Machine Translation

CS 585, Fall 2016

Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2016

Brendan O'Connor

College of Information and Computer Sciences
University of Massachusetts Amherst

[Slides borrowed from http://mt-class.org/ and J&M 2nd. ed]

Final project

- 12/6 (Tues): we need your poster PDF if you want it printed (will send instructions to submit)
- 12/13 (Tues): poster session, 2:30-3:45 in CS 150/151
  - Poster boards are 2ft wide, 3ft tall
- 12/22: Final report due [[CHANGED from lecture. not 12/16.]]

- (This week: one last, writing-based HW: practice writing about research papers.)

MT is long-sought
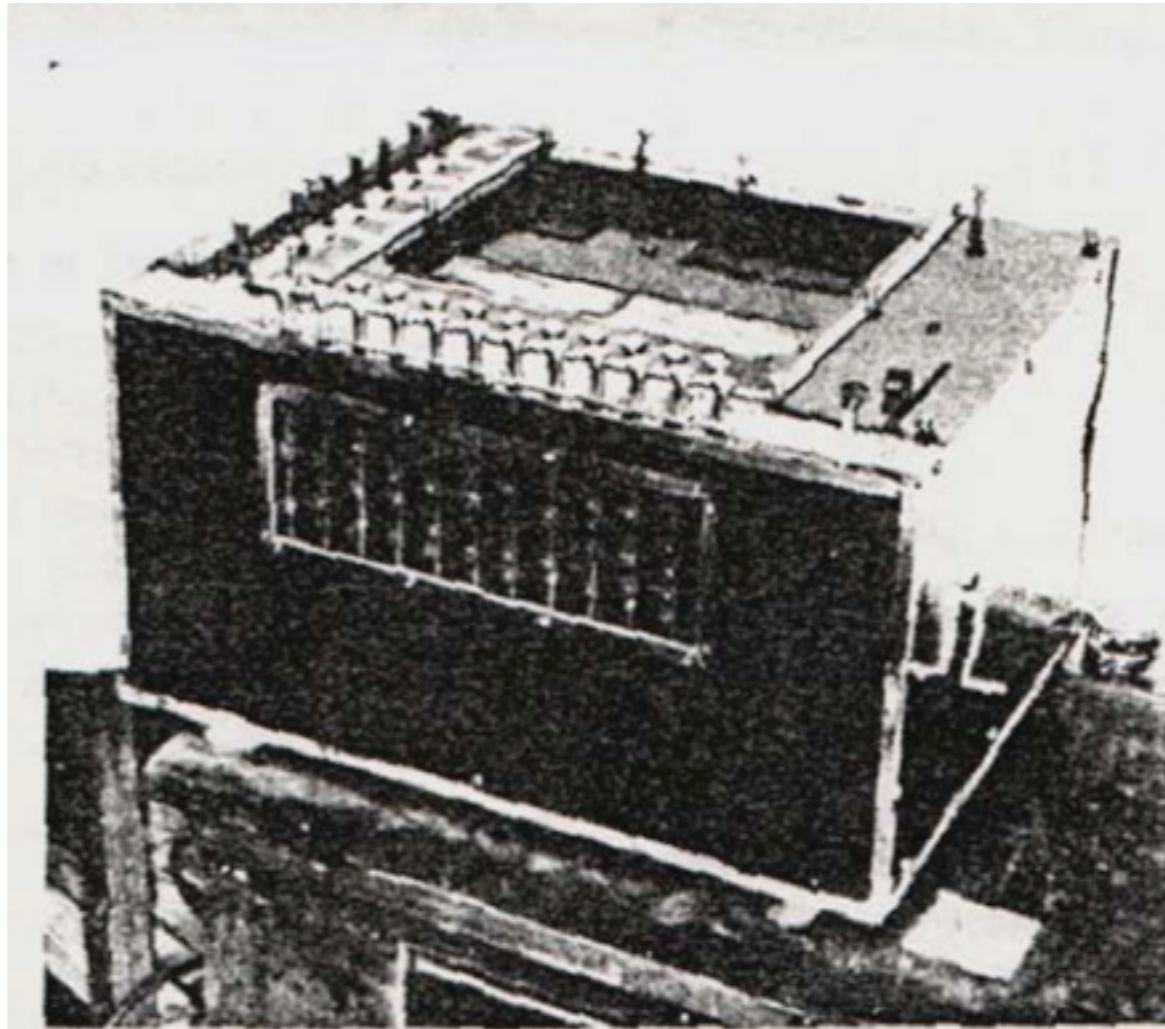
Georges Artrouni's "mechanical brain", a translation device patented in France in 1933. (Image from Corbé by way of [John Hutchins](#))

The memory was the core of the device. It consisted of a paper band 40 cm wide, which could be up to 40 meters in length, moving over two rolling drums and held in position by perforations on the edges. The dictionary entries were recorded in normal orthographic form (i.e. not coded) line by line in five columns. The first column was for the source language word (or term), the other columns for equivalents in other languages and for other useful information.
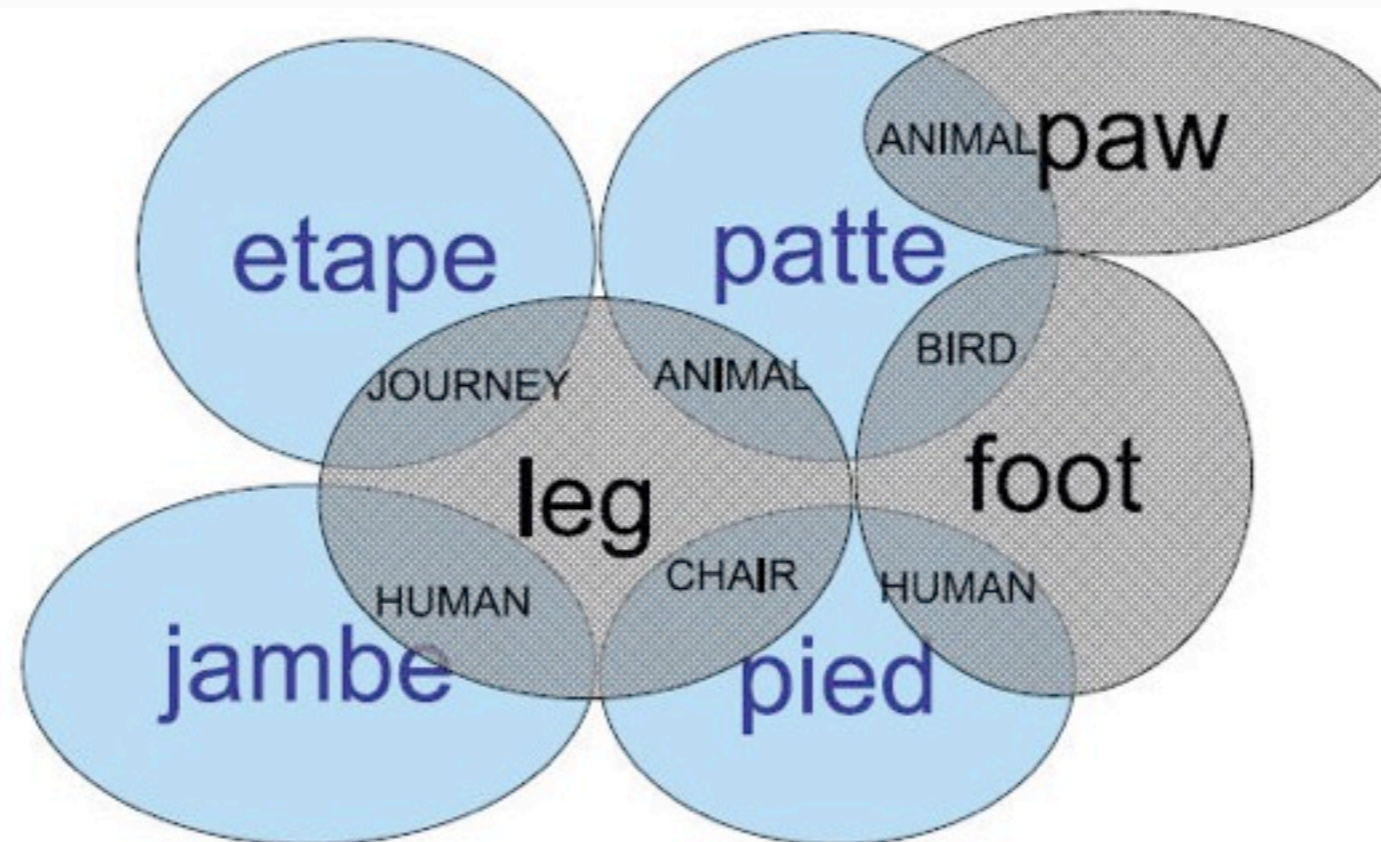
3

MT is hard

- Word order, word meanings
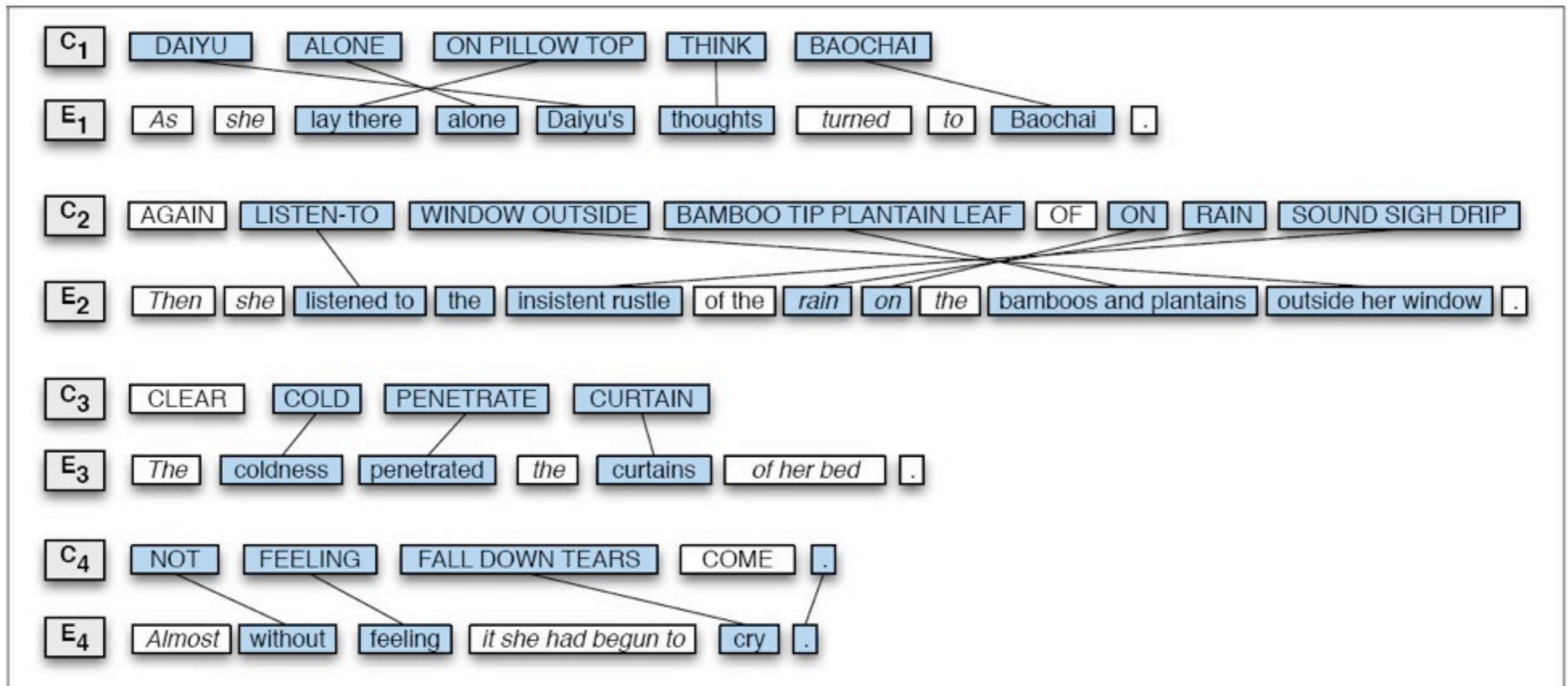
MT is hard

- Word meaning:
  many-to-many and context dependent

- *Translation* itself is hard: metaphors, cultural references, etc.

5

# MT goals

- Motivation: Human translation is expensive

- Rough translation vs. none

- Interactive assistance for human translators
  - e.g. Lilt
    - https://www.youtube.com/watch?v=YZ7G3gQgpfI
    - https://lilt.com/app/projects/details/1887/edit-document/2306
  - [compare to bilingual dictionary]

6

MT paradigms

- Rule-based *transfer rules*
  - Manually program lexicons/rules
  - SYSTRAN (AltaVista Babelfish; originally from 70s)

- Statistical MT
  - Word-to-word, phrase-to-phrase probs
  - Learn translation rules from data,
    search for high-scoring translation outputs
    - Phrase or syntactic transformations
  - Key research in the early 90s
  - Google Translate (mid 00s)
  - Open-source: Moses

- Neural MT
  - Research in early 10s; very recently deployed
  - Latent representations of words/phrases

7

Vauquois Triangle

8
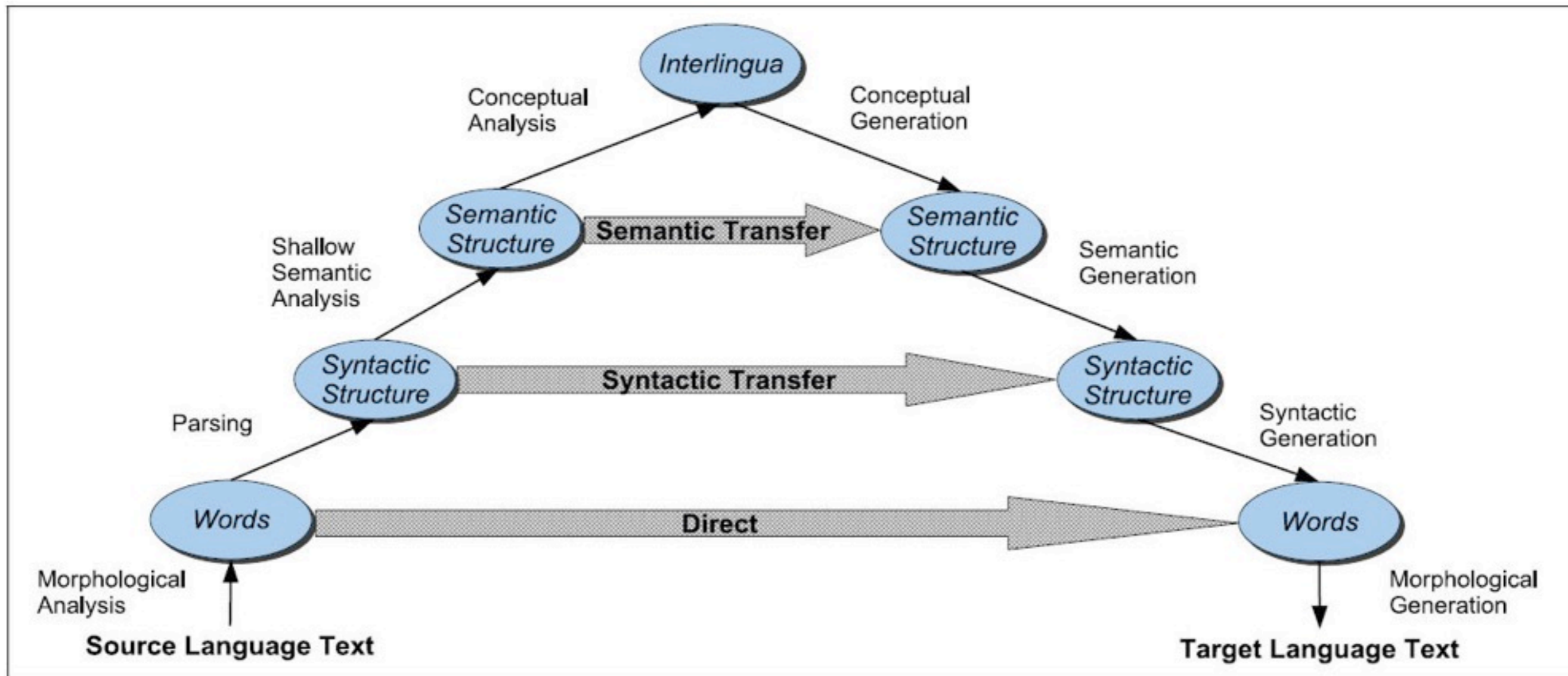
Direct (word-based) transfer
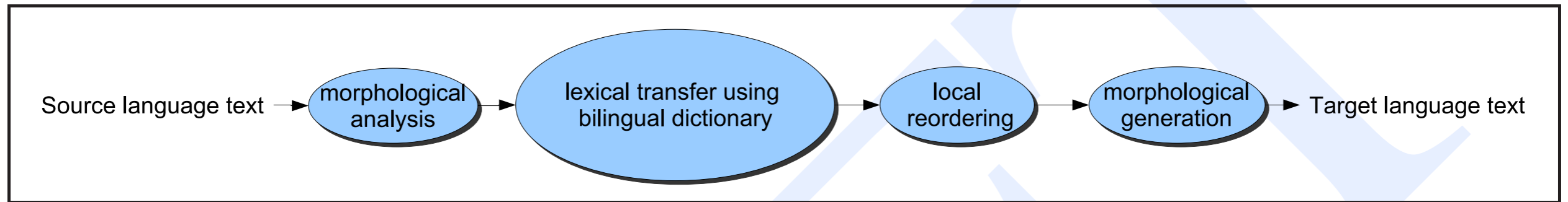
Input:	Mary didn't slap the green witch
After 1: Morphology	Mary DO-PAST not slap the green witch
After 2: Lexical Transfer	Maria PAST no dar una bofetada a la verde bruja
After 3: Local reordering	Maria no dar PAST una bofetada a la bruja verde
After 4: Morphology	Maria no dió una bofetada a la bruja verde

Syntactic transfer

Nominal $\Rightarrow$ Nominal	
Adj Noun Noun Adj	

Figure 25.10   A simple transformation that reorders adjectives and nouns
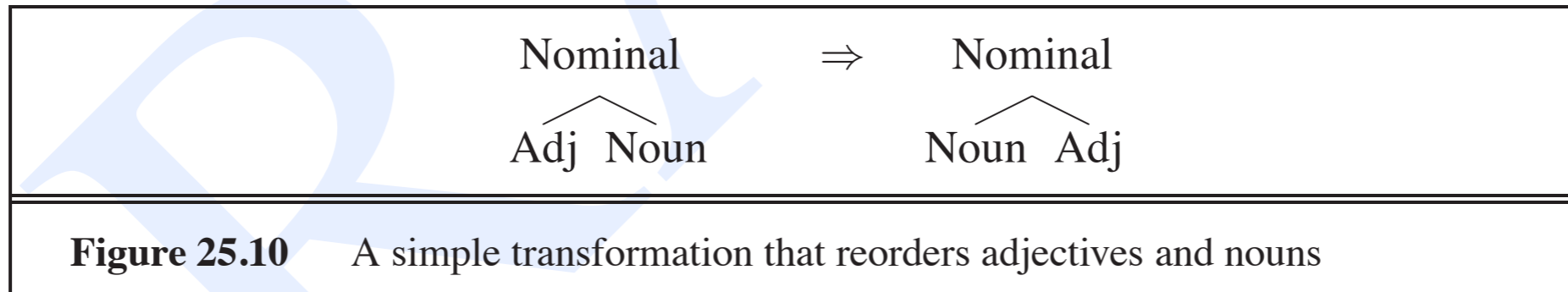
English to Spanish:

1. $NP \rightarrow Adjective_1\ Noun_2$ $\Rightarrow$ $NP \rightarrow Noun_2\ Adjective_1$

Chinese to English:

2. $VP \rightarrow PP[+Goal]\ V$ $\Rightarrow$ $VP \rightarrow V\ PP[+Goal]$

English to Japanese:

3. $VP \rightarrow V\ NP$ $\Rightarrow$ $VP \rightarrow NP\ V$
4. $PP \rightarrow P\ NP$ $\Rightarrow$ $PP \rightarrow NP\ P$
5. $NP \rightarrow NP_1\ Rel.\ Clause_2$ $\Rightarrow$ $NP \rightarrow Rel.\ Clause_2\ NP_1$

Interlingua

"Mary did not slap the green witch"

```
⎡ EVENT        SLAPPING                                          ⎤
⎢ AGENT        MARY                                              ⎥
⎢ TENSE        PAST                                              ⎥
⎢ POLARITY     NEGATIVE                                          ⎥
⎢              ⎡ WITCH                                       ⎤   ⎥
⎢              ⎢ DEFINITENESS  DEF                           ⎥   ⎥
⎢ THEME        ⎢                                             ⎥   ⎥
⎣              ⎣ ATTRIBUTES    [ HAS-COLOR  GREEN ]          ⎦   ⎦
```
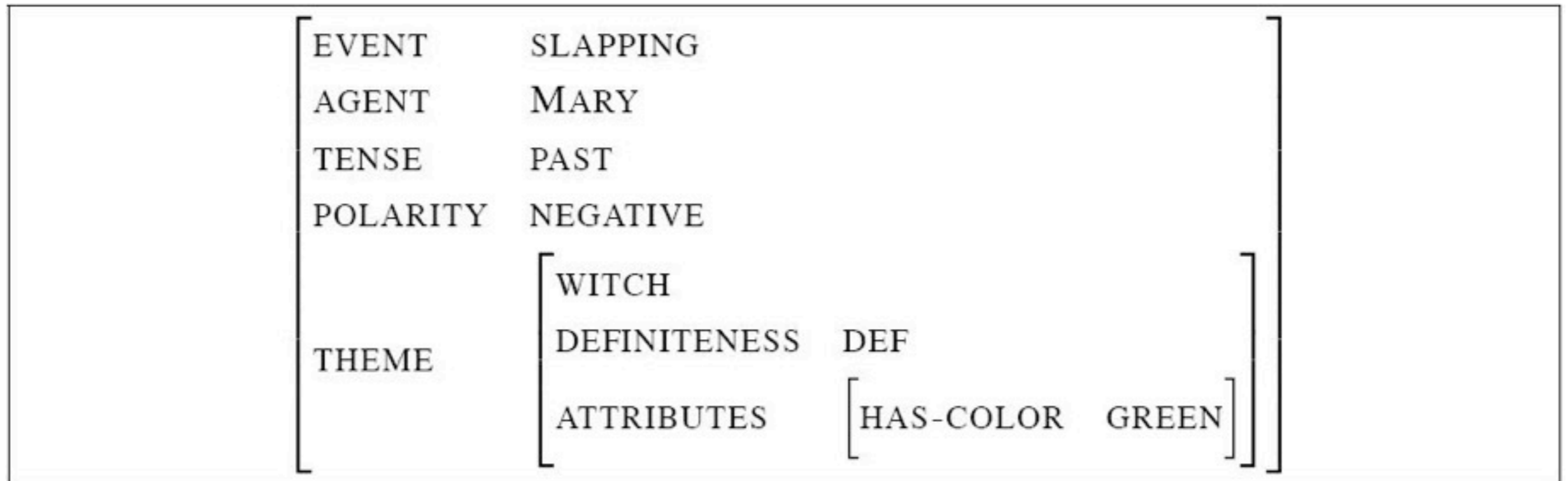
- More like classic logic-based AI
- Works in narrow domains
- Broad domain currently fails
  - Coverage: Knowledge representation for all possible semantics?
  - Can you parse to it?
  - Can you generate from it?

11

Rules are hard

- Coverage
- Complexity (context dependence)
- Maintenance

```
function DIRECT_TRANSLATE_MUCH/MANY(word) returns Russian translation

if preceding word is how return skol'ko
else if preceding word is as return stol'ko zhe
else if word is much
    if preceding word is very return nil
    else if following word is a noun return mnogo
else /* word is many */
    if preceding word is a preposition and following word is a noun return mnogii
    else return mnogo
```

12

Machine learning for MT

- MT as ML: Translation is something people do naturally. Learn rules from data?

- Parallel data: (source, target) text pairs

  - E.g. 20 million words of European Parliament proceedings http://www.statmt.org/europarl/

- Training: learn parameters to predict {source => target}

- Test time: given source sentence, search for high-scoring target (e.g. beam search)

Noisy channel model

Original text → Hypothesized transmission process → Observed text

Inference problem
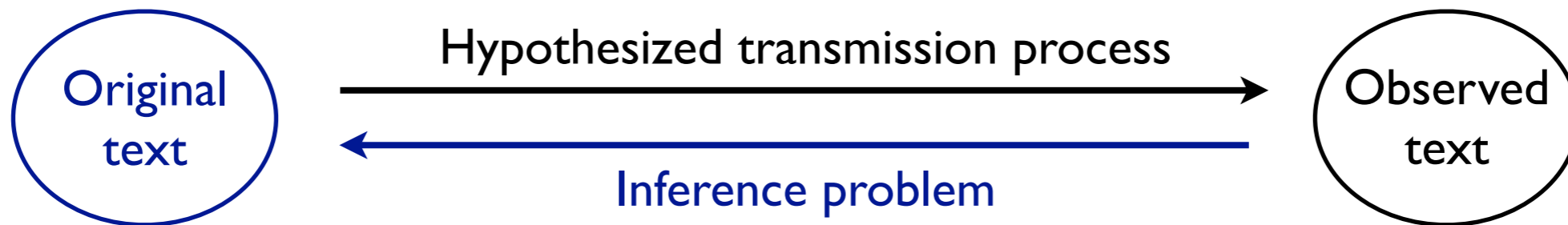
Machine translation

P(target text | source text) $\propto$ P(source text | target text) P(target text)

Noisy channel model

Original text → Observed text

Hypothesized transmission process

Inference problem
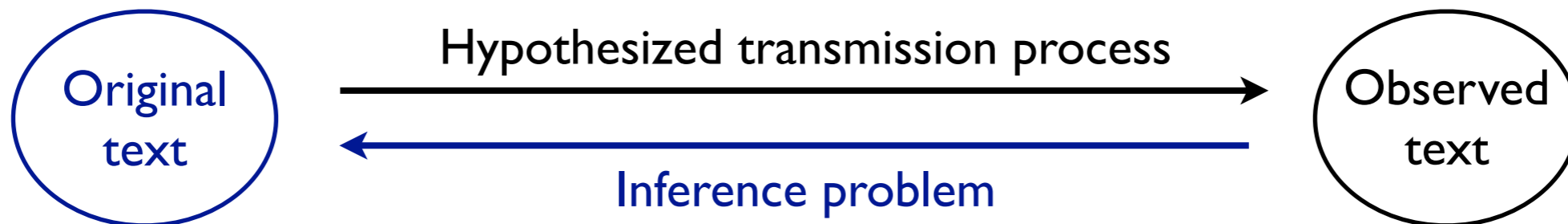
> One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'
>
> -- Warren Weaver (1955)

Machine translation

P(target text | source text) $\propto$ P(source text | target text) P(target text)

Statistical MT

- Pioneered at IBM, early 1990s
  (Forerunner of 90s-era statistical revolution in NLP)

The COLING Paper Review

The validity of statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950. (cf. Hutchins, MT: Past, Present, Future, Ellis Horwood, 1986, pp. 30ff. and references therein) The crude force of computers is not science. The paper is simply beyond the scope of COLING.

Historical notes: http://cs.jhu.edu/~post/bitext/

15

Statistical MT

- Pioneered at IBM, early 1990s
  (Forerunner of 90s-era statistical revolution in NLP)

- Noisy channel model borrowed from
  speech recognition processing

"Every time I fire a linguist,
the performance of the speech recognizer goes up"
[Fred Jelinek]

IBM Models

- [Brown et al. 1993, "The Mathematics of Statistical Machine Translation: Parameter Estimation"]

- *Lexical translations*: each source word has word-level translations to target language

- *Alignments*: hypothesizes that individual input words get translated to outputs (potentially in different order)

- Training
  - Problem: don't know which came from which!
  - Solution: use the *EM algorithm*

Example: learning with parallel data

Sentence pair 3 is much more challenging. So far, we have

erok   sprok   izok   hihok   ghirok

totat   dat   arrat   vat   hilat

The Centauri word *izok* would be translated as either *totat*, *arrat*, or *vat*, yet when you look at *izok* in sentence pair 6, none of those three words appear in the Arcturan. Therefore, *izok* appears to be ambiguous. The word *hihok*, however, is fixed in sentence pair 11 as *arrat*. Both sentence pairs 3 and 12 have *izok hihok* sitting directly on top of *arrat vat*; so, in all possibility, *vat* seems a reasonable translation for (ambiguous) *izok*. Sentence pairs 5, 6, and 9 suggest that *quat* is its other translation. Through process of elimination, you connect the words *erok* and *totat*, finishing off the analysis:

erok   sprok   izok hihok   ghirok

totat   dat   arrat   vat   hilat

Automating Knowledge Acquisition for Machine Translation

Kevin Knight

Notice that aligning the sentence pairs helps you to build the translation dictionary and that building the translation dictionary also helps you decide on correct alignments. You might call this the *decipherment method*.

18

Example: learning with parallel data

1a. Garcia and associates.
1b. Garcia y asociados.

2a. Carlos Garcia has three associates.
2b. Carlos Garcia tiene tres asociados.

3a. his associates are not strong.
3b. sus asociados no son fuertes.

4a. Garcia has a company also.
4b. Garcia tambien tiene una empresa.

5a. its clients are angry.
5b. sus clientes están enfadados.

6a. the associates are also angry.
6b. los asociados tambien están enfadados.

7a. the clients and the associates are enemies.
7b. los clientes y los asociados son enemigos.

8a. the company has three groups.
8b. la empresa tiene tres grupos.

9a. its groups are in Europe.
9b. sus grupos están en Europa.

10a. the modern groups sell strong pharmaceuti-cals.
10b. los grupos modernos venden medicinas fuertes.

11a. the groups do not sell zanzanine.
11b. los grupos no venden zanzanina.

12a. the small groups are not modern.
12b. los grupos pequeños no son modernos.

19

Lexical Translation

- How do we translate a word? Look it up in the dictionary

$$Haus : house, home, shell, household$$

- Multiple translations

  - Different word senses, different registers, different inflections (?)

  - *house, home* are common

  - *shell* is specialized (the Haus of a snail is a shell)

How common is each?

Translation	Count
house	5000
home	2000
shell	100
household	80

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$

- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences

$$\mathbf{e} = \langle e_1, e_2, \dots, e_m \rangle \qquad \mathbf{f} = \langle f_1, f_2, \dots, f_n \rangle$$

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$

- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences

- Lexical translation makes the following **assumptions**:

  - Each word in e_i in $\mathbf{e}$ is generated from exactly one word in $\mathbf{f}$

  - Thus, we have an *alignment* a_i that indicates which word e_i "came from", specifically it came from f_{a_i}.

  - Given the alignments $\mathbf{a}$, translation decisions are conditionally independent of each other and depend *only* on the aligned source word f_{a_i}.

Lexical Translation

$$\mathbf{e} = \langle e_1, e_2, ...e_m \rangle \qquad \mathbf{f} = \langle f_1, f_2, ...f_n \rangle$$

$$\mathbf{a} = \langle a_1, a_2, ...a_m \rangle \quad \text{each } a_i \in \{0, 1, ..., n\}$$

Modeling assumptions

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in \{0, 1, .., n\}^m} p(\mathbf{a} \mid \mathbf{f}, m) \times \prod_{i=1}^{m} p(e_i \mid f_{a_i})$$

[Alignment] x [Translation | Alignment]

Alignment

- Alignments can be visualized in by drawing links between two sentences, and they are represented as vectors of positions:

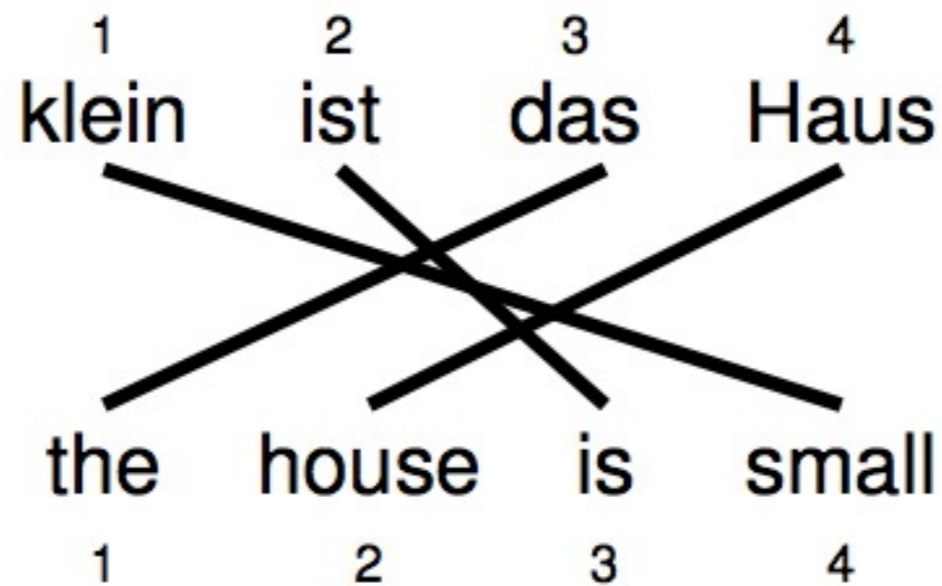$$\mathbf{f} = \begin{array}{cccc} \overset{1}{\text{das}} & \overset{2}{\text{Haus}} & \overset{3}{\text{ist}} & \overset{4}{\text{klein}} \end{array}$$

$$\mathbf{e} = \begin{array}{cccc} \underset{1}{\text{the}} & \underset{2}{\text{house}} & \underset{3}{\text{is}} & \underset{4}{\text{small}} \end{array}$$

$$\mathbf{a} = (1, 2, 3, 4)^{\top}$$

Reordering

- Words may be reordered during translation.

$$\mathbf{a} = (3, 4, 2, 1)^{\top}$$

Word Dropping

- A source word may not be translated at all
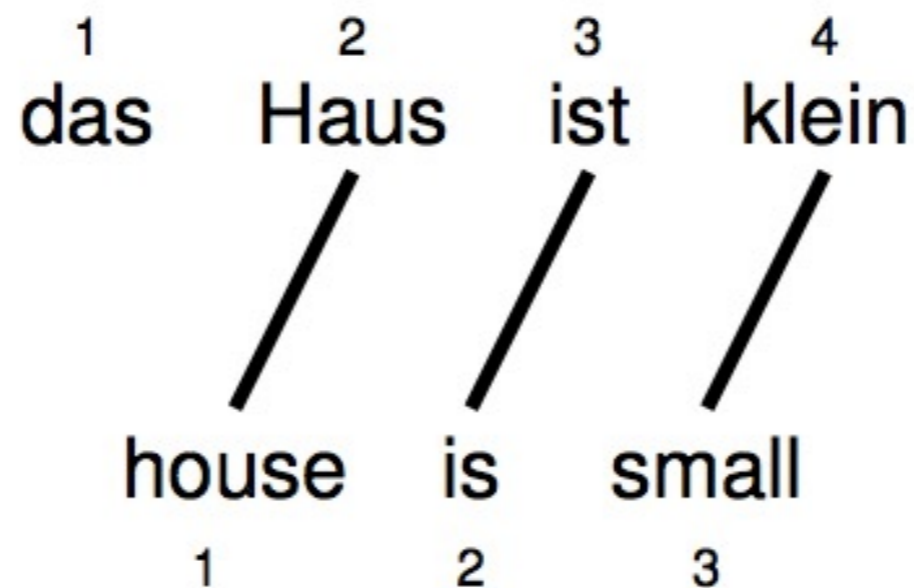
$$\mathbf{a} = (2, 3, 4)^{\top}$$

Word Insertion

- Words may be inserted during translation

  English just does not have an equivalent

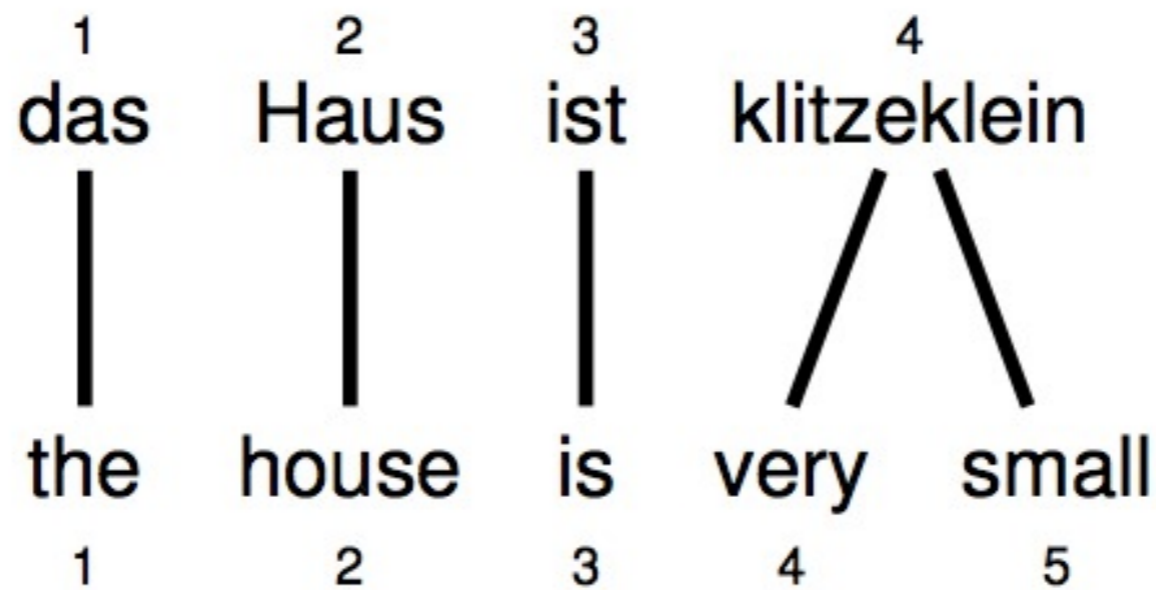 But it must be explained - we typically assume every source sentence contains a NULL token
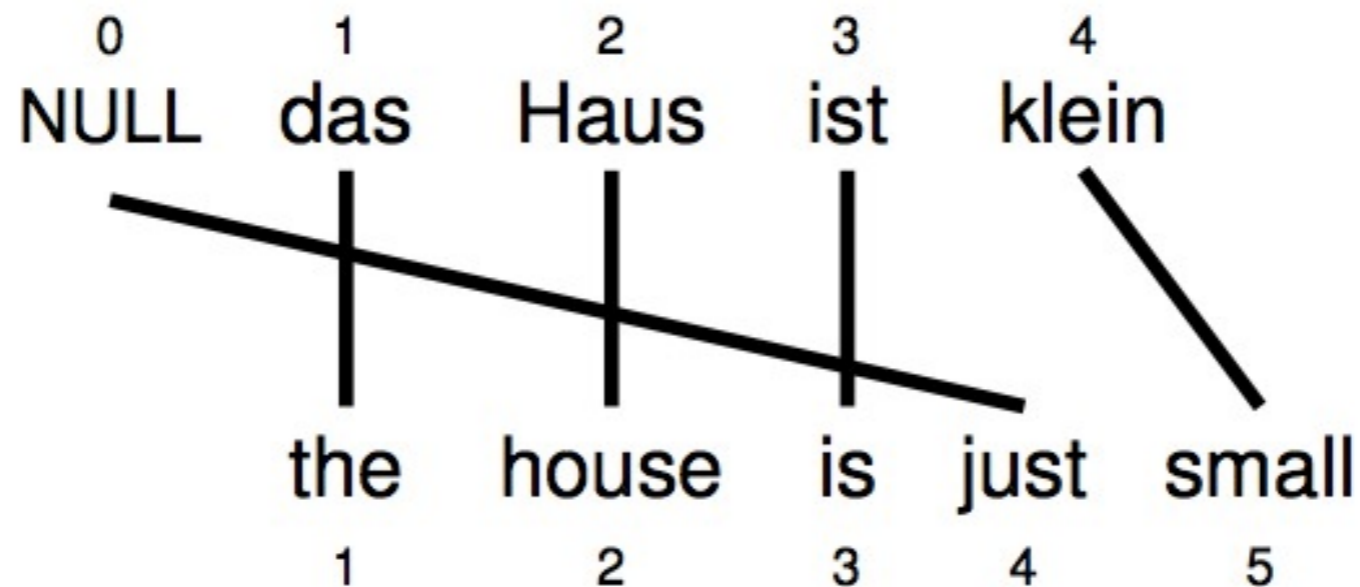
$$\mathbf{a} = (1, 2, 3, 0, 4)^{\top}$$

One-to-many Translation

- A source word may translate into **more than one** target word

$$\mathbf{a} = (1, 2, 3, 4, 4)^{\top}$$

Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$

Many-to-one Translation

- **More than one source word** may **not** translate as a unit in lexical translation

$$
\begin{array}{cccc}
1 & 2 & 3 & 4 \\
\text{das} & \text{Haus} & \text{brach} & \text{zusammen}
\end{array}
$$

$$
\begin{array}{ccc}
\text{the} & \text{house} & \text{collapsed} \\
1 & 2 & 3
\end{array}
$$

$\mathbf{a} = ???$

[IBM Model 1 can't do this]

IBM Model 1: Inference and learning

- Inferring alignments: assume lexical translations are independent conditional on alignments. That implies it's easy to compute

$$p(\mathbf{a} \mid \mathbf{e}, \mathbf{f}, \theta)$$

- How do we learn translation parameters?

$$\arg\max_{\theta} p(\mathbf{e} \mid \mathbf{f}, \theta)$$

- Chicken and egg problem:
  If we knew alignments, translation parameters would be trivial (just counting):

$$\arg\max_{\theta} p(\mathbf{e} \mid \mathbf{a}, \mathbf{f}, \theta)$$

EM Algorithm

- pick some random (or uniform) parameters

- Repeat until you get bored (~ 5 iterations for lexical translation models)

  - using your current parameters, compute "expected" alignments for every target word token in the training data

  $$p(a_i \mid \mathbf{e}, \mathbf{f}) \qquad \text{(on board)}$$

  - keep track of the expected number of times f translates into e throughout the whole corpus

  - keep track of the expected number of times that f is used as the source of any translation

  - use these expected counts as if they were "real" counts in the standard MLE equation

EM for Model 1

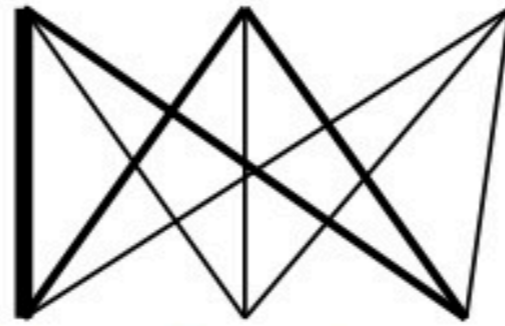... la maison ... la maison blue ... la fleur ...
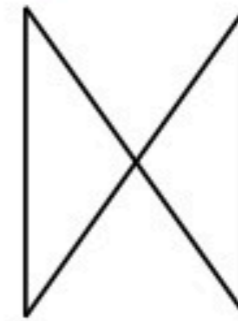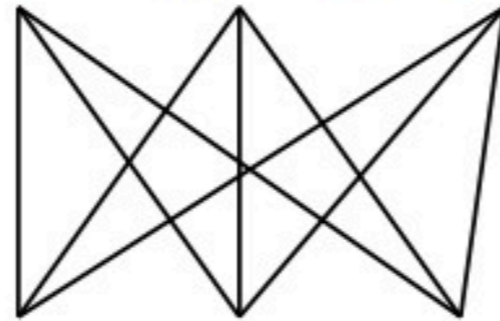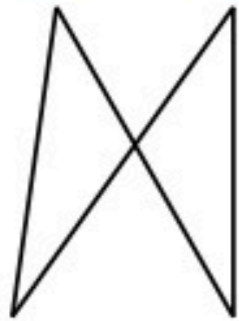
... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely

- Model learns that, e.g., la is often aligned with the

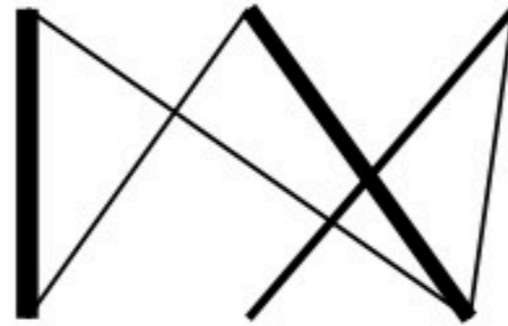# EM for Model 1

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- After one iteration

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

EM for Model 1

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

- After another iteration

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

EM for Model 1

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

- Convergence

- Inherent hidden structure revealed by EM

EM for Model 1

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

```
p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...
```

- Parameter estimation from the aligned corpus

Convergence

das	Haus	das	Buch	ein	Buch
the	house	the	book	a	book

e	f	initial	1st it.	2nd it.	3rd it.	...	final
the	das	0.25	0.5	0.6364	0.7479	...	1
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.5	0.6364	0.7479	...	1
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.5	0.4286	0.3466	...	0
a	ein	0.25	0.5	0.5714	0.6534	...	1
the	haus	0.25	0.5	0.4286	0.3466	...	0
house	haus	0.25	0.5	0.5714	0.6534	...	1

EM algorithm

- Very general meta-algorithm when we have
  - observed data **x**
  - latent (hidden) variables **z**
  - parameters θ
- and
  - we want $\text{argmax}_\theta P(x \mid \theta)$ but it's intractable
  - but this is easy: $\text{argmax}_\theta P(x,z \mid \theta)$
- EM: iterate
  - E-step: Infer $P(z \mid x, \theta)$ [[make your best guess]]
  - M-step: Infer with the usual MLE but with *weighted* counts from the E-step
- Many applications in NLP:
  - Unsupervised training of HMMs, NB, topic models...
  - Semi-supervised learning: see only some of the labels

39

MT Evaluation

- Problem: {source => target} has very large output space
- Ideally: bilingual humans judge every (source, translation) pair. Typically Likert scale for:
  - Faithfulness
  - Fluency [[monolingual humans can do this one]]
- *BLEU score:* an automatic metric
  - Given (source, target) gold-standard, score translation
  - ~Precision of translation's ngrams: are they in the gold-standard translation?
  - Brevity penalty (so can't game it with short sentences)
- Problem: there are multiple legitimate ways to say the same thing!
  - => Use multiple alternate gold-standard translations

40

Multiple Reference Translations

Reference translation 1:
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 .

Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

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.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:
US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

- Statistical MT:
  - Phrase-based models (extension of lexical models) are common (e.g. Moses system)
  - Syntactic and other models too

- ...but! These all just changed!

- Neural MT
  - e.g. Google's 2016 paper
  - Use NN representations of words and sentences
  - Alignment models reborn as *attention-based neural networks*: choose which source words to look at, when translating next target word

- How to incorporate linguistic knowledge and constraints?

42