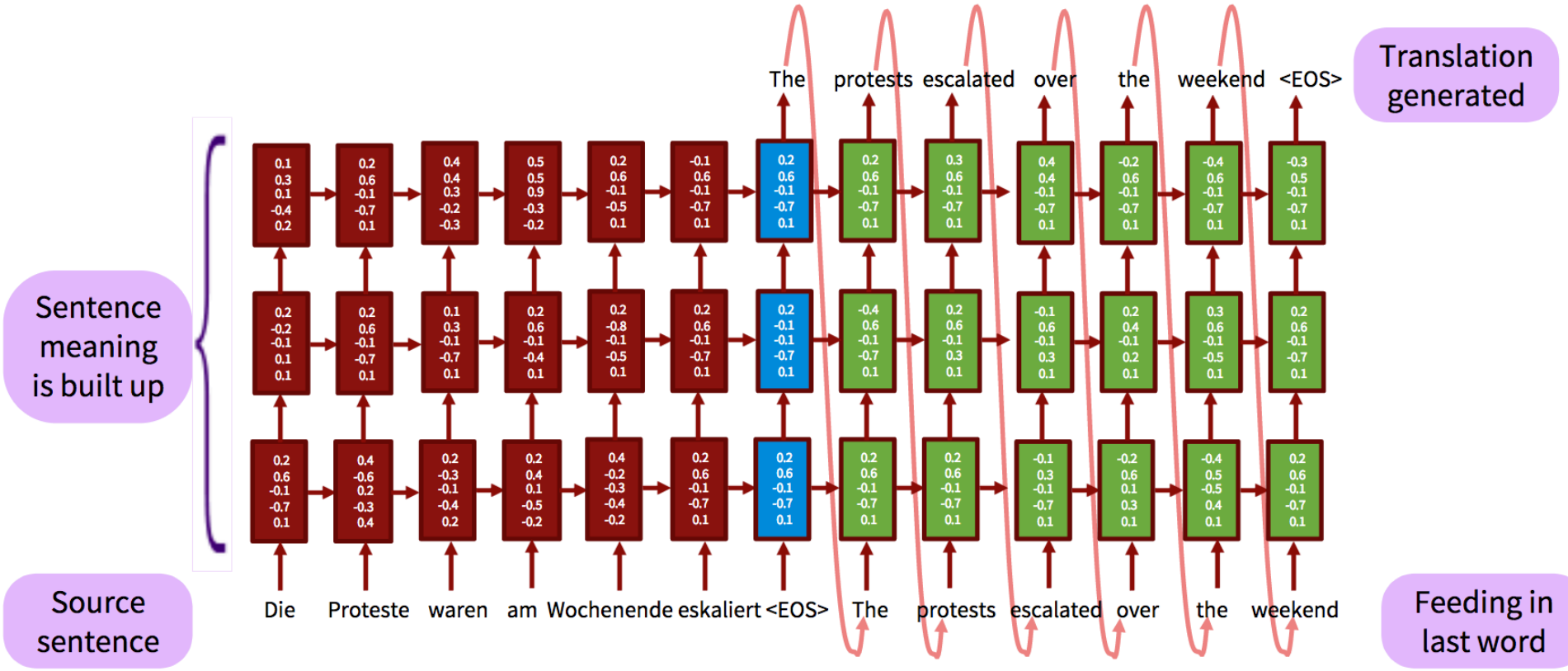
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- It is difficult to compare translations that cover different fractions of the foreign sentence, so maintain multiple priority queues (stacks), one for each subset of foreign words currently translated.

- this prevents us from pruning solutions that cover all words: we prune *per stack*

- Finally, return best scoring translation in the queue of translations that cover all of the words in F.

SEQUENCE TO SEQUENCE NEURAL MT



- Encoding/decoding with (recurrent) LSTMs

Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2014. Sequence to Sequence Learning with Neural Networks, Proc. NIPS 2014
Tutorial: see https://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf

ADVANTAGES OF NEURAL MT

- End-to-End training: **all** parameters are simultaneously optimized
- Distributed dense-vector representations: exploits word similarities
- 'Infinite' context: neural models better make use of long-range contexts

MT EVALUATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Manual:
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - Error categorization

- Testing in an application that uses MT as one sub-component
  - Question answering from foreign language documents
  - Cross Language Information Retrieval

- Automatic:
  - **BLEU (Bilingual Evaluation Understudy)**
  - **YiSi**

BLEU EVALUATION METRIC

Reference (human) translation:

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 .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is in [0,1])
  - What percentage of machine n-grams can be found in the reference translation?
  - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the the")
- Brevity penalty
  - Can't just type out single word "the" (precision 1.0!)
- Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

BLEU EVALUATION METRIC

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Reference (human) translation:

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 .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

BLEU4 formula

(counts n-grams up to length 4)

exp (1.0 * log p1 +

0.5 * log p2 +

0.25 * log p3 +

0.125 * log p4 –

max(words-in-reference /

words-in-machine – 1, 0))

