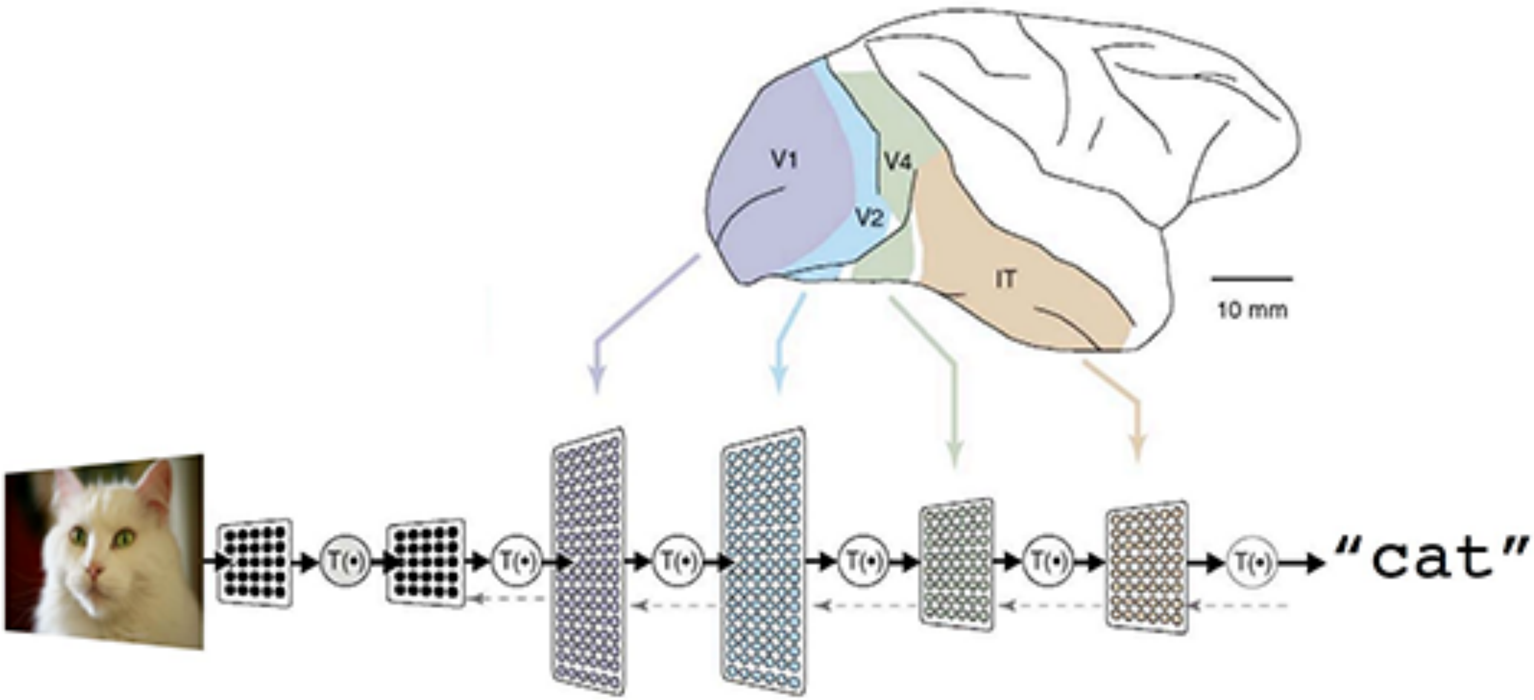
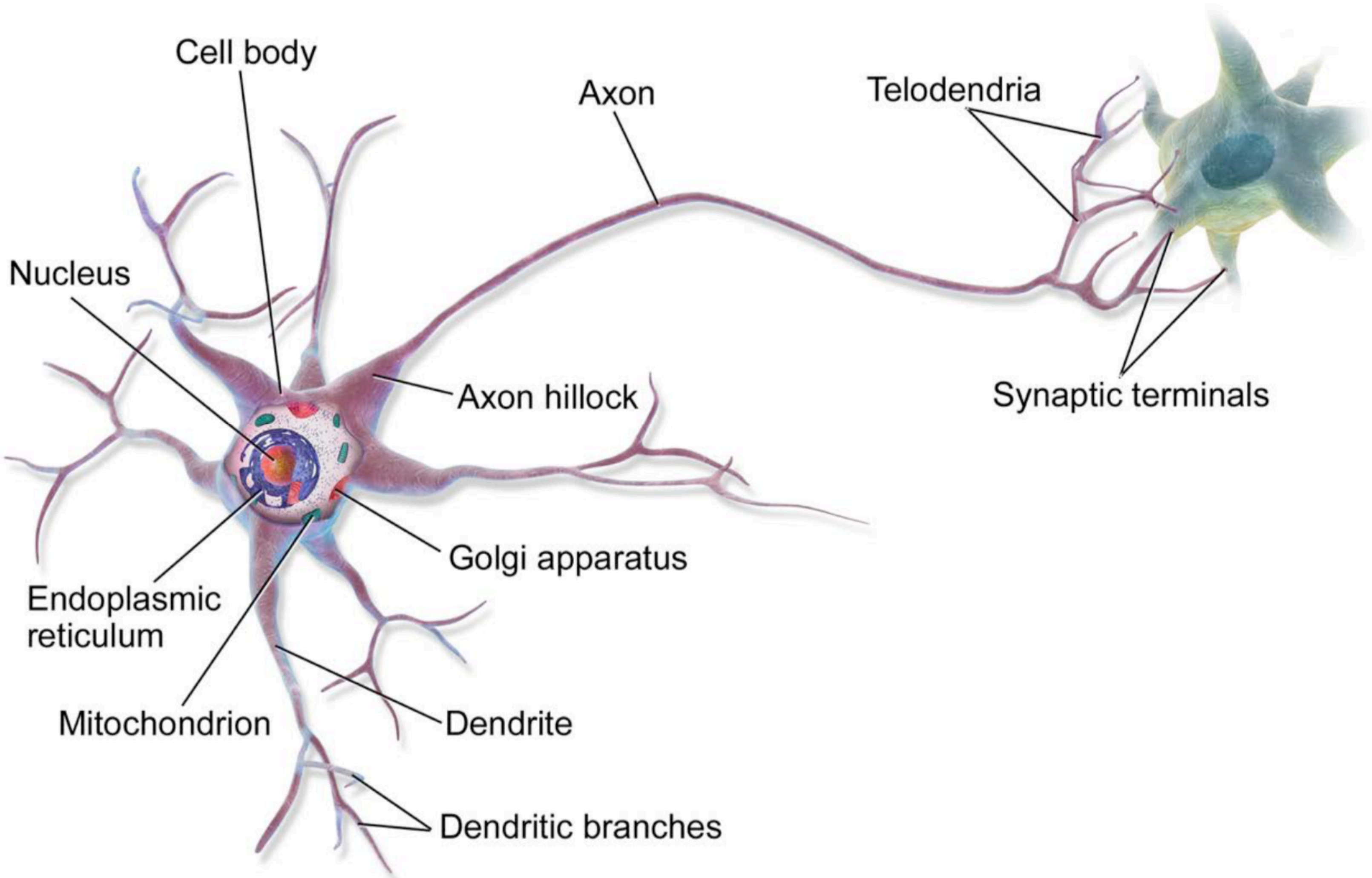
Introduction to Neural networks: Neural Machine Translation - and earlier MT architectures

DAAD Training 2021

IÜD, Universität Heidelberg

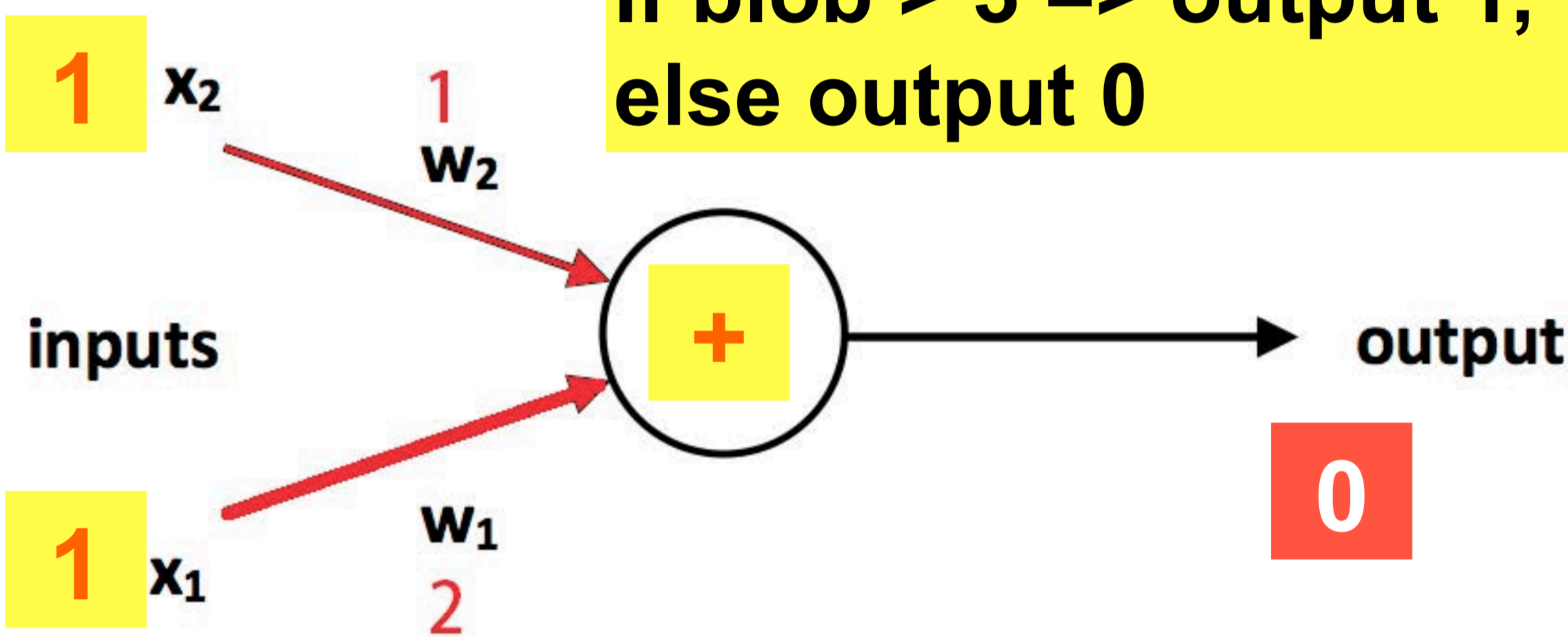
Neural MT: brain metaphor see: Koehn, P. 2020 Neural machine translation. Cambridge University Press

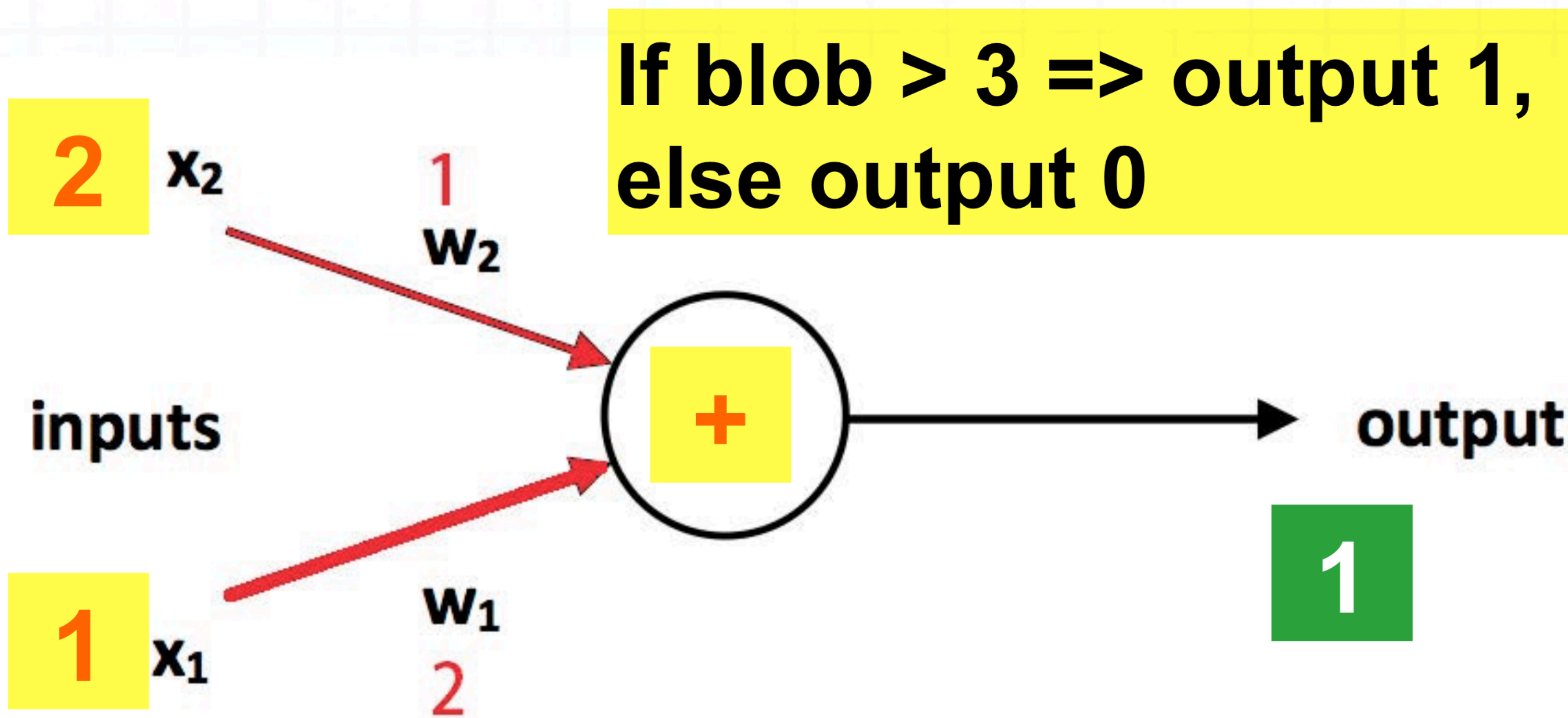


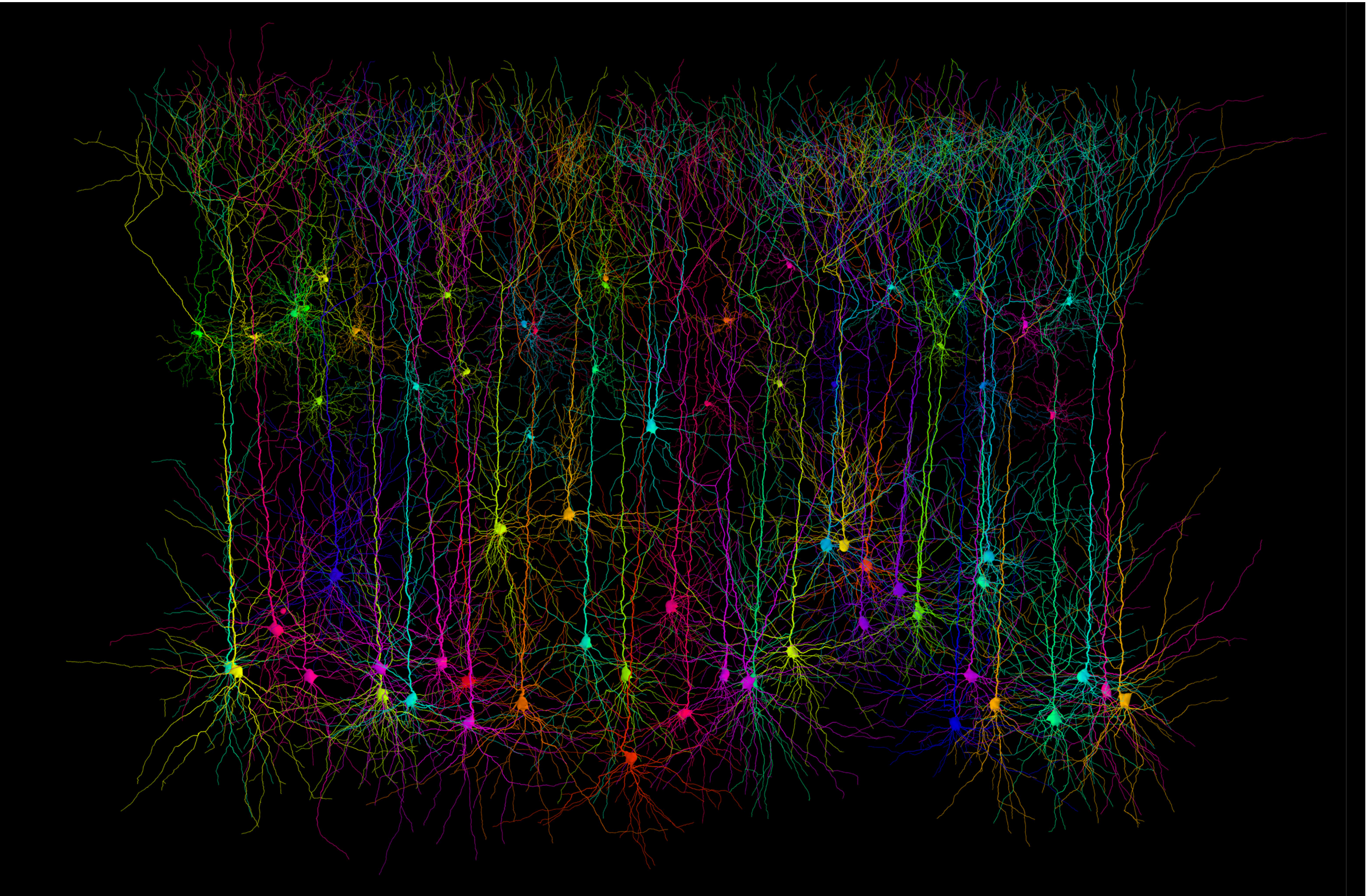


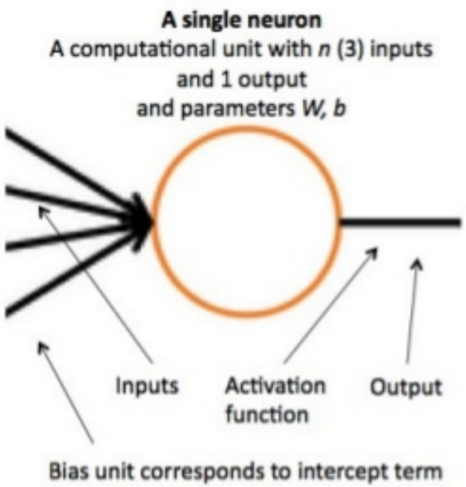
Cell body
Axon
Telodendria
Nucleus
Axon hillock
Synaptic terminals
Golgi apparatus
Endoplasmic reticulum
Mitochondrion
Dendrite
Dendritic branches

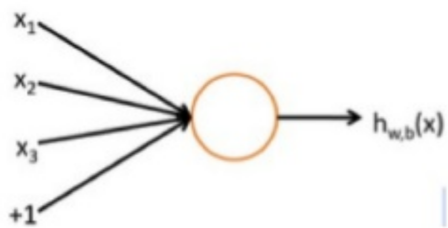
If blob > 3 => output 1, else output 0

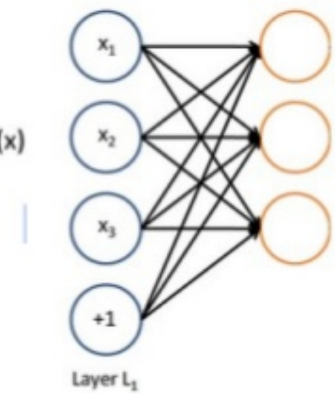


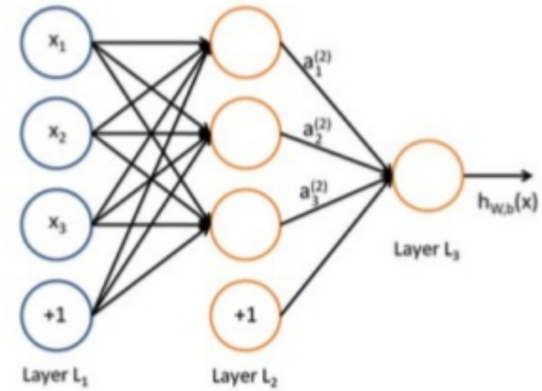

If blob > 3 => output 1,
else output 0
2
x_2
1
w_2
inputs
+
output
1
1
x_1
w_1
2

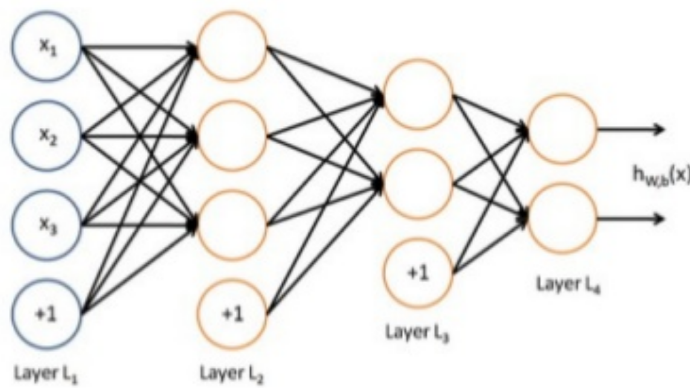



A single neuron
A computational unit with n (3) inputs
and 1 output
and parameters W, b
Inputs
Activation function
Output
Bias unit corresponds to intercept term

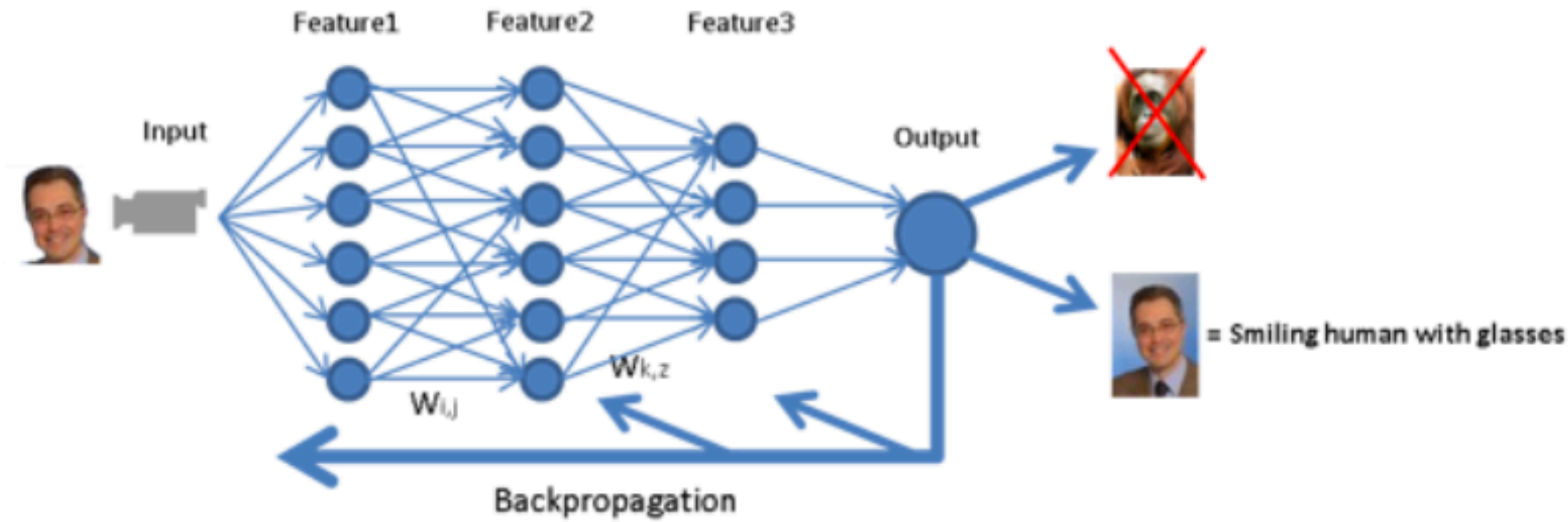

x_1
x_2
x_3
+1
$h_{W,b}(x)$

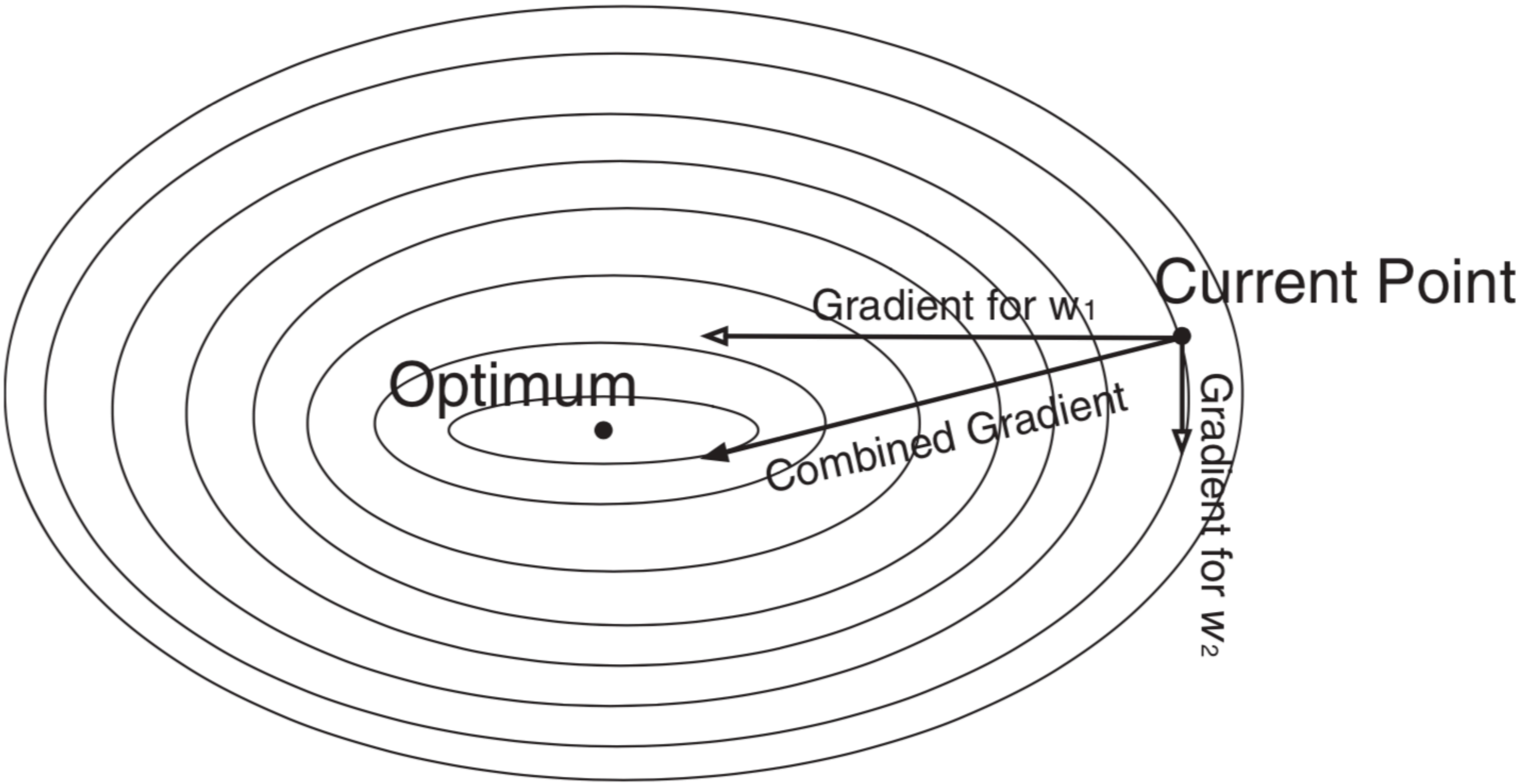

x_1
x_2
x_3
+1
Layer L_1


x_1
x_2
x_3
+1
+1
$a_1^{(2)}$
$a_2^{(2)}$
$a_3^{(2)}$
$h_{W,b}(x)$
Layer L_1
Layer L_2
Layer L_3


x_1
x_2
x_3
+1
+1
+1
$h_{W,b}(x)$
Layer L_1
Layer L_2
Layer L_3
Layer L_4

3/34

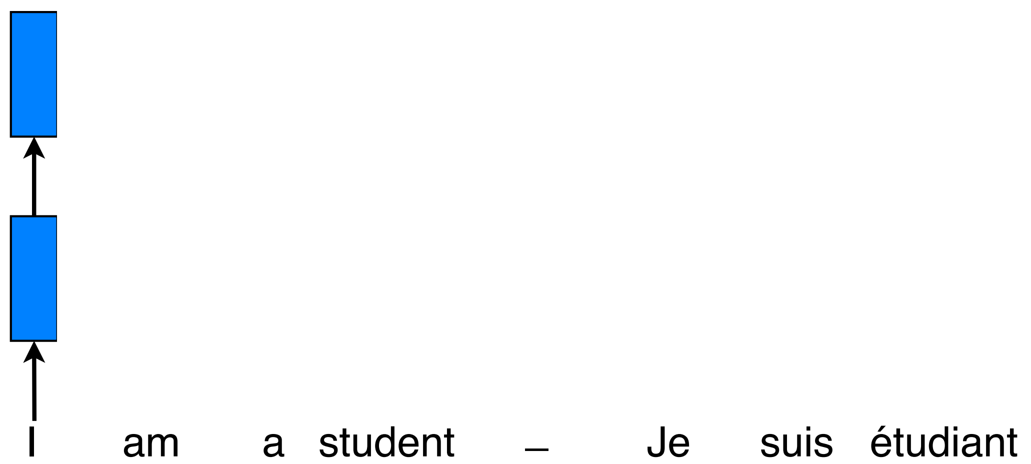

Feature1
Feature2
Feature3
Input
Output
= Smiling human with glasses
$W_{k,z}$
$W_{i,j}$
Backpropagation



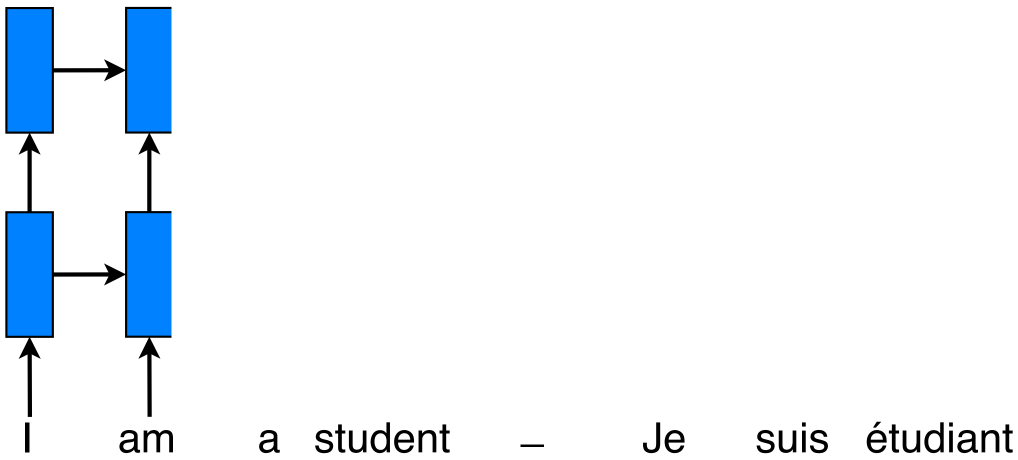
Current Point
Gradient for w1
Optimum
Combined Gradient
Gradient for w2

I am a student _ Je suis étudiant

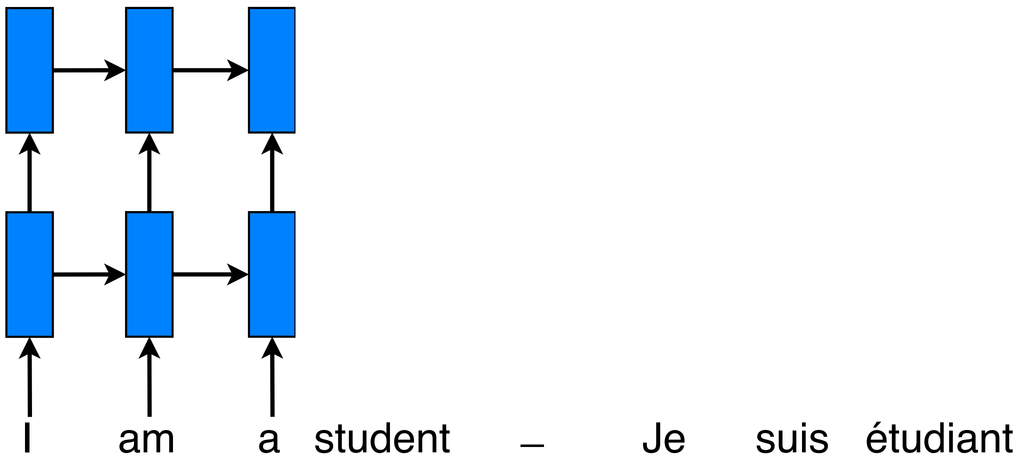
- Big RNNs trained end-to-end.



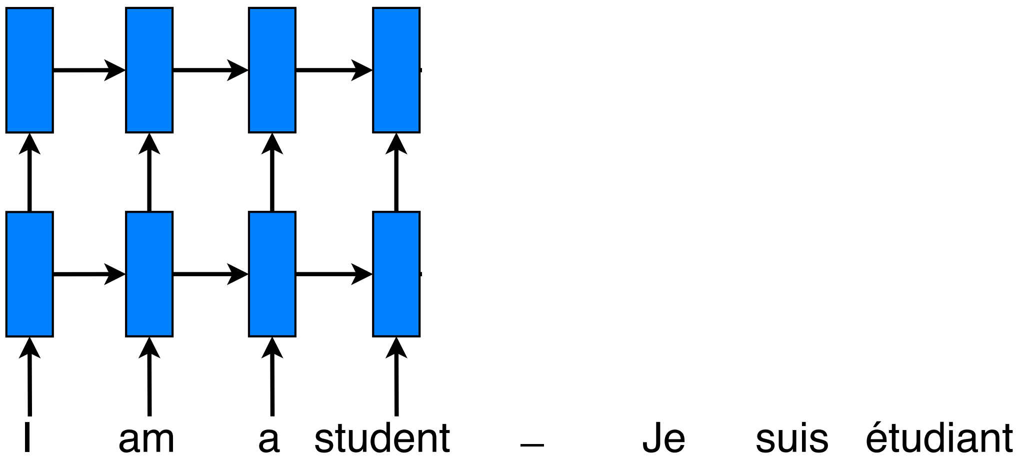
- Big RNNs trained end-to-end.



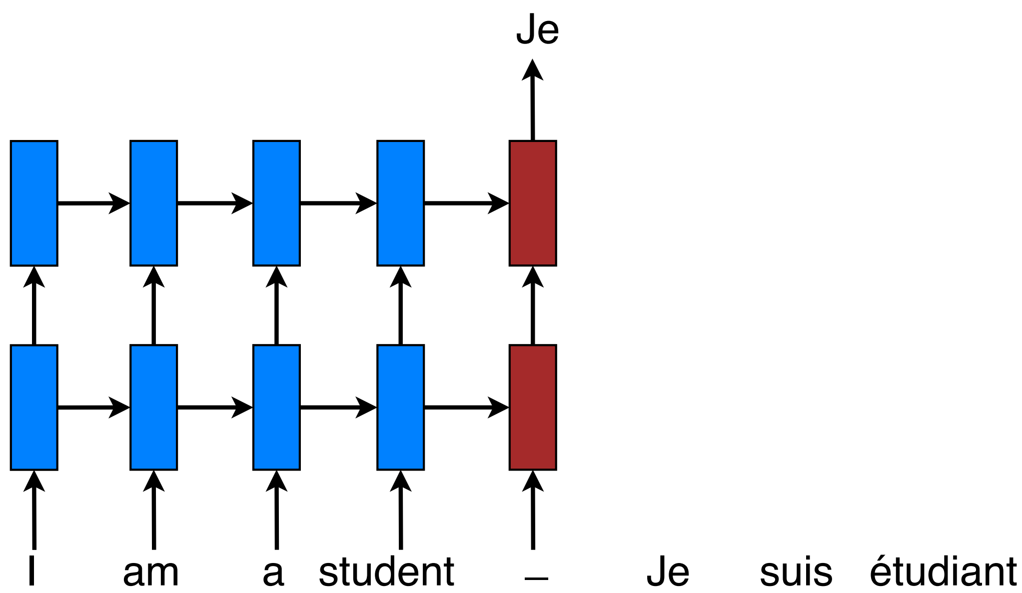
- Big RNNs trained end-to-end.



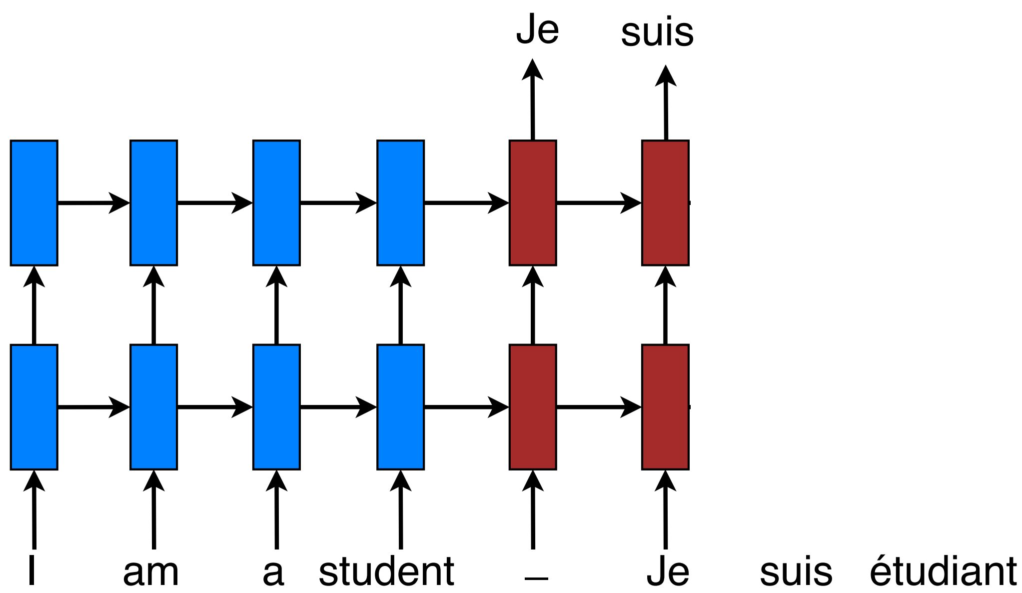
- Big RNNs trained end-to-end.



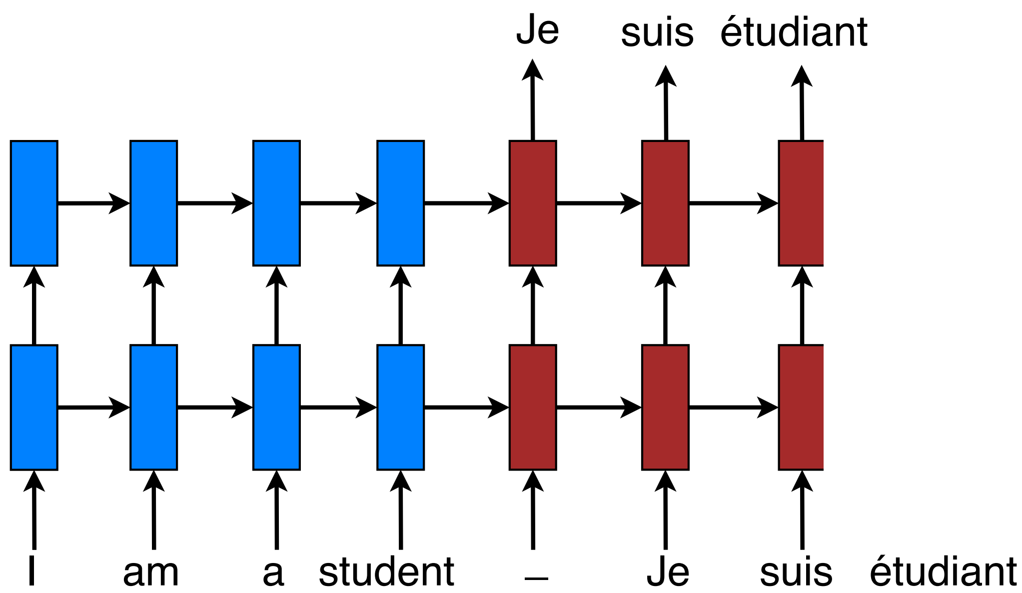
- Big RNNs trained end-to-end.



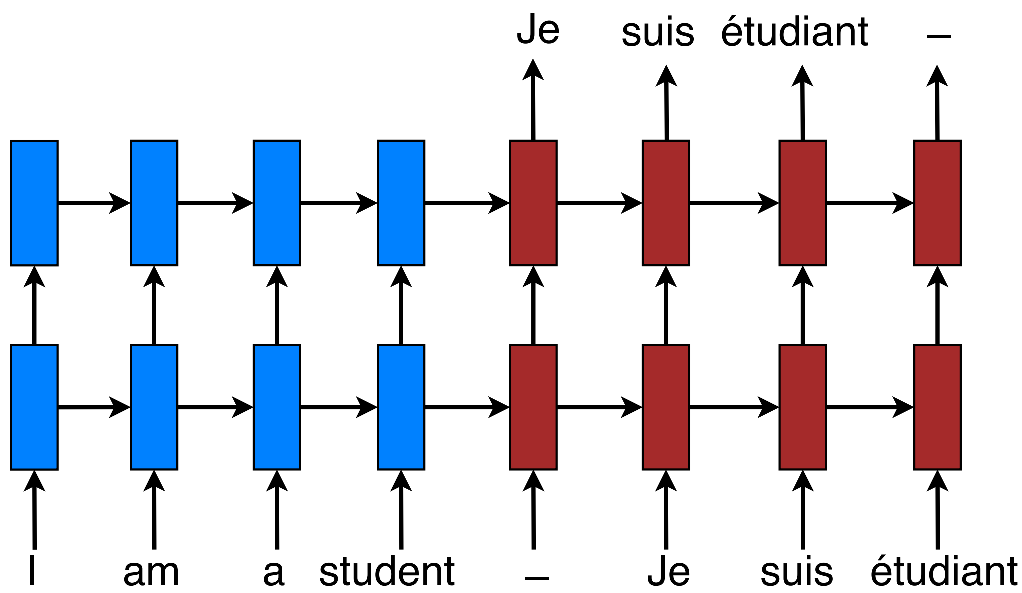
- Big RNNs trained end-to-end: encoder-decoder.



- Big RNNs trained end-to-end: encoder-decoder.

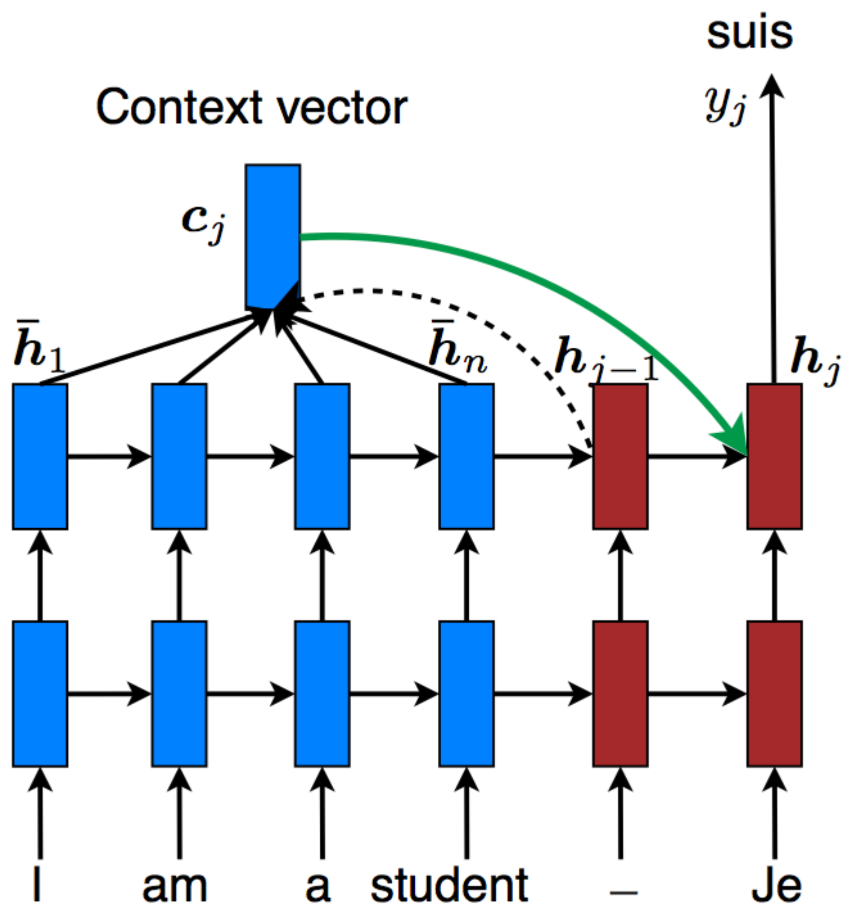


- Big RNNs trained end-to-end: encoder-decoder.



- Big RNNs trained end-to-end: encoder-decoder.
  - Generalize well to long sequences.
  - Small memory footprint.
  - Simple decoder.

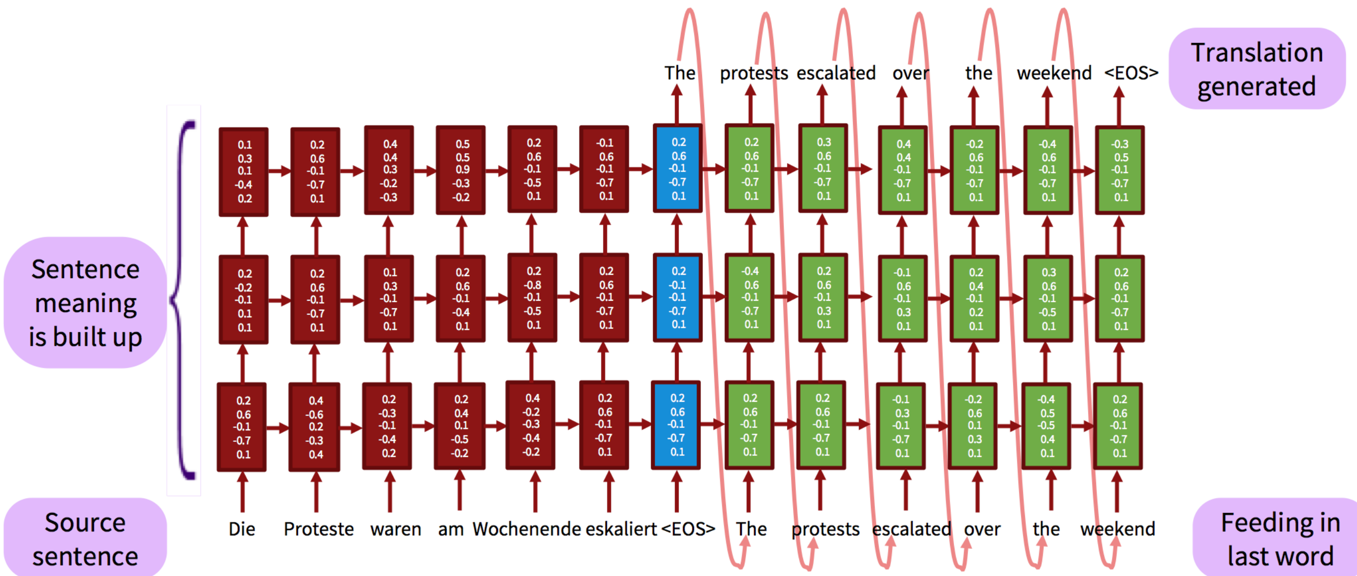
Attention mechanism



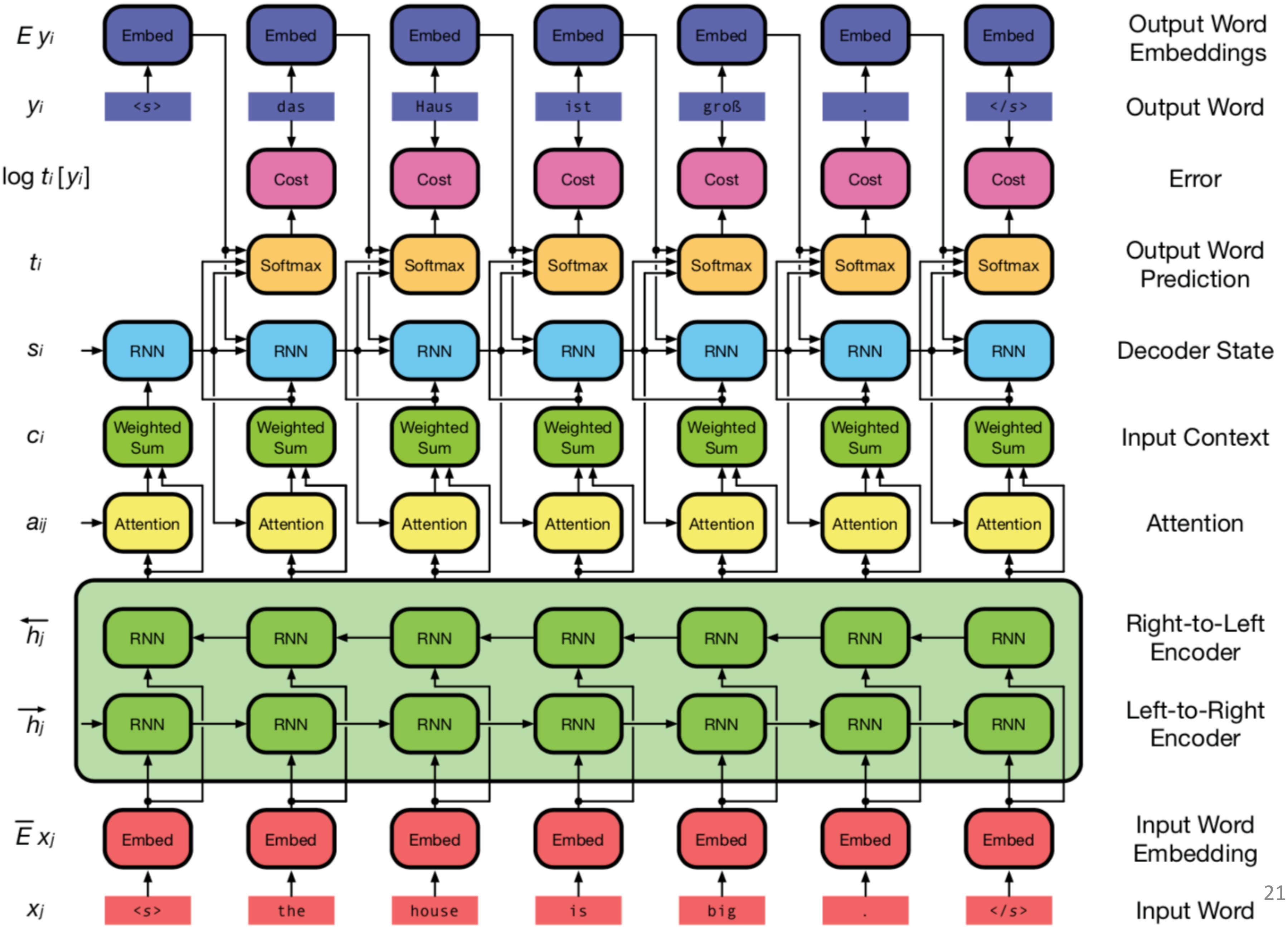
Further developments

Modern Sequence Models for NMT

[Sutskever et al. 2014, Bahdanau et al. 2014, et seq.]
following [Jordan 1986] and more closely [Elman 1990]



A deep recurrent neural network


E yi
yi
- log ti [yi]
ti
si
ci
aij
hj
hj
E xj
xj
Embed
<s>
das
Haus
ist
groß
.
</s>
Cost
Softmax
RNN
Weighted Sum
Attention
the
house
is
big
Output Word Embeddings
Output Word
Error
Output Word Prediction
Decoder State
Input Context
Attention
Right-to-Left Encoder
Left-to-Right Encoder
Input Word Embedding
Input Word

Decoder (translation in real time)

ich glaube aber auch, er ist clever genug um seine Aussagen vage genug zu halten, so dass sie auf verschiedene Art und Weise interpretiert werden können.

Output Word Predictions

Best		Alternatives
but	(42.1%)	*however* (25.3%), *I* (20.4%), *yet* (1.9%), *and* (0.8%), *nor* (0.8%), ...
I	(80.4%)	*also* (6.0%), , (4.7%), *it* (1.2%), *in* (0.7%), *nor* (0.5%), *he* (0.4%), ...
also	(85.2%)	*think* (4.2%), *do* (3.1%), *believe* (2.9%), , (0.8%), *too* (0.5%), ...
believe	(68.4%)	*think* (28.6%), *feel* (1.6%), *do* (0.8%), ...
he	(90.4%)	*that* (6.7%), *it* (2.2%), *him* (0.2%), ...
is	(74.7%)	*'s* (24.4%), *has* (0.3%), *was* (0.1%), ...
clever	(99.1%)	*smart* (0.6%), ...
enough	(99.9%)	
to	(95.5%)	*about* (1.2%), *for* (1.1%), *in* (1.0%), *of* (0.3%), *around* (0.1%), ...
keep	(69.8%)	*maintain* (4.5%), *hold* (4.4%), *be* (4.2%), *have* (1.1%), *make* (1.0%), ...
his	(86.2%)	*its* (2.1%), *statements* (1.5%), *what* (1.0%), *out* (0.6%), *the* (0.6%), ...
statements	(91.9%)	*testimony* (1.5%), *messages* (0.7%), *comments* (0.6%), ...
vague	(96.2%)	*v@@* (1.2%), *in* (0.6%), *ambiguous* (0.3%), ...
enough	(98.9%)	*and* (0.2%), ...
so	(51.1%)	, (44.3%), *to* (1.2%), *in* (0.6%), *and* (0.5%), *just* (0.2%), *that* (0.2%), ...
they	(55.2%)	*that* (35.3%), *it* (2.5%), *can* (1.6%), *you* (0.8%), *we* (0.4%), *to* (0.3%), ...
can	(93.2%)	*may* (2.7%), *could* (1.6%), *are* (0.8%), *will* (0.6%), *might* (0.5%), ...
be	(98.4%)	*have* (0.3%), *interpret* (0.2%), *get* (0.2%), ...
interpreted	(99.1%)	*interpre@@* (0.1%), *constru@@* (0.1%), ...
in	(96.5%)	*on* (0.9%), *differently* (0.5%), *as* (0.3%), *to* (0.2%), *for* (0.2%), *by* (0.1%),
different	(41.5%)	*a* (25.2%), *various* (22.7%), *several* (3.6%), *ways* (2.4%), *some* (1.7%), ...
ways	(99.3%)	*way* (0.2%), *manner* (0.2%), ...
.	(99.2%)	</s> (0.2%), , (0.1%), ...

N-best list for translations: *‘Er wollte nie an irgendeiner Art von Auseinandersetzung teilnehmen’*

He never wanted to participate in any kind of confrontation.
He never wanted to take part in any kind of confrontation.
He never wanted to participate in any kind of argument.
He never wanted to take part in any kind of argument.
He never wanted to participate in any sort of confrontation.
He never wanted to take part in any sort of confrontation.
He never wanted to participate in any sort of argument.
He never wanted to take part in any sort of argument.
He never wanted to participate in any kind of controversy.
He never wanted to take part in any kind of controversy.
He never intended to participate in any kind of confrontation.
He never intended to take part in any kind of confrontation.
He never wanted to take part in some sort of confrontation.
He never wanted to take part in any sort of controversy.

BLEU / NIST – automated MT evaluation metrics

- Evaluating the quality of Machine Translation
  - precision vs. recall

Reference proximity methods

- Assumption of Reference Proximity (ARP):
  - "*…the closer the machine translation is to a professional human translation, the better it is*" (Papineni et al., 2002: 311)
- Finding a distance between 2 texts
  - Minimal edit distance
  - Word & word sequences overlap (N-gram distance)
- BLEU (BiLingual Evaluation Understudy) implements N-gram overlap (precision of N-gram matches)
  - Try it out: http://corpus.leeds.ac.uk/corpuslabs/lab201801cgibleu/

Proximity to human reference (1)

- MT "Systran": *The 38 heads of* undertaking put in examination *in the* file were the subject of hearings [...] in *the* tread of "*political*" *confrontation*.

- Human translation "Expert": *The 38 heads of* companies questioned *in the* case had been heard [...] following *the* "*political*" *confrontation*.

- MT "Candide": *The 38* counts *of* company put into consideration *in the case had* the object of hearings [...] in *the* path of confrontal "*political*."

Proximity to human reference (2)

- MT “Systran”: *The 38 heads of* undertaking put in examination *in the* file were the subject of hearings [...] in *the* tread of "*political*" *confrontation*.

- Human translation “Expert”: *The 38 heads of* companies questioned *in the* case had been heard [...] following *the* "*political*" *confrontation*.

- MT “Candide”: *The 38* counts *of* company put into consideration *in the case had* the object of hearings [...] in *the* path of confrontal "*political*."

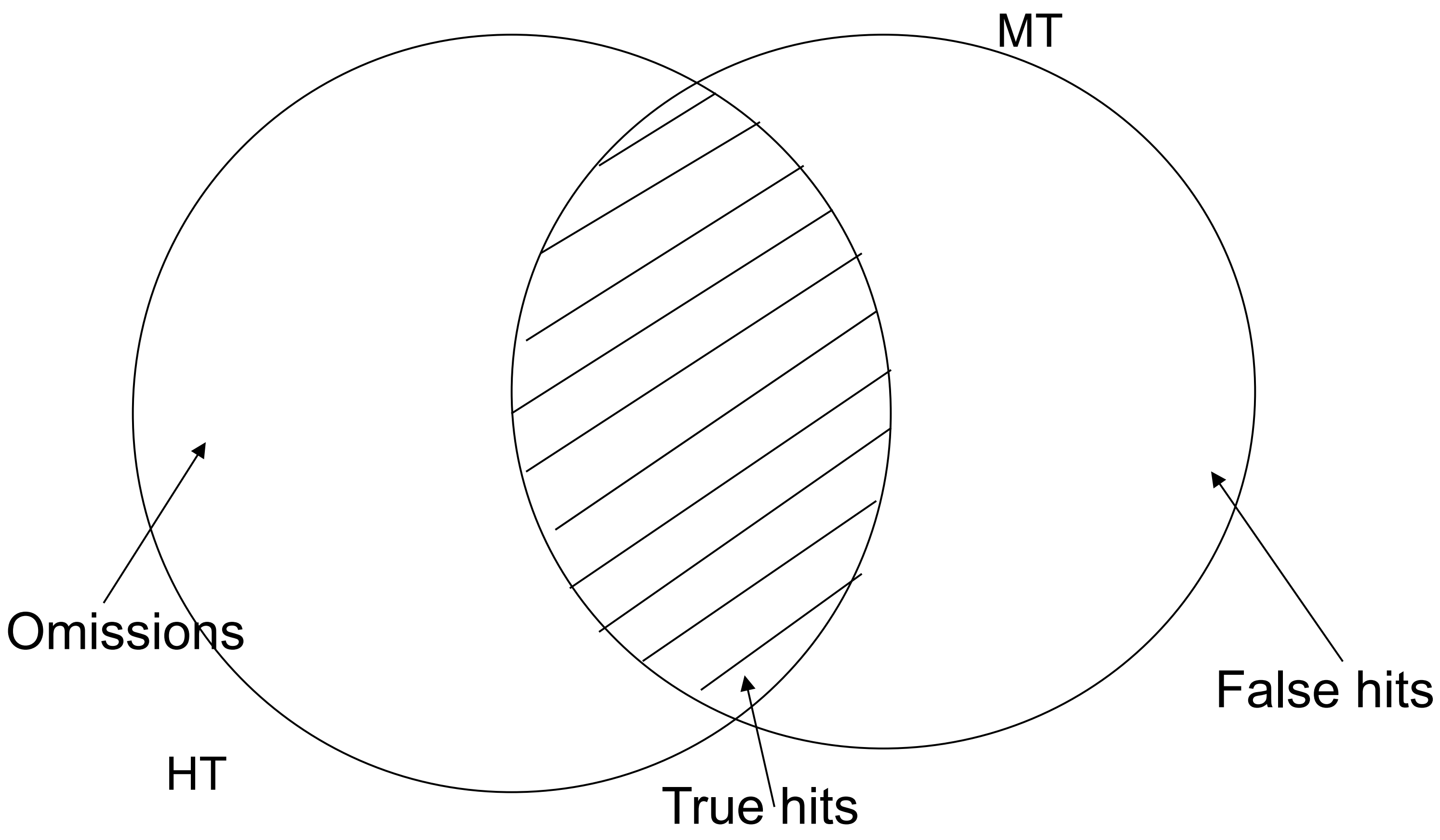
Proximity to human reference (3)

- MT “Systran”: *The 38 heads of* undertaking put in examination *in the* file were the subject of hearings [...] in *the* tread of "*political*" *confrontation*.

- Human translation “Expert”: *The 38* ***heads*** *of* companies questioned *in the* **case had** been heard [...] following *the* "*political*" ***confrontation***.

- MT “Candide”: *The 38* counts *of* company put into consideration *in the case had* the object of hearings [...] in *the* path of confrontal "*political*."

Matches of N-grams

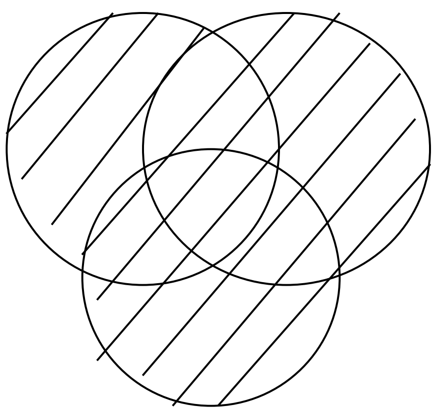


Legitimate translation variation (LTV)

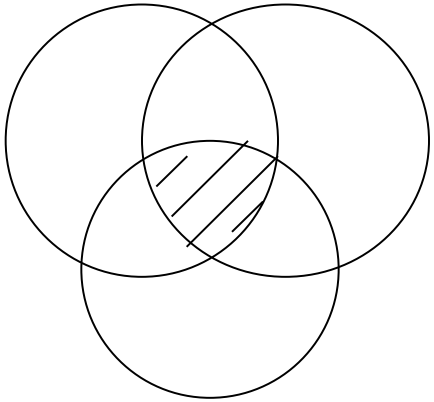
- to which human translation should we compute the edit distance?
- is it possible to integrate both human translations into a reference set?

N-grams: Union and Intersection

- Union ~Precision



Intersection ~Recall



Experiments with BLEU

- http://corpus.leeds.ac.uk/corpuslabs/lab201801cgibleu/

- Try simple sentence first (test and ref need to be in the same language, e.g.,:
  - Test: `the fat cat is sitting on a very warm rug`
  - Reference: `the fat cat is sitting on a very warm mat`
- MT-translate real text with Google, Bing, Systran into your target language
  - Which system gets best scores?

BLEU extensions: METEOR, etc.

- NIST (National Institute of STandard's)
    - improved version of BLEU: repetitions (NIST 2005),
- METEOR
    - integrates additional linguistic features, such synonyms and stems, or dictionary forms of the inflected words found in the evaluated texts) (Banerjee and Lavie, 2005),
- WNM (Weighted N-gram model)
    - takes into account statistical salience scores, and assigns more weight to topic-specific terms and named entities (Babych and Hartley, 2004)

Types of automated MT evaluation

- Automatic Evaluation is more recent: first methods appeared in the late 90-ies
  - Performance methods
    - Measuring performance of some system which uses degraded MT output
  - Reference proximity methods
    - Measuring distance between MT and a “gold standard” translation

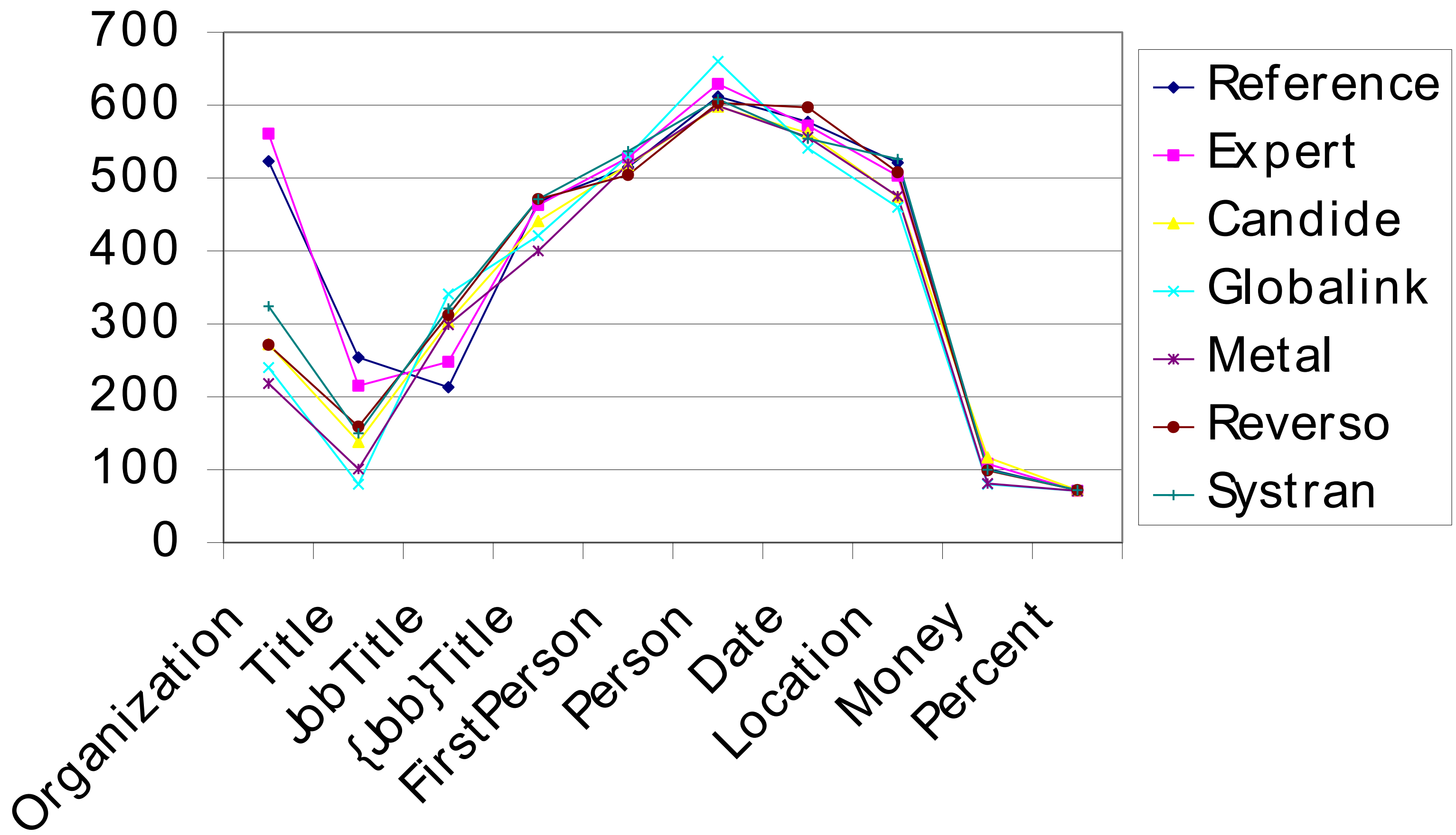
Performance methods

- A pragmatic approach to MT: similar to performance-based human evaluation
  - *"…can someone using the translation carry out the instructions as well as someone using the original?"* (Hutchins & Somers, 1992: 163)
- Different from human performance evaluation
  - 1. Tasks are carried out by an automated system
  - 2. Parameter(s) of the output are automatically computed

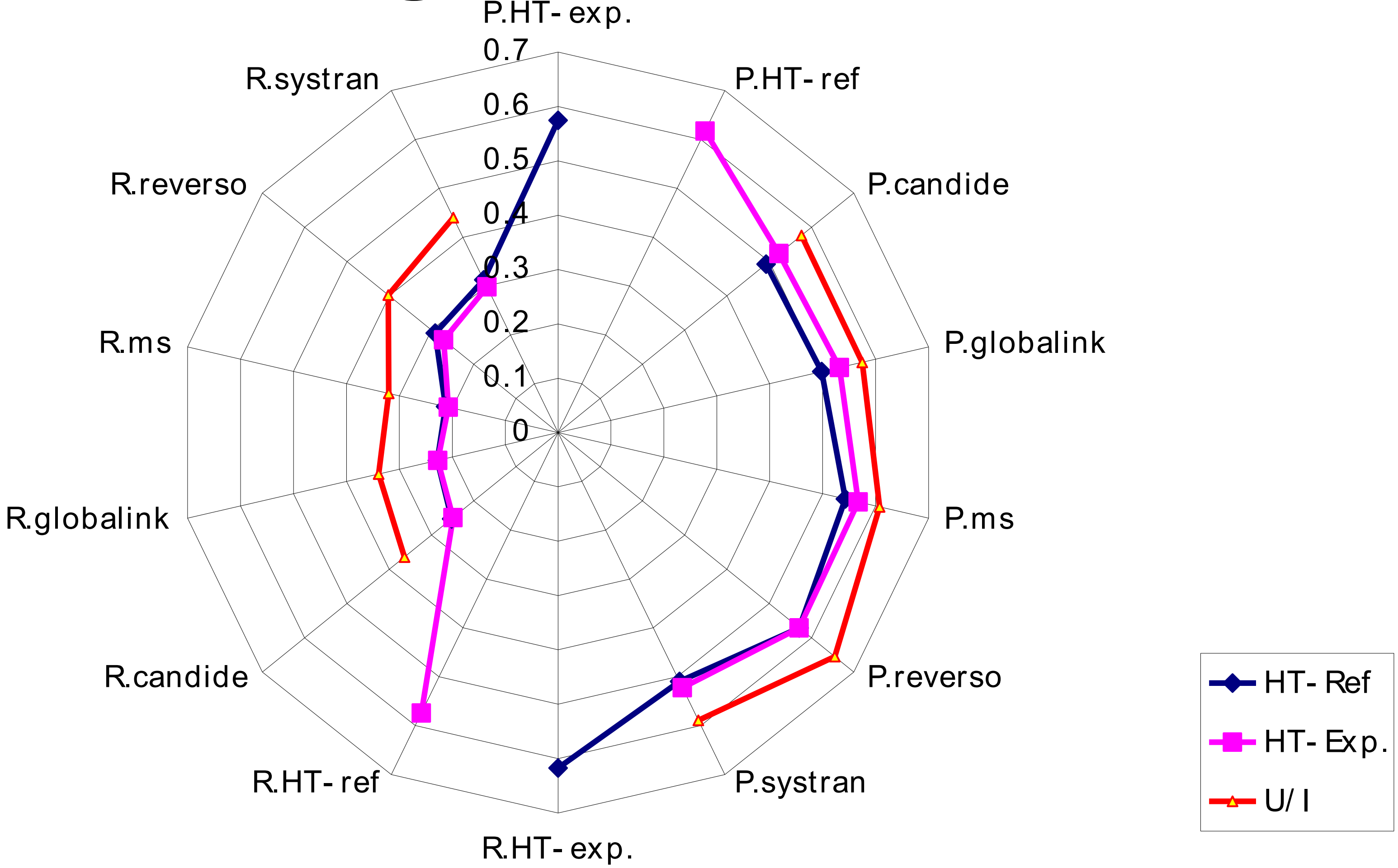
Performance-based methods: an example

- Open-source NER system for English (ANNIE) www.gate.ac.uk
  - the number of extracted Organisation Names gives an indication of Adequacy
  - ORI: … *le chef de la <u>diplomatie égyptienne</u>*
  - HT: *the* <Title>Chief</Title> *of the* <Organization>*<u>Egyptian Diplomatic Corps</u>* </Organization>
  - MT-Systran: *the* <JobTitle> *chief* </JobTitle> *of the <u>Egyptian diplomacy</u>*

NE recognition on MT output

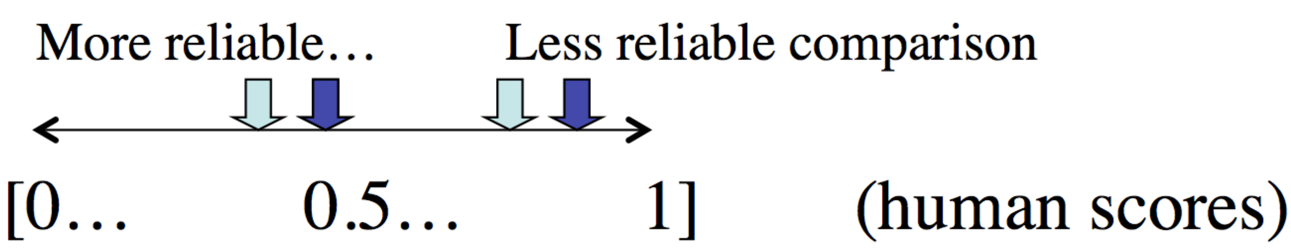


Precision (P) and Recall (R): Organisation names

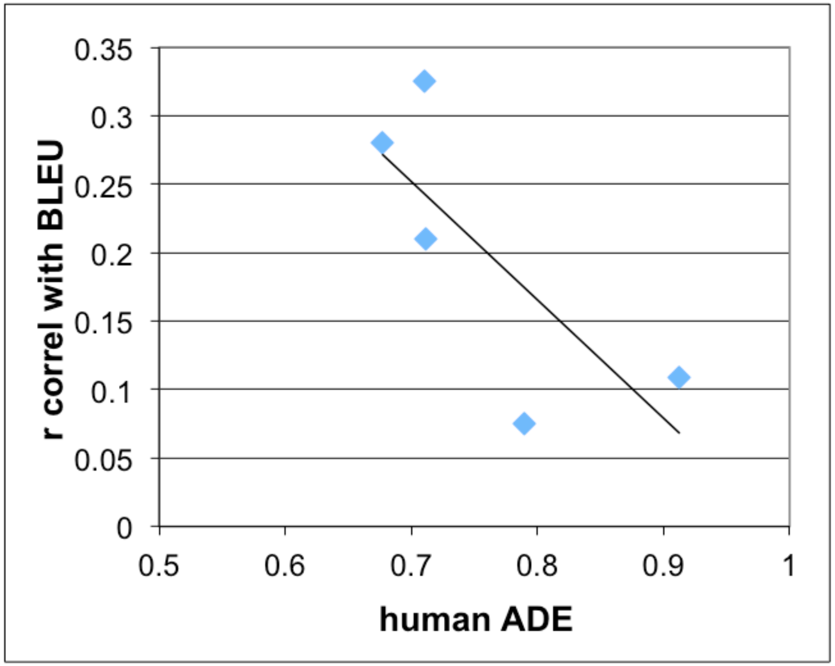


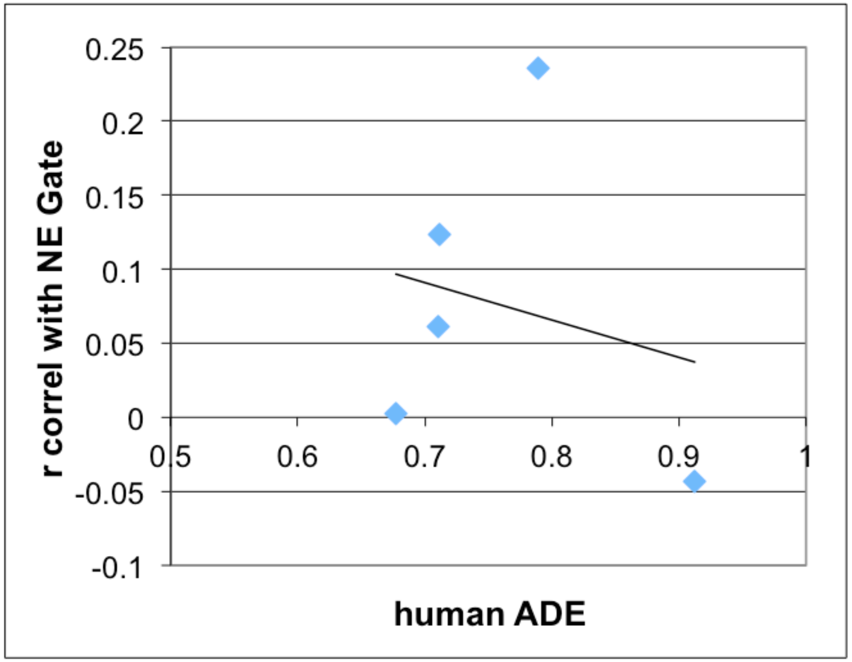
Varying sensitivity as a possible limit of automated metrics

- Ideally automated metrics should have homogeneous sensitivity across the entire human quality scale
- If sensitivity declines at a certain area on the scale, automated scores become less meaningful / reliable
  - For comparing easy / difficult segments generated by the same MT system
  - For distinguishing between systems at that area, e.g.:



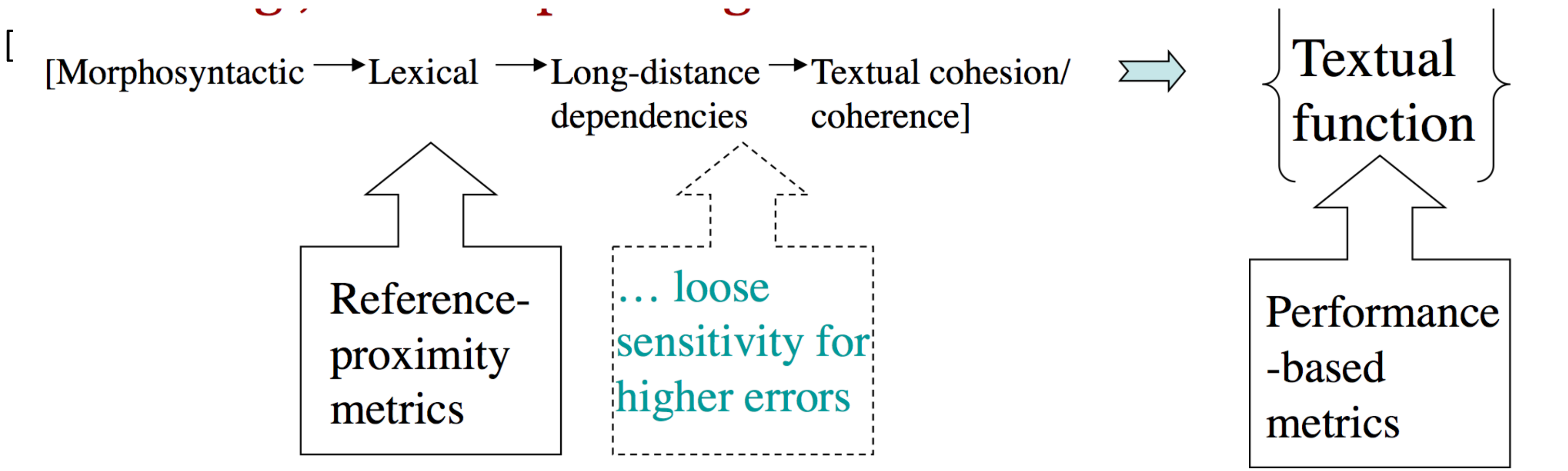
Correlation auto vs. human scores at different Adequacy ranges





Interpreting sensitivity

- Reference proximity metrics use *structural* model
  - Non-sensitive for errors on higher level (better MT)
  - Optimal correlation for certain error types
- Performance-based metrics use *functional* model
  - Potentially can capture degradation at any level
  - E.g., better capture legitimate variation



MT Architectures

- Rule-based
- Data-driven
  - Statistical & Neural
- Hybrid

→ typical errors and implications for the collaborative translation workflow

Comparison (Van Brussel et al., 2018)

	RBMT	PBMT	NMT
Accuracy	1309	741	472
Fluency	1831	1531	719
Total	**3140**	**2272**	**1191**

Table 1: Total number of errors

	RBMT	PBMT	NMT
Correct sentences	81	130	217
In %	12%	20%	33%

Table 2 : Correct sentences in MT output

Comparison (Van Brussel et al., 2018)

Accuracy errors	RBMT	PBMT	NMT
Mistranslation	972	483	330
DNT	116	14	22
Untranslated	65	69	44
Addition	61	39	2
Omission	43	115	62
Mechanical	52	21	12
Total	**1309**	**741**	**472**

Table 3: Overview of the number of accuracy errors

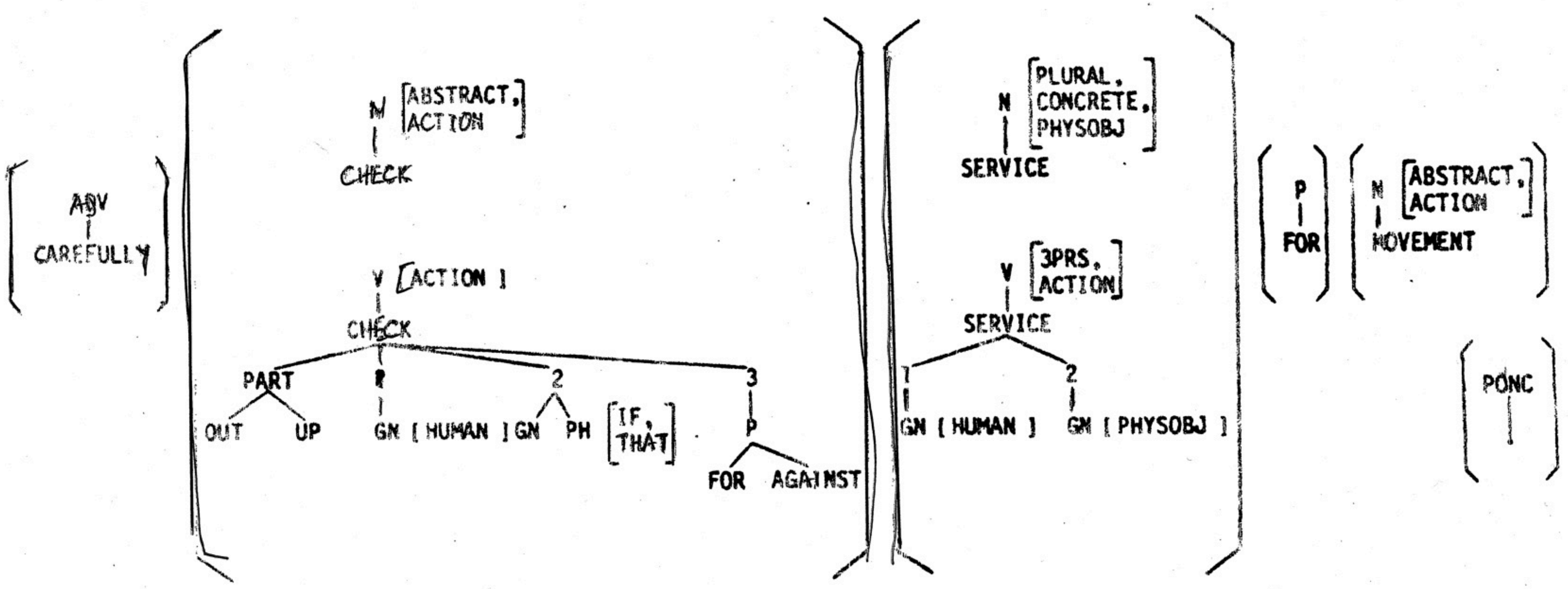
Comparison (Van Brussel et al., 2018)

	RBMT	PBMT	NMT
Total omissions	**43**	**115**	**62**
Content words	6	80	53
Function words	37	35	9
% Content words	0,14 %	69,96 %	85,48 %
Visibility	40	89	19
Invisibility	3	26	43
% Invisibility	7 %	23 %	69 %

Table 6: Subdivision of omission errors based on their type and visibility

Rule-based MT: TAUM

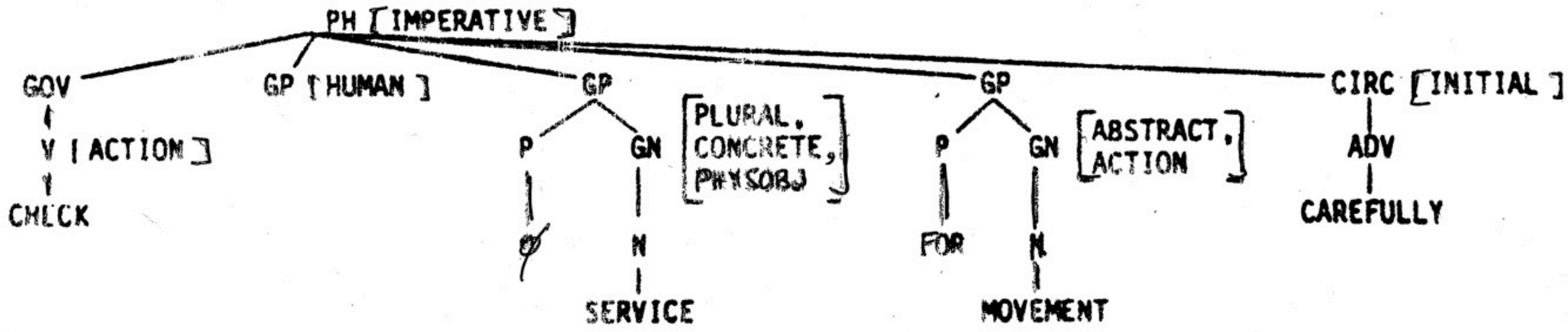
Dictionary lookup



En: ‘carefully check services for movement’

Rule-based MT: TAUM

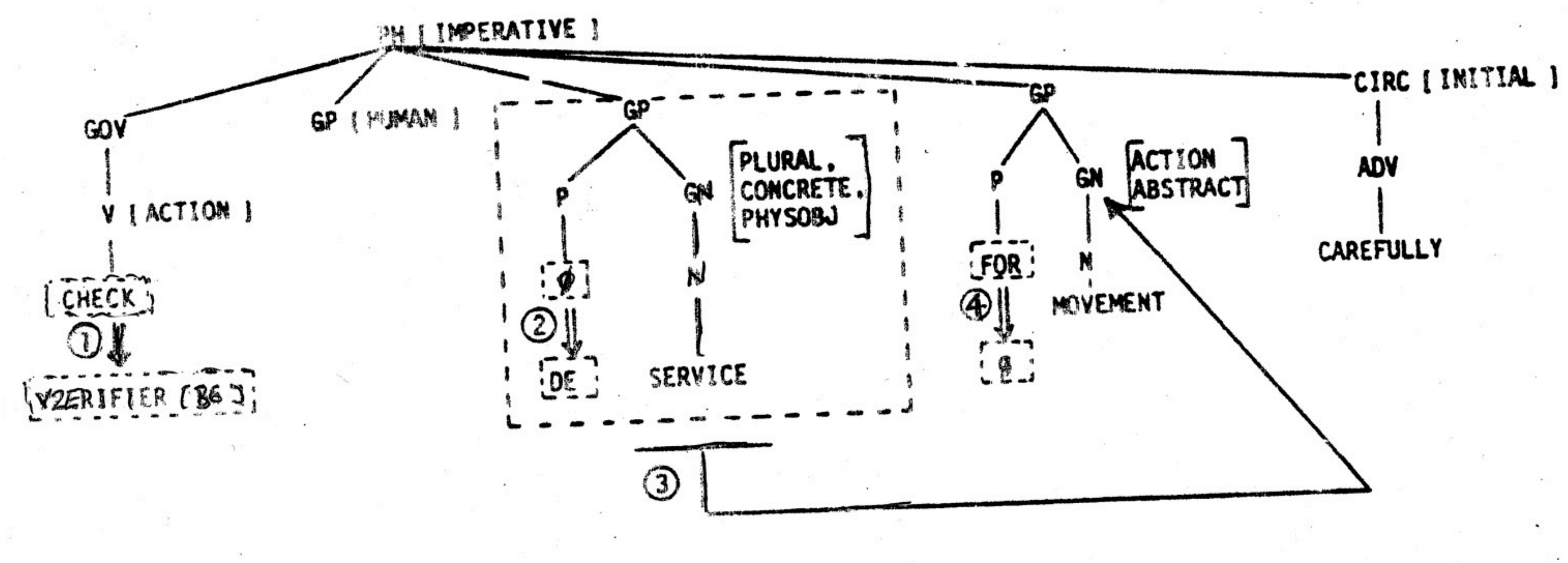
4.5 Syntactico-Semantic Analysis



En: ‘carefully check services for movement’

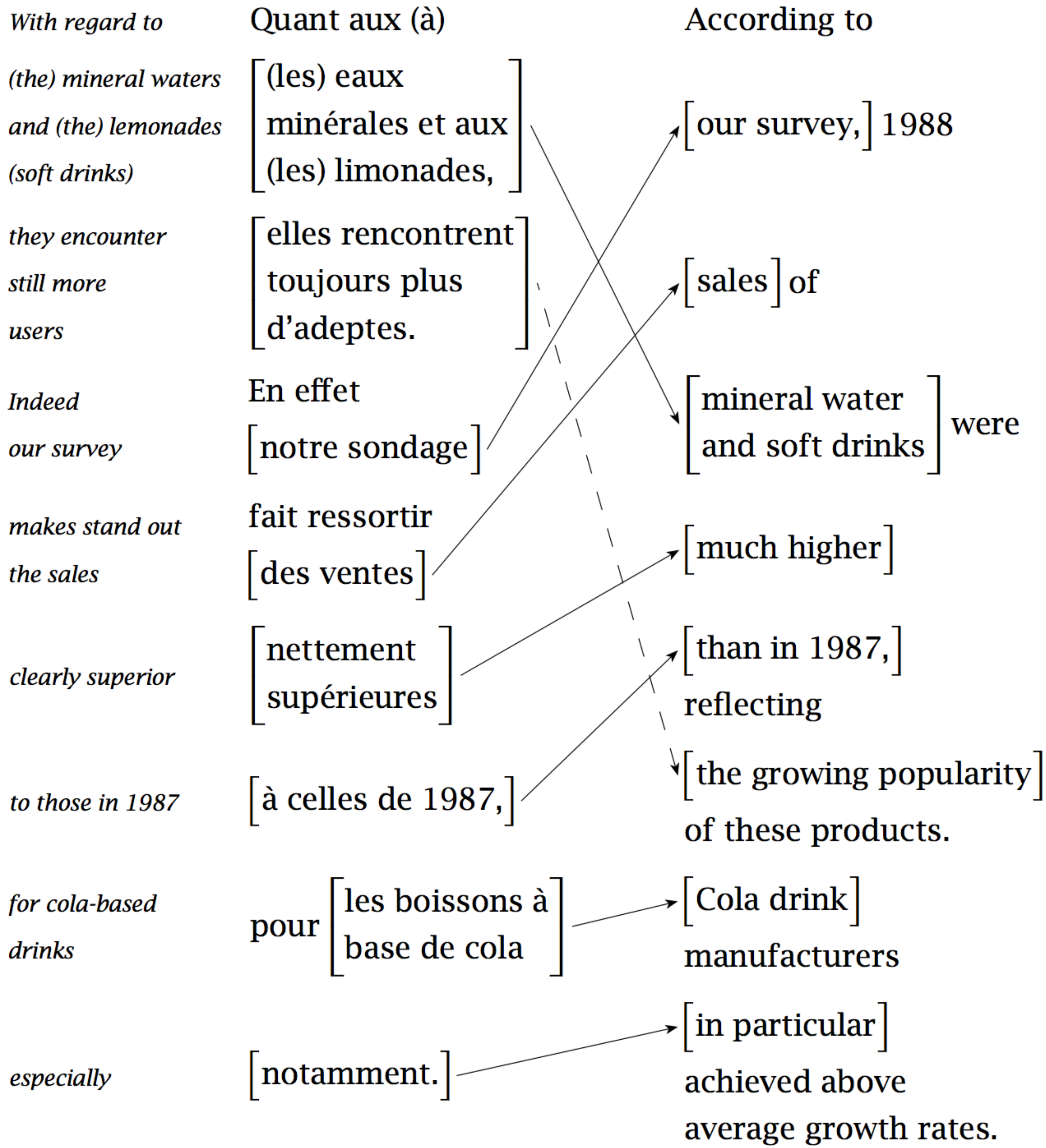
Rule-based MT: TAUM

lexical and syntactic transfer



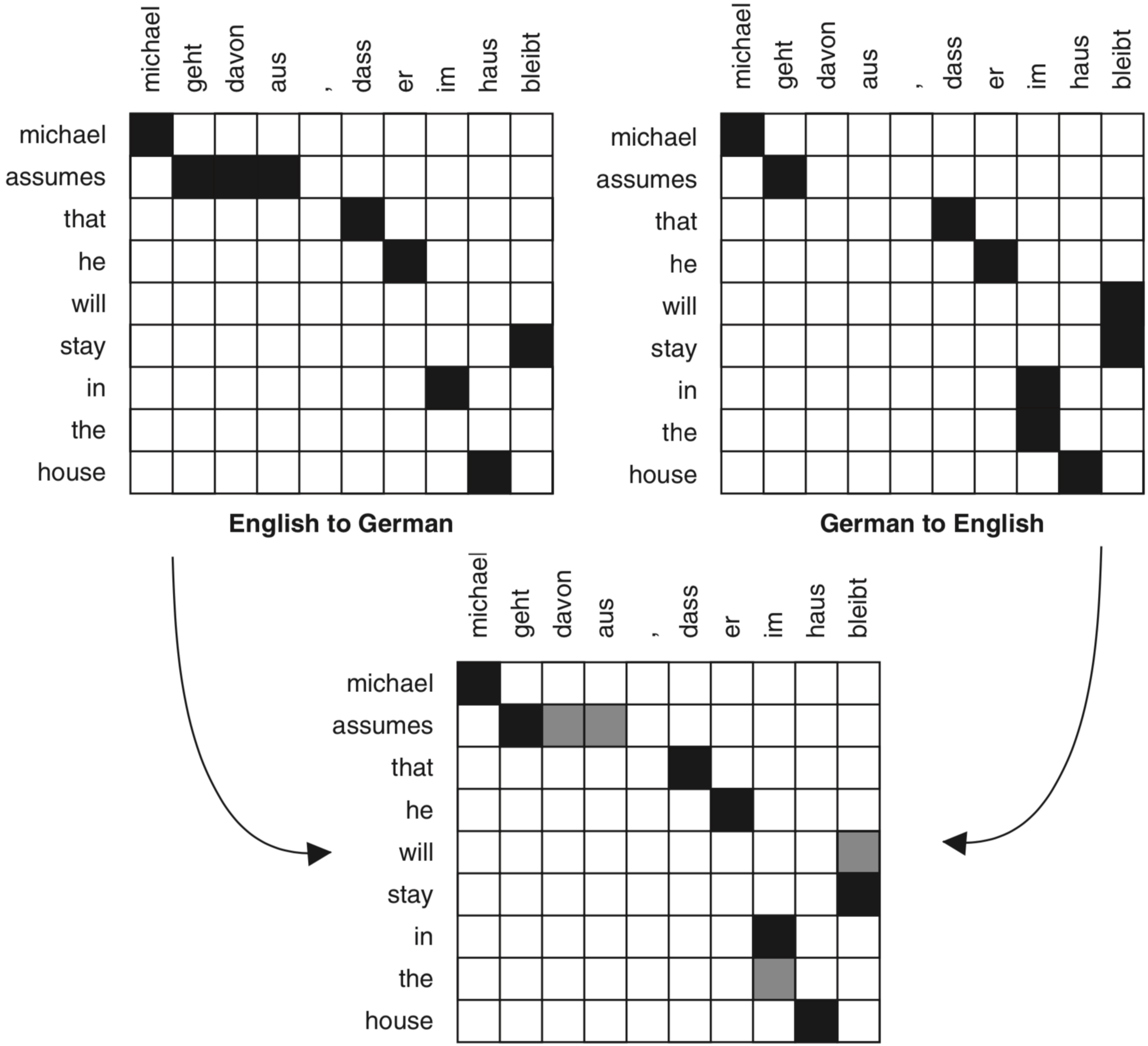
En: ‘carefully check services for movement’ →

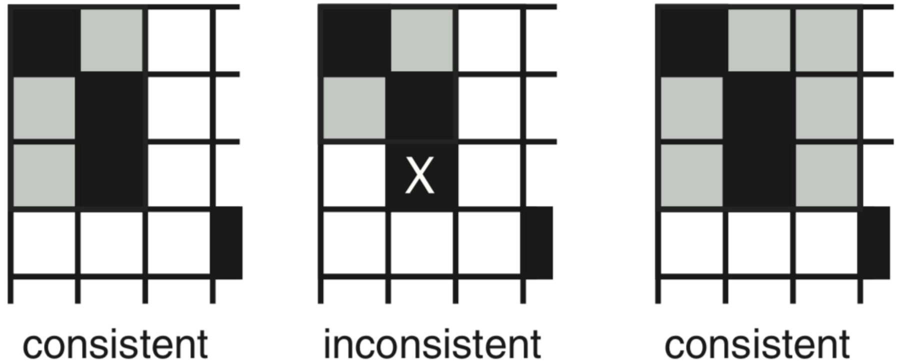
Fr: ‘vérifiez soigneusement le mouvement des servitudes’

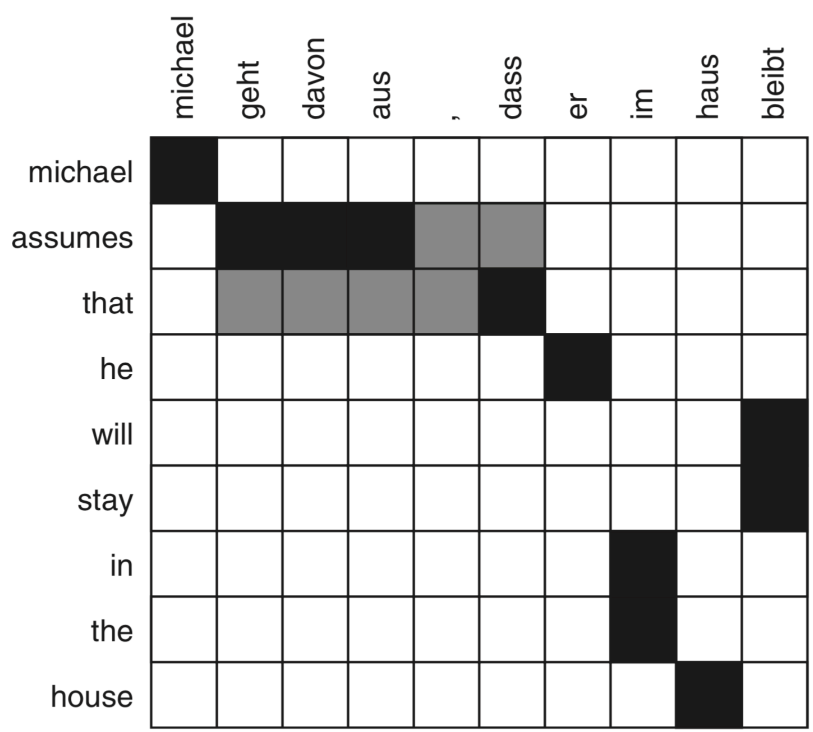


Statistical MT (Moses…)

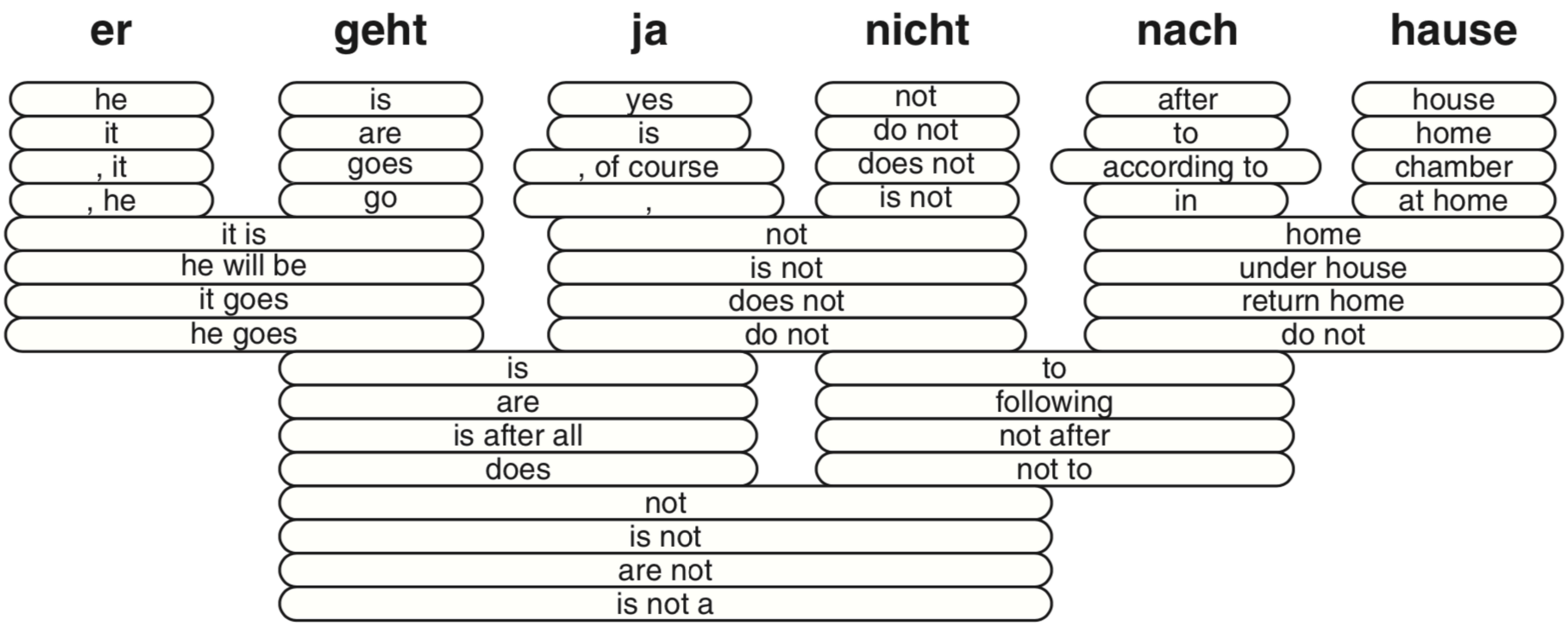
- Off-line Stage: Alignment
- See: Koehn, P. 2010 Statistical machine translation. Cambridge University Press


michael geht davon aus , dass er im haus bleibt
michael assumes that he will stay in the house
English to German
German to English


X
consistent
inconsistent
consistent


michael
geht
davon
aus
,
dass
er
im
haus
bleibt
michael
assumes
that
he
will
stay
in
the
house

Statistical PBMT: Lookup in the phrase table



Statistical PBMT: decoding at runtime