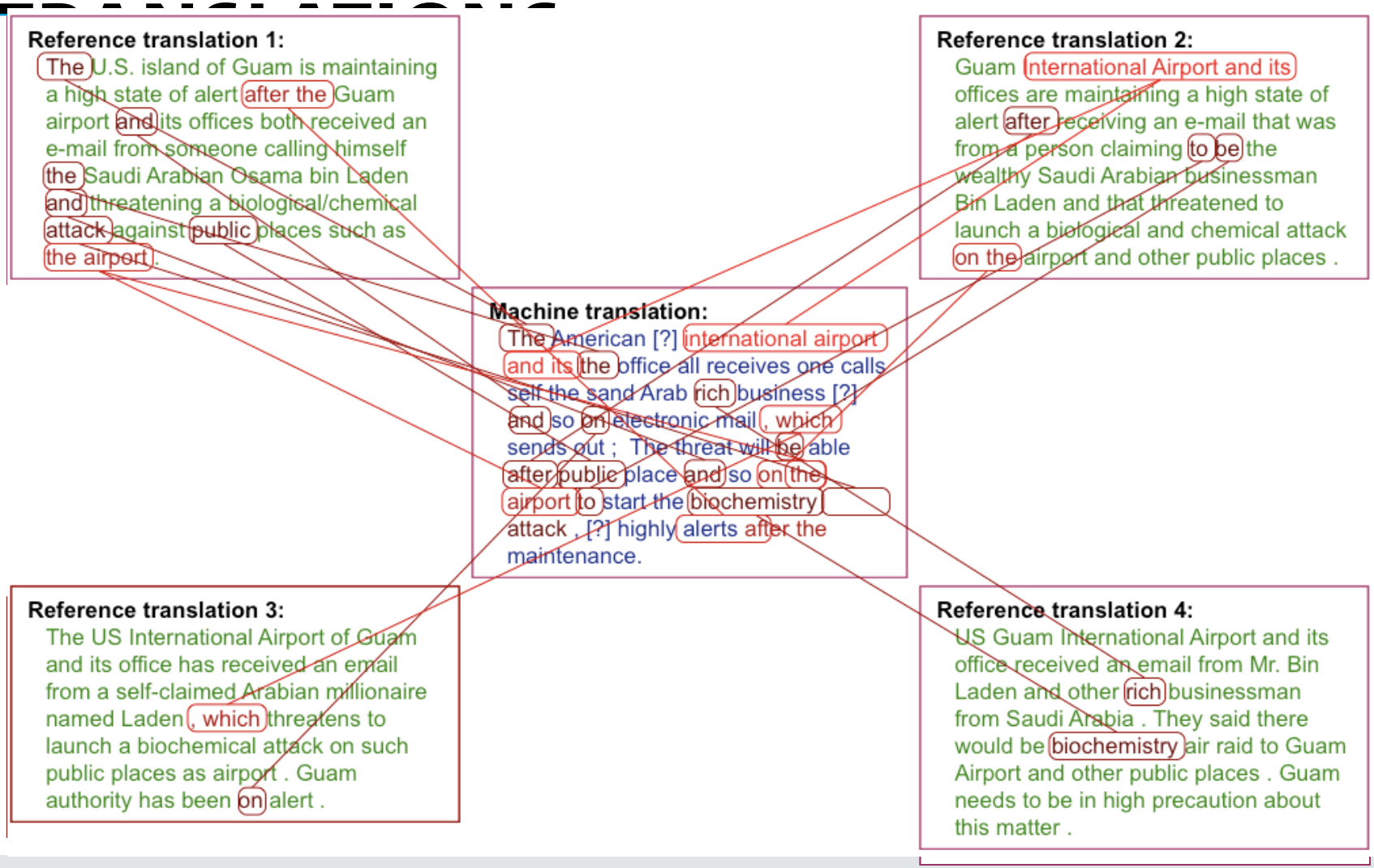
p1 = 1-gram precision

p2 = 2-gram precision

p3 = 3-gram precision

p4 = 4-gram precision

MULTIPLE REFERENCE TRANSLATIONS



BLEU CONSIDERS MT_1 TO BE A DESCENT TRANSLATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Input Il visit son pays natal demain pour célébrer l'anniversaire de sa maman

MT_1 He is out his country to his mother 's birthday

Reference Tomorrow, he visits his original country to celebrate his mother 's birthday

n-gram matchings	
1-gram matches:	7
2-gram matches:	4
3-gram matches:	**2**
4-gram matches:	**1**

BLEU CONSIDERS MT_1 TO BE BETTER THAN MT_2

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Input Il visite son pays natal demain pour célébrer l'anniversaire de sa maman

MT_1 He is out his country to his mother's birthday

MT_2 He goes back home to celebrate the birthday of his mama

Reference Tomorrow, he visits his original country to celebrate his mother's birthday

n-gram matchings **MT_1**	
1-gram matches:	7
2-gram matches:	4
3-gram matches:	**2**
4-gram matches:	**1**

n-gram matchings **MT_2**	
1-gram matches:	5
2-gram matches:	1
3-gram matches:	**0**
4-gram matches:	**0**

WHAT MAKES A TRANSLATION USEFUL?

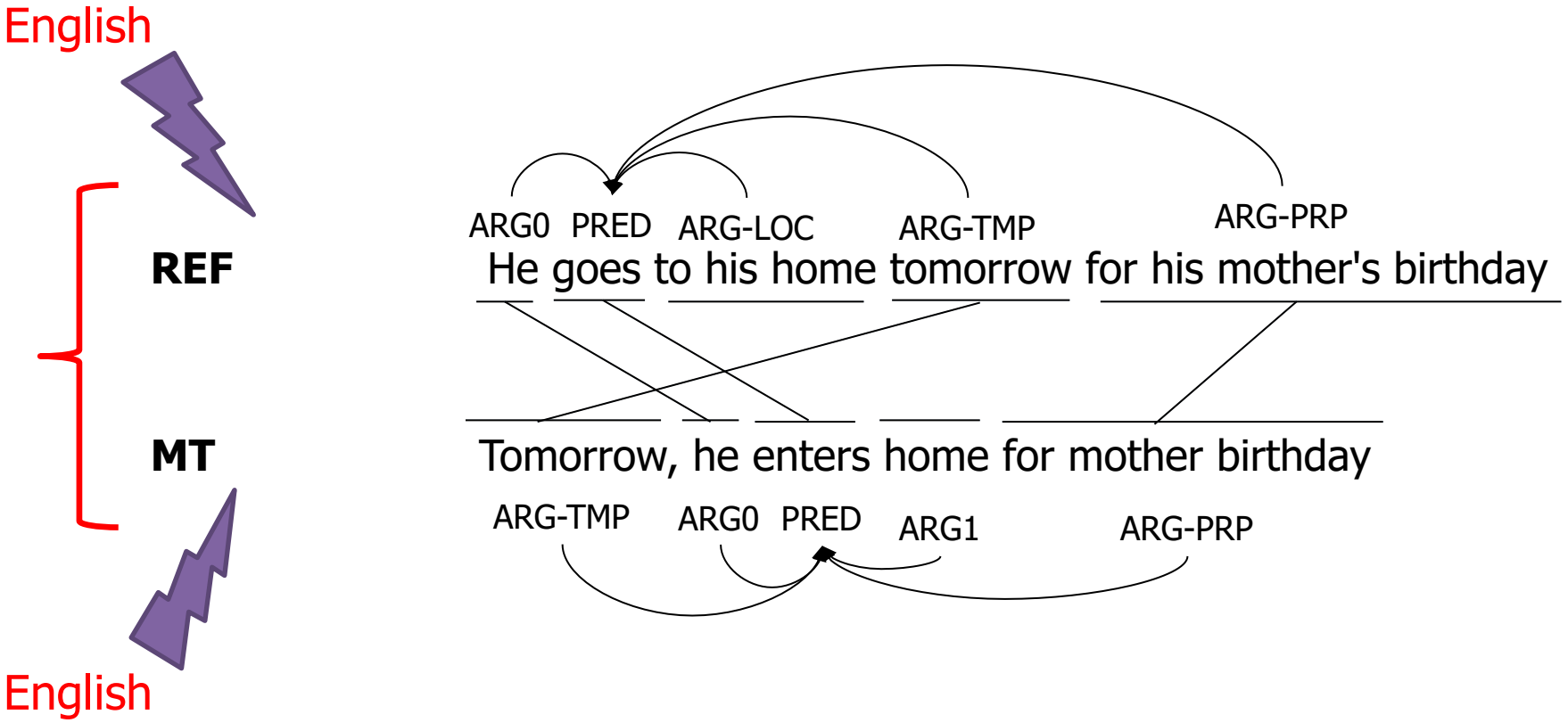
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

How well is

who did what to whom, for whom, when, where, why and how

preserved in translation?

SEMANTIC BASED EVALUATION



SEMANTIC BASED EVALUATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MT_2 He goes back home to celebrate the birthday of his mama

who verb where why

who what where why

Reference Tomorrow, he visits his original country to celebrate his mother's birthday

SRL matchings	
Predicate matches:	2
Argument matches:	8
	0
	0

SEMANTIC BASED EVALUATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Input Il visit son pays natal demain pour célébrer l'anniversaire de sa maman

MT_1 He is out his country to his mother's birthday

MT_2 He goes back home to celebrate the birthday of his mama

Reference Tomorrow, he visits his original country to celebrate his mother's birthday

SRL matchings **MT_1**	
Predicate matches:	0
Argument matches:	0

SRL matchings **MT_2**	
Predicate matches:	2
Argument matches:	8

IMMEDIATE FEEDBACK

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG