Neural Machine Translation and Beyond

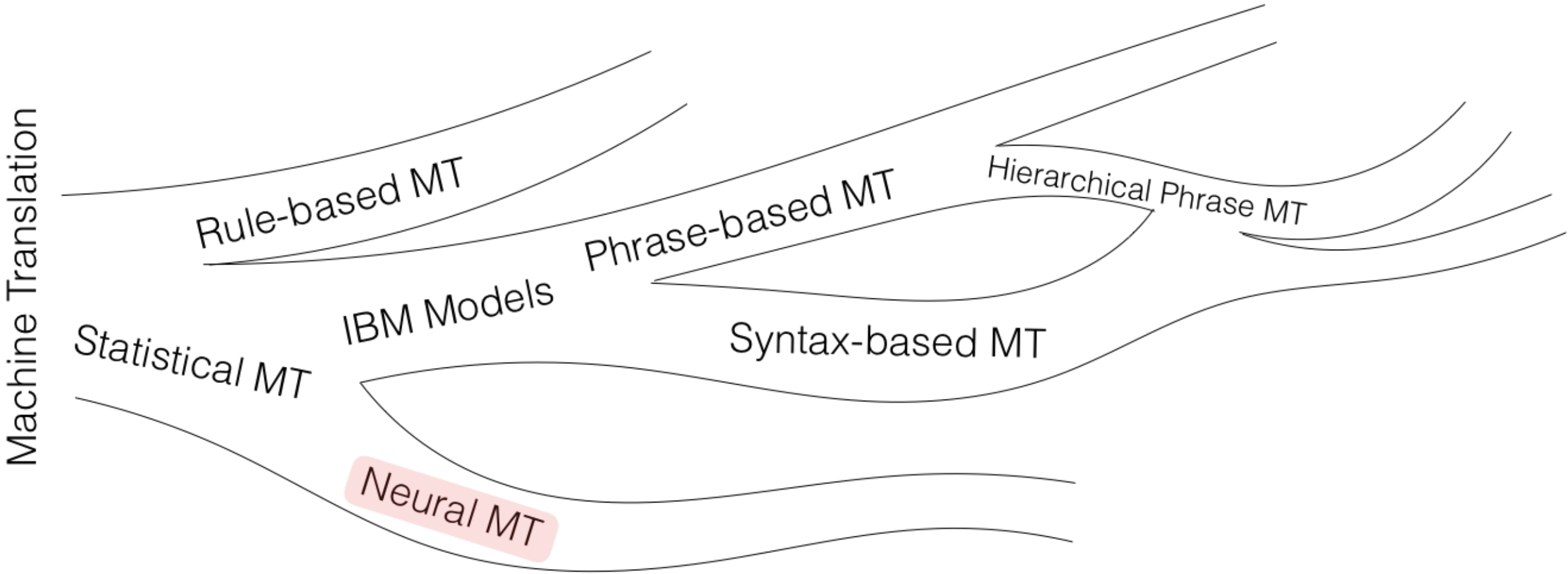
Kyunghyun Cho

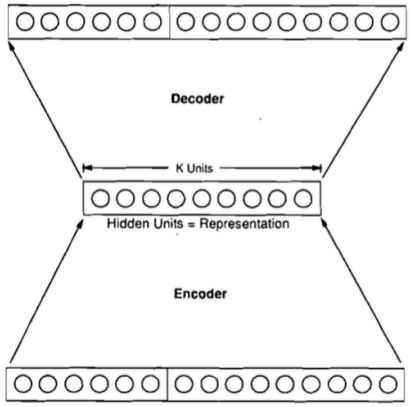
Courant Institute & Center for Data Science, New York University

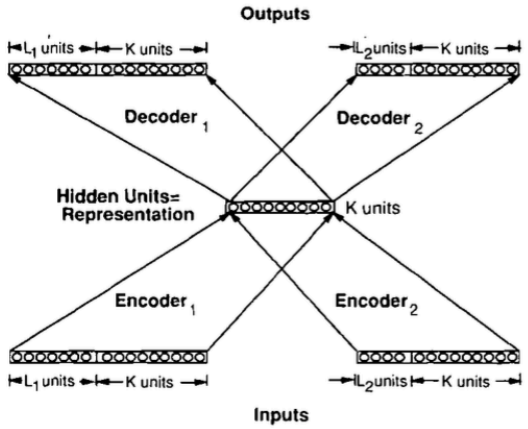
Facebook AI Research

A Brief History

For a comprehensive history of machine translation, go find **Andy Way**


Machine Translation
Rule-based MT
Statistical MT
IBM Models
Phrase-based MT
Hierarchical Phrase MT
Syntax-based MT
Neural MT



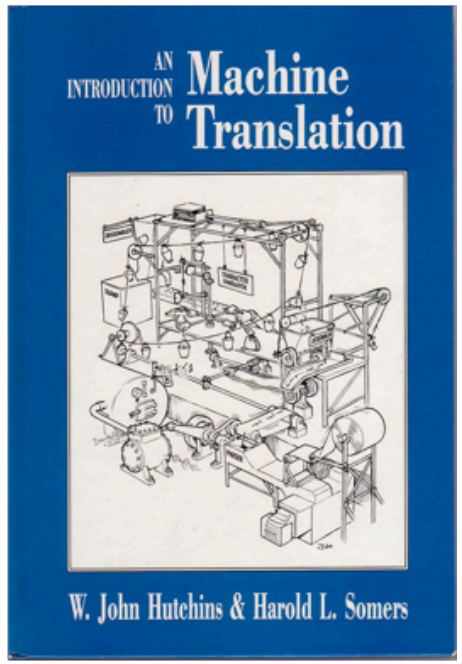


- [Allen 1987 IEEE 1st ICNN]
- 3310 En-Es pairs constructed on 31 En, 40 Es words, max 10/11 word sentence; 33 used as test set
- Binary encoding of words – 50 inputs, 66 outputs; 1 or 3 hidden 150-unit layers. Ave WER: 1.3 words

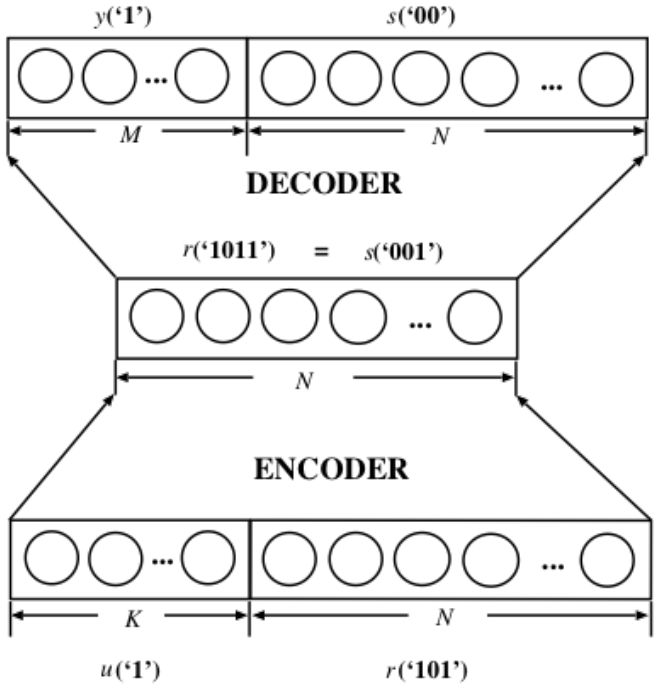
- [Chrisman 1992 *Connection Science*]
- Dual-ported RAAM architecture [Pollack 1990 *Artificial Intelligence*] applied to corpus of 216 parallel pairs of simple En-Es sentences:
- Split 50/50 as train/test, 75% of sentences correctly translated!

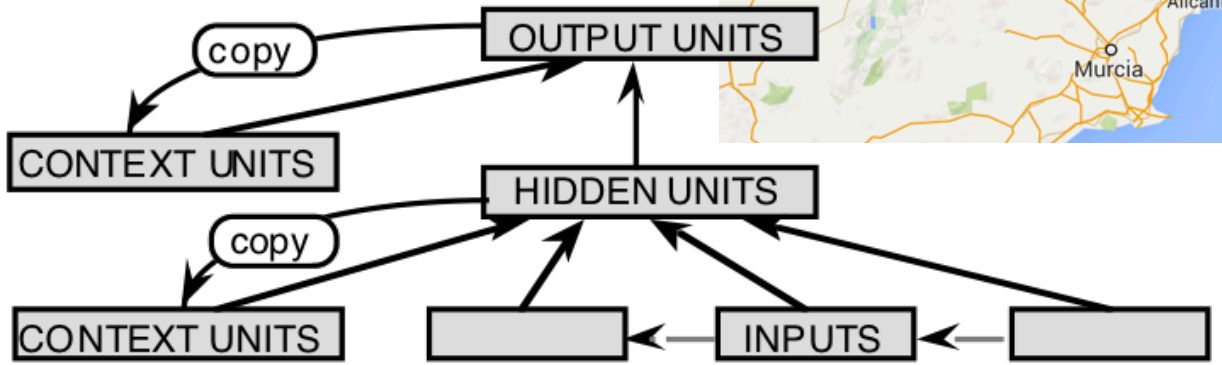
The relevance of the ***connectionist model to natural language processing is clear enough****. The traditional stratificational approach to parsing and generation (morphology, syntax, semantics) .. is not seriously accepted .. as* ***a psychologically real model of how humans understand and communicate****.*

Hutchins and Somers (1992)



Brief resurrection in 1997: Spain





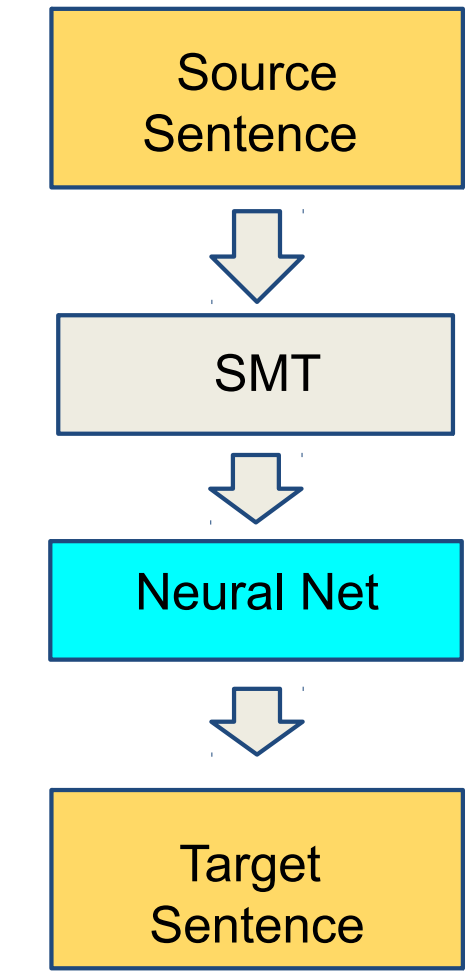
*"We propose .. **Recursive Hetero-Associative Memory** which .. may be applied **to learn general translations from examples** in which different sentences may have the same translation."*

– Forcada & Ñeco, 1997

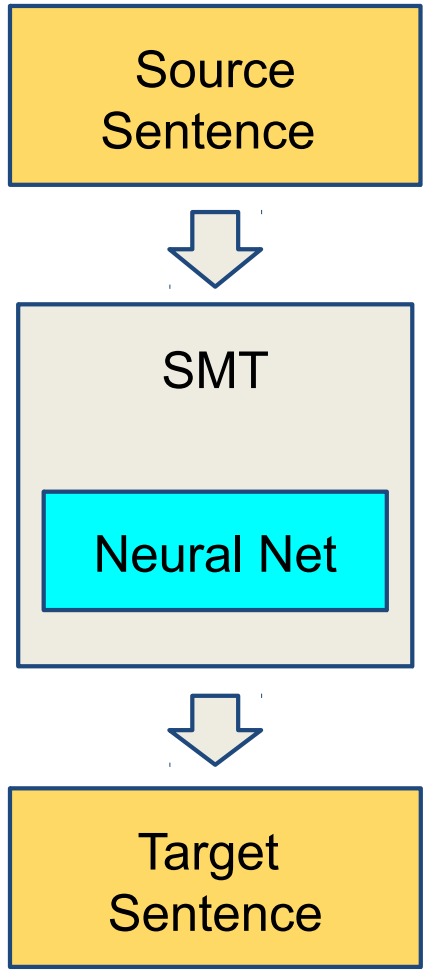
*"Based on these encouraging performances, future work dealing with more complex limited-domain translations seems to be feasible. **However, the size of the neural nets required for such applications (and consequently, the learning time) can be prohibitive**"*

- Castano & Casacuberta, 1997

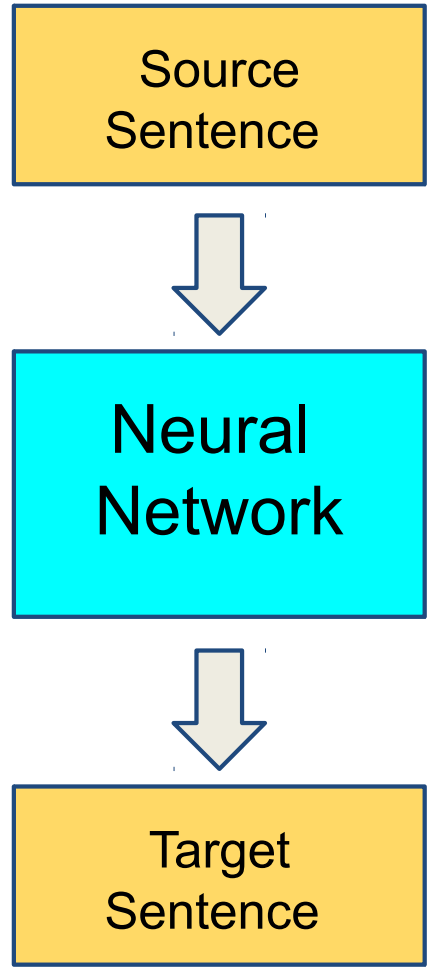
Modern neural machine translation



(Schwenk et al. 2006)

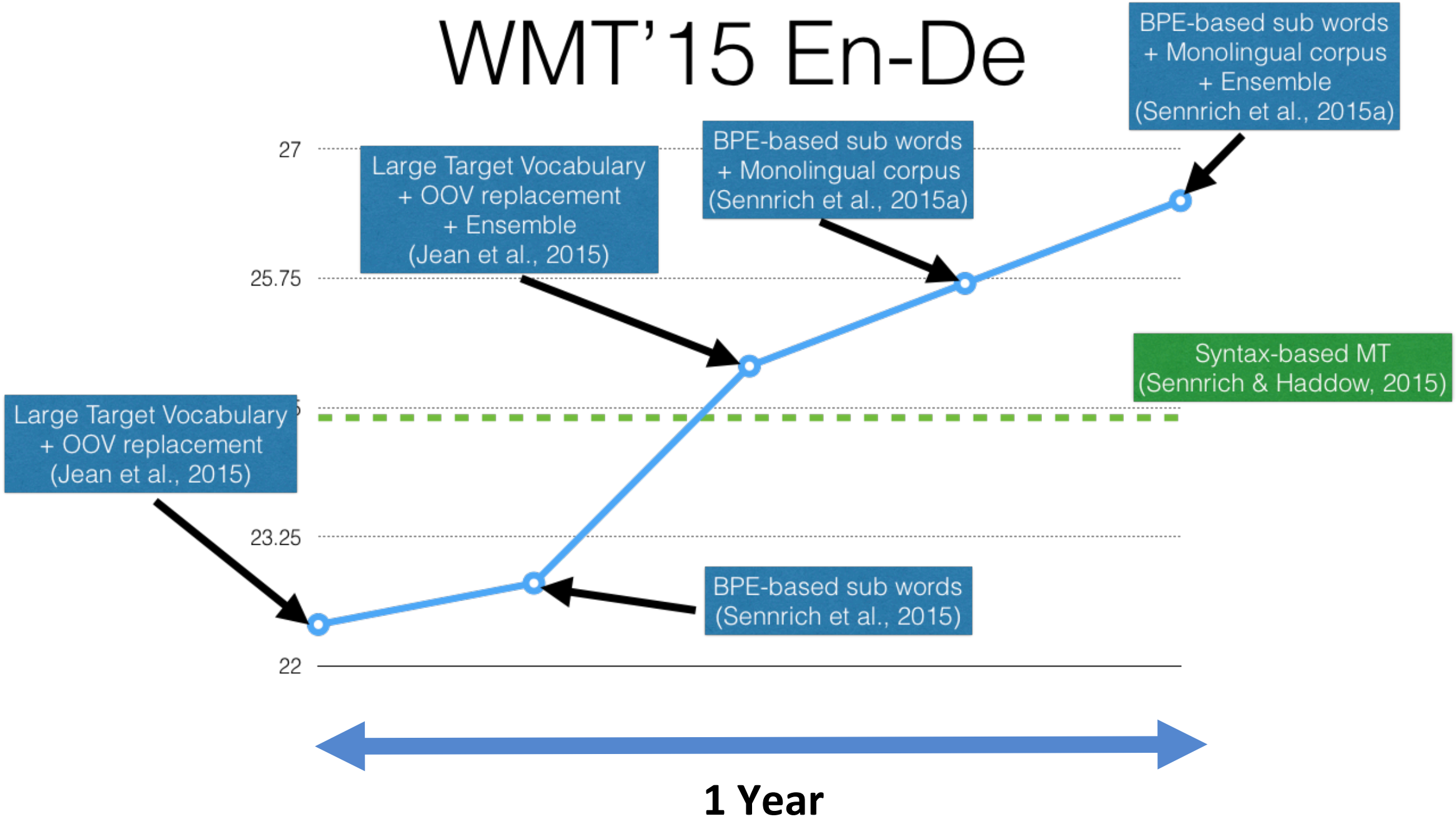


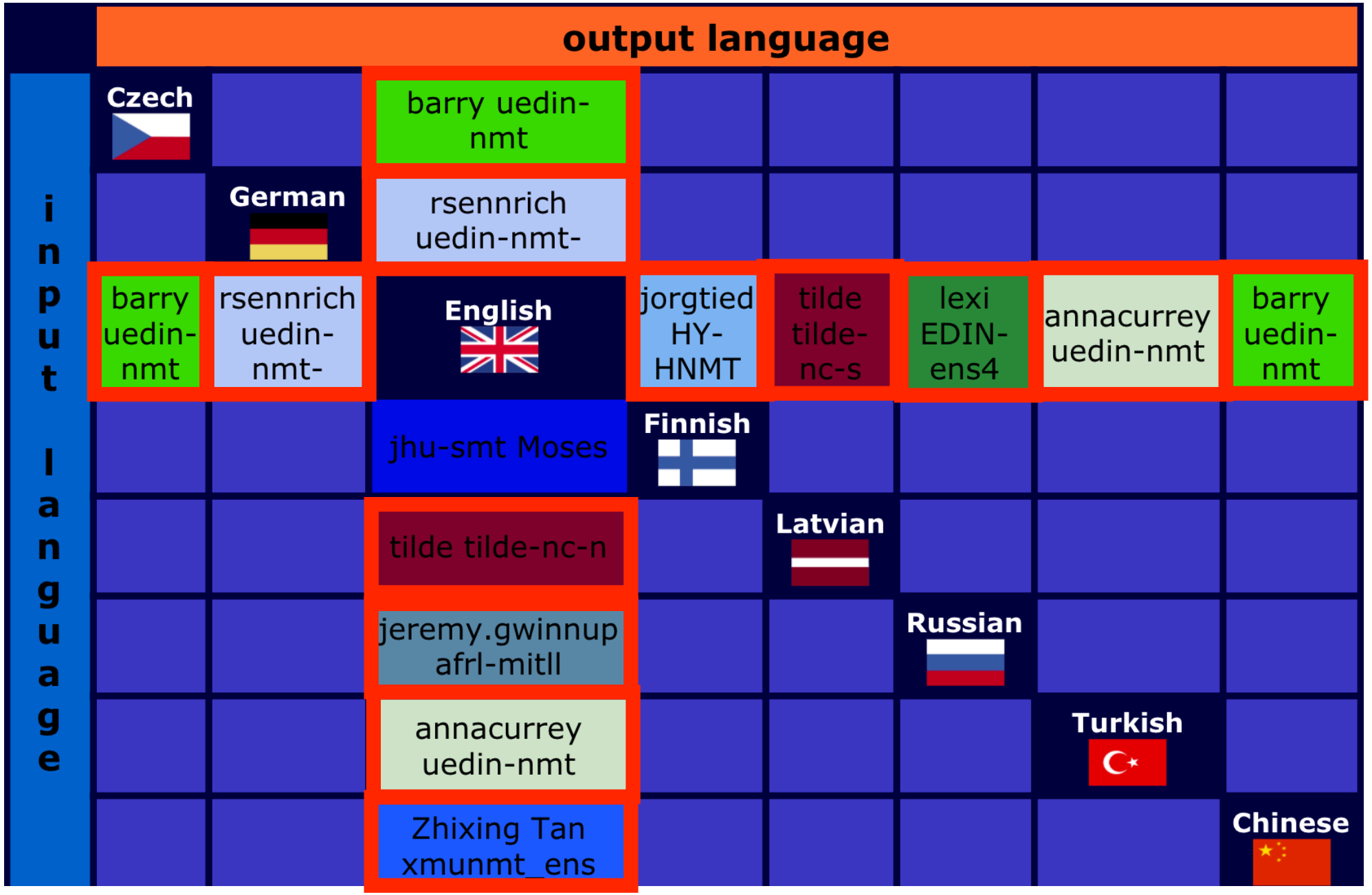
(Devlin et al. 2014)



Neural MT

WMT'15 En-De


BPE-based sub words
+ Monolingual corpus
+ Ensemble
(Sennrich et al., 2015a)
27
Large Target Vocabulary
+ OOV replacement
+ Ensemble
(Jean et al., 2015)
BPE-based sub words
+ Monolingual corpus
(Sennrich et al., 2015a)
25.75
Syntax-based MT
(Sennrich & Haddow, 2015)
Large Target Vocabulary
+ OOV replacement
(Jean et al., 2015)
23.25
BPE-based sub words
(Sennrich et al., 2015)
22
1 Year



WMT 2017: news translation task

A better single-pair translation system has never been “the” goal of neural MT

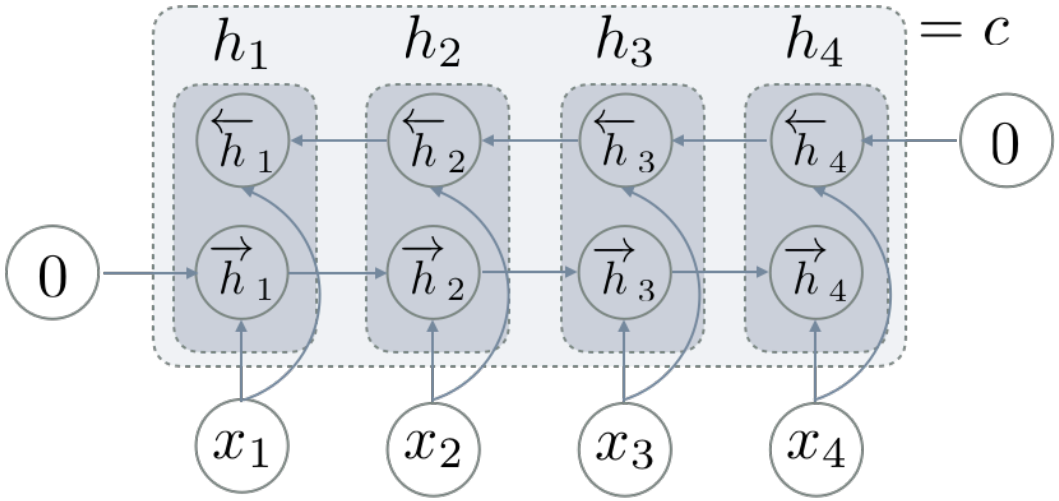
Multilingual translation via continuous representation

What if we can project sentences in multiple languages into a single vector space?

What does NMT do?

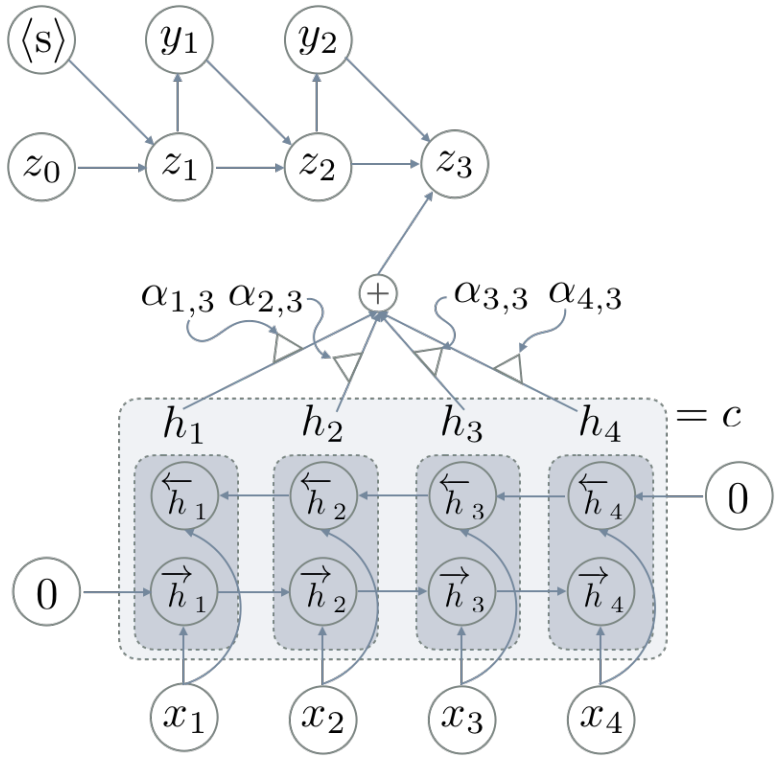
Encoder

- Project a source sentence into a set of continuous vectors



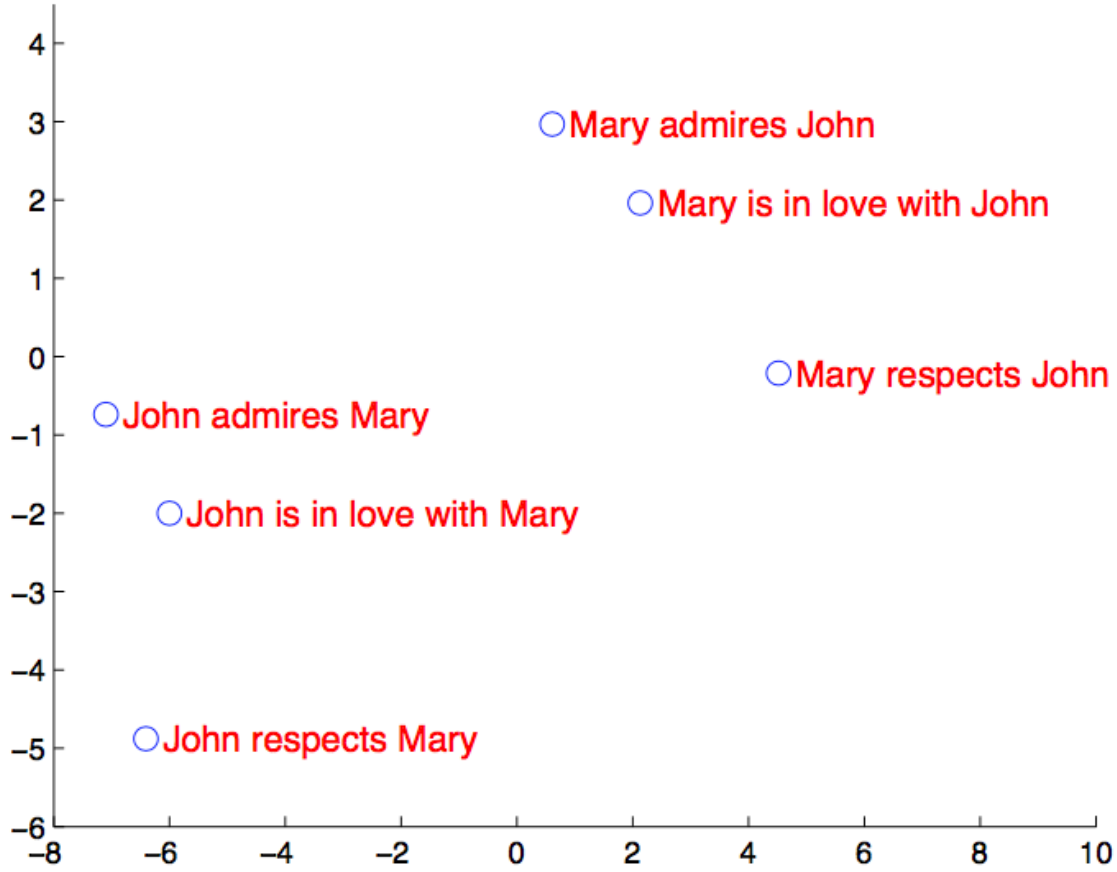
Decoder+Attention

- Decode a target sentence from a set of "source" continuous vectors



What is this “continuous vector space”?

- Similar sentences are near each other in this vector space
- Multiple dimensions of similarity are encoded simultaneously



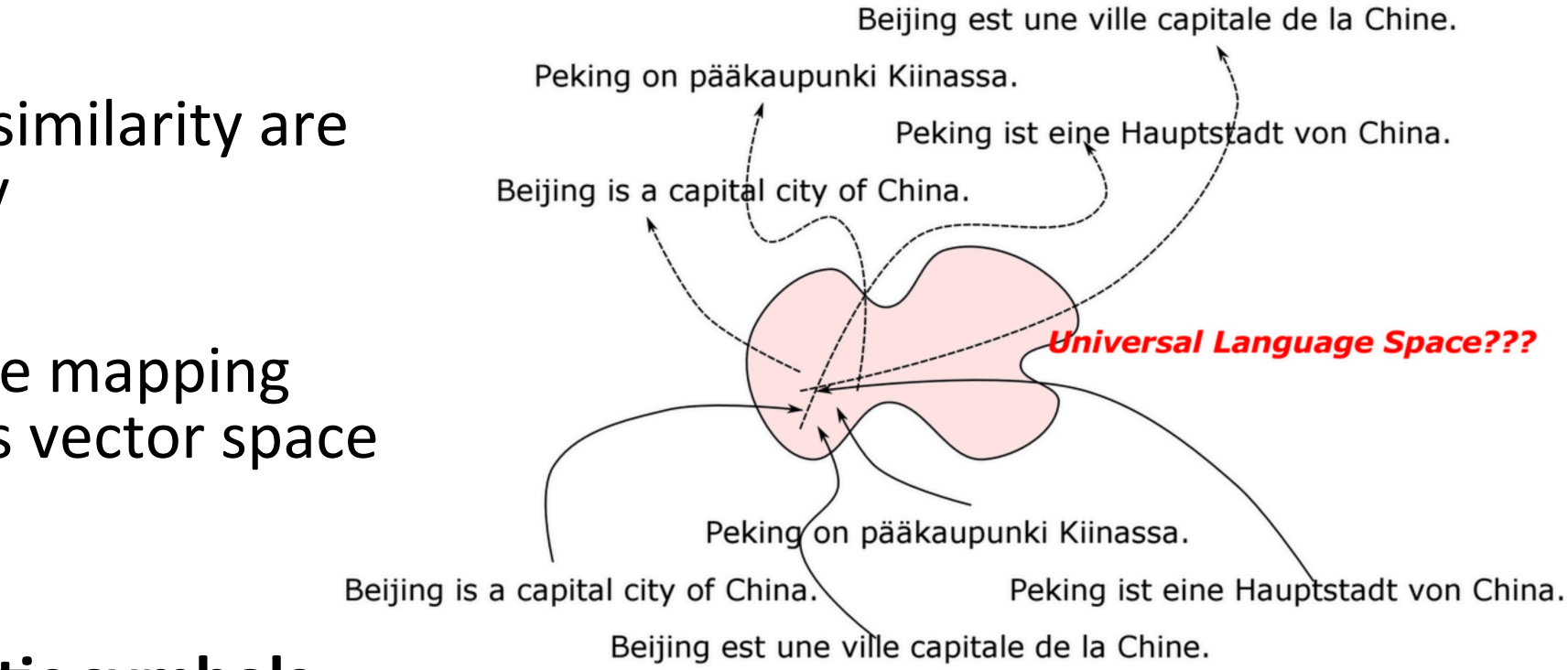
(Sutskever et al., 2014)

What is this “continuous vector space”?

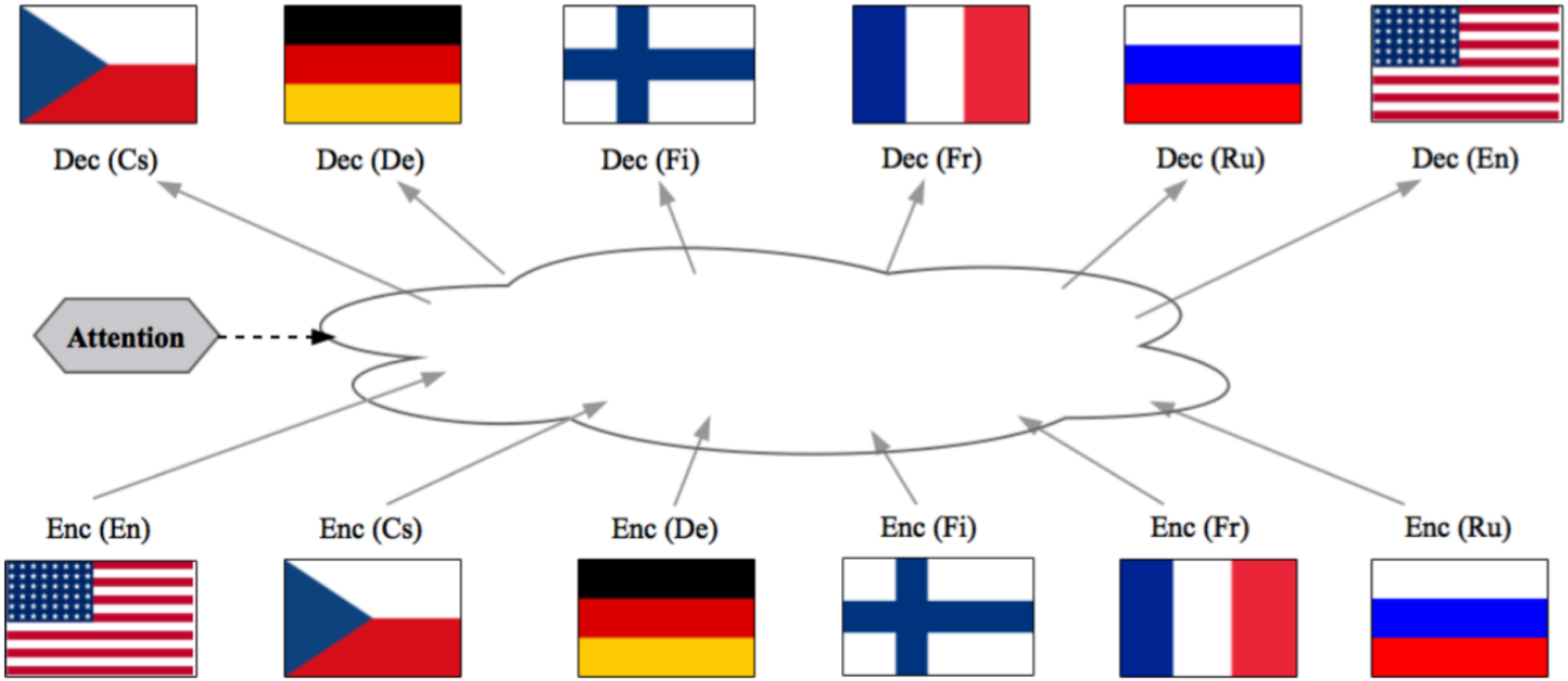
- Similar sentences are near each other in this vector space
- Multiple dimensions of similarity are encoded simultaneously

- (Trainable) near-bijective mapping between the continuous vector space and the sentence space

- **Stripped of hard linguistic symbols**



What is this “continuous vector space”?

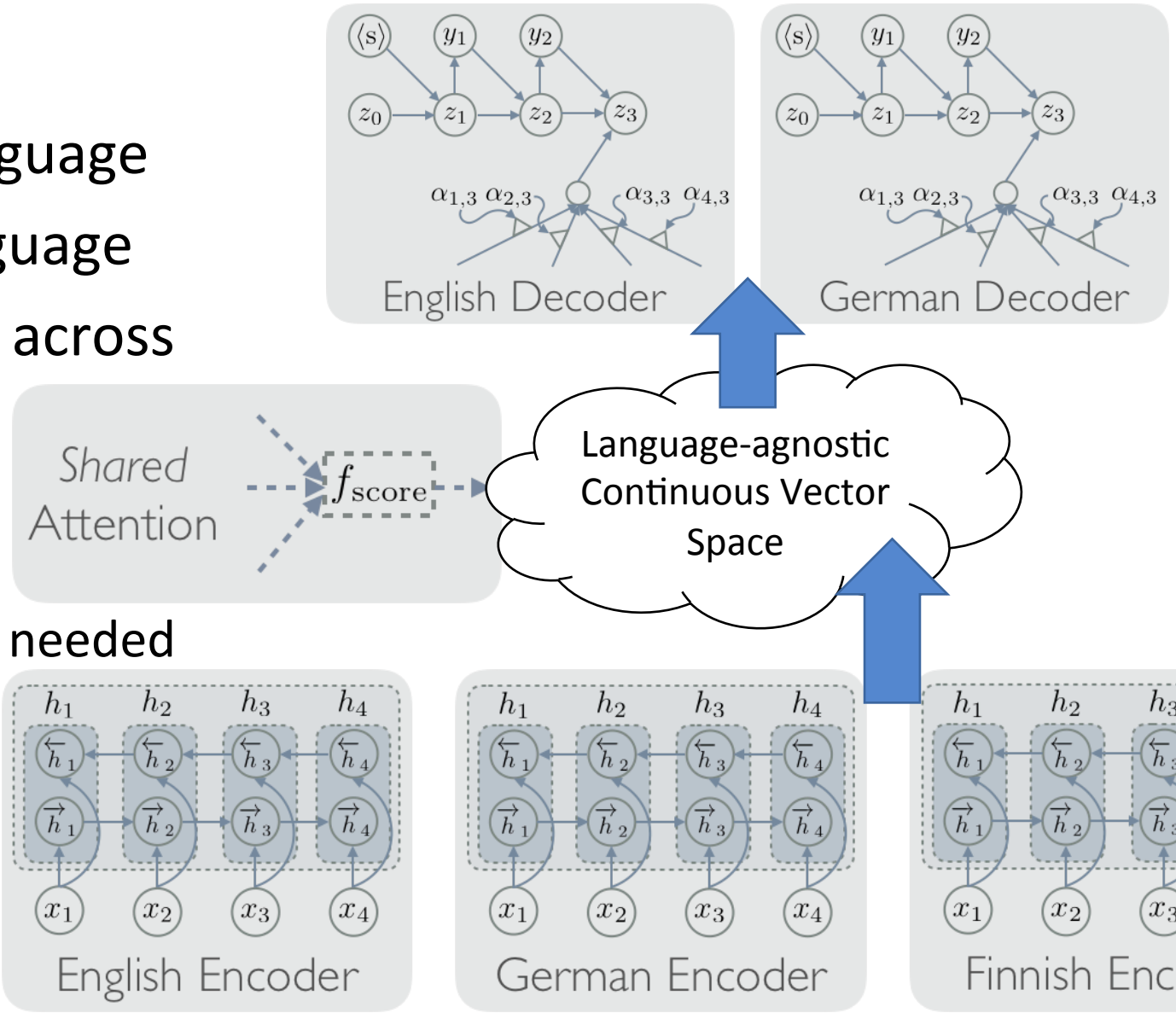


- Can this continuous vector space be shared across multiple languages?

(Firat et al., 2016; Luong et al., 2015; Dong et al., 2015)

Multi-way, multilingual machine translation (1)

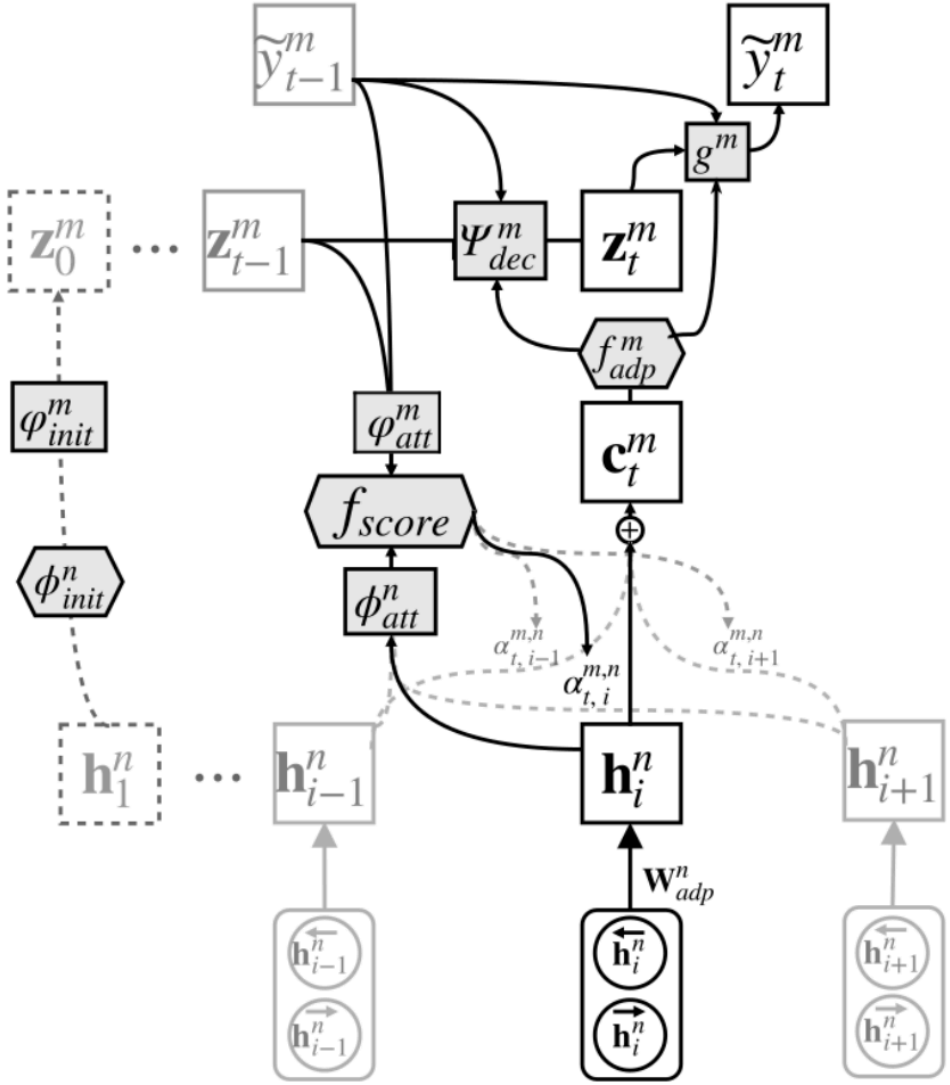
- One encoder per source language
- One decoder per target language
- Attention/alignment shared across all the language pairs
- Only bilingual parallel corpora necessary
  - No multi-way parallel corpus needed



(Firat et al., 2016)

Multi-way, multilingual machine translation (2)

- Neural nets are like lego
- Build one encoder per source
- Build one decoder per target
- Build one attention mechanism
- Given a sentence pair $(X^{\text{source}}, Y^{\text{target}})$
  - $H = \text{encoder}^{\text{source}}(X^{\text{source}})$
  - $Y = \text{decoder}^{\text{target}}(H, \text{attention})$

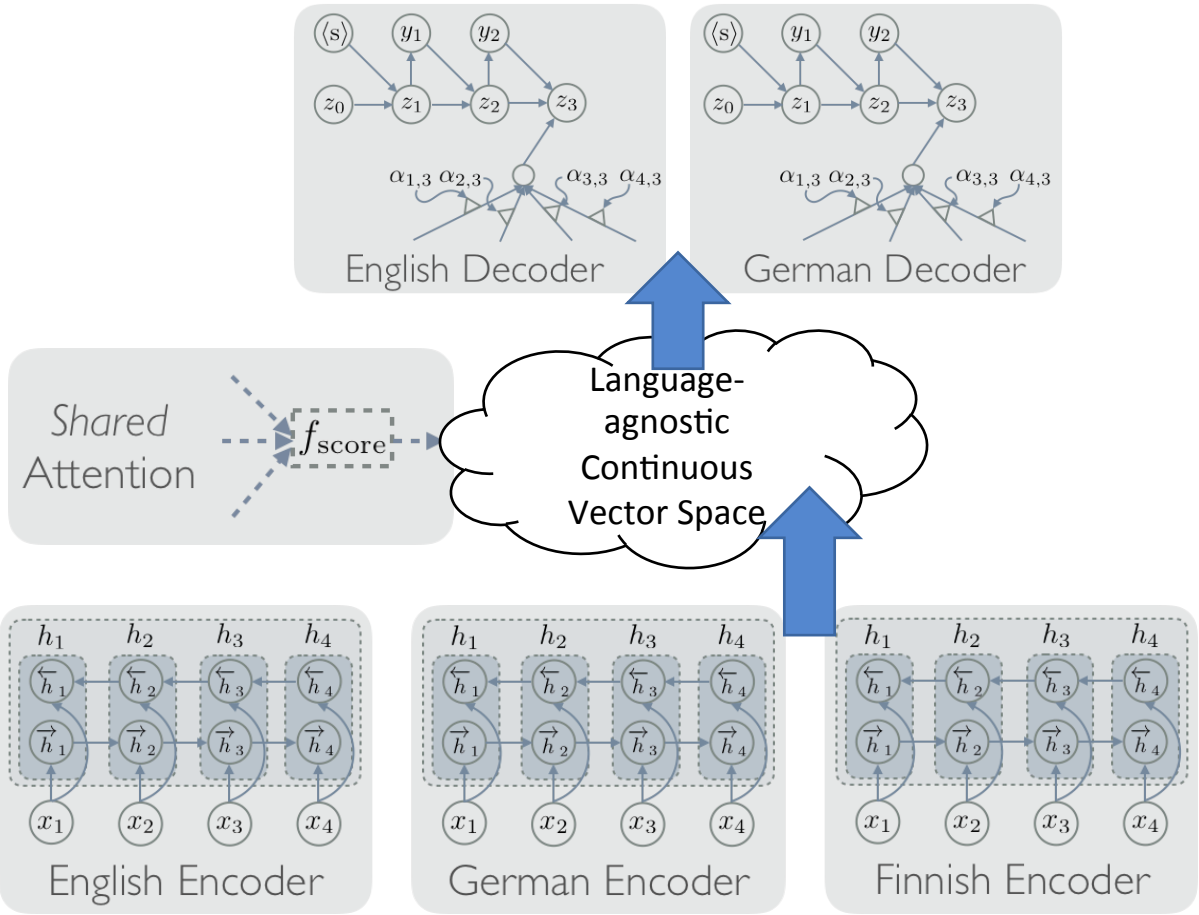


(Firat et al., 2016)

Multi-way, multilingual machine translation (3)

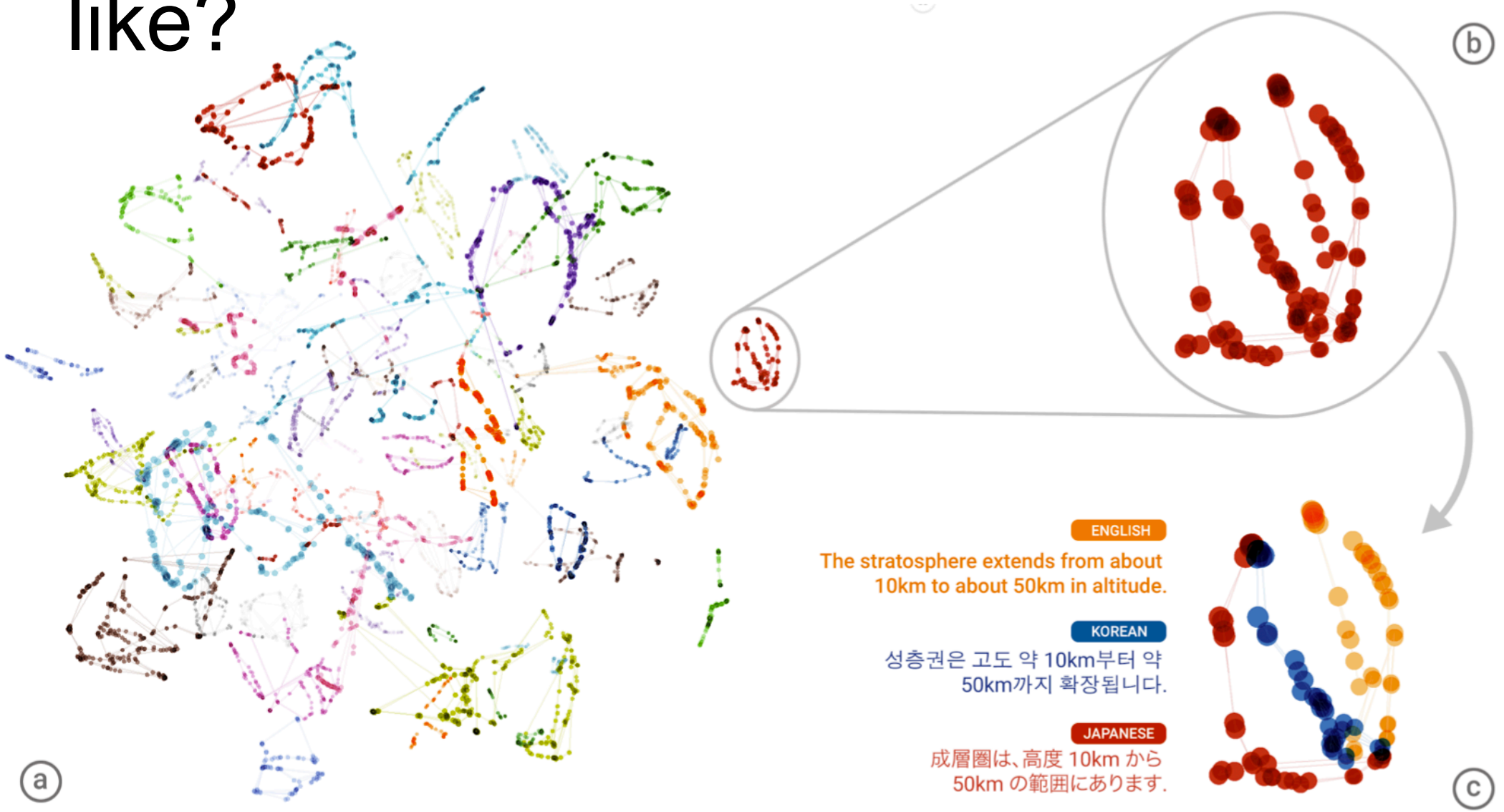
- Sentence-level positive language transfer
- Helps low-resource language pairs
- Why?
  1. Better structural constraint on the continuous vector space
  2. Regularization

- *Real-valued vector-based interlingua?*



(Firat et al., 2016)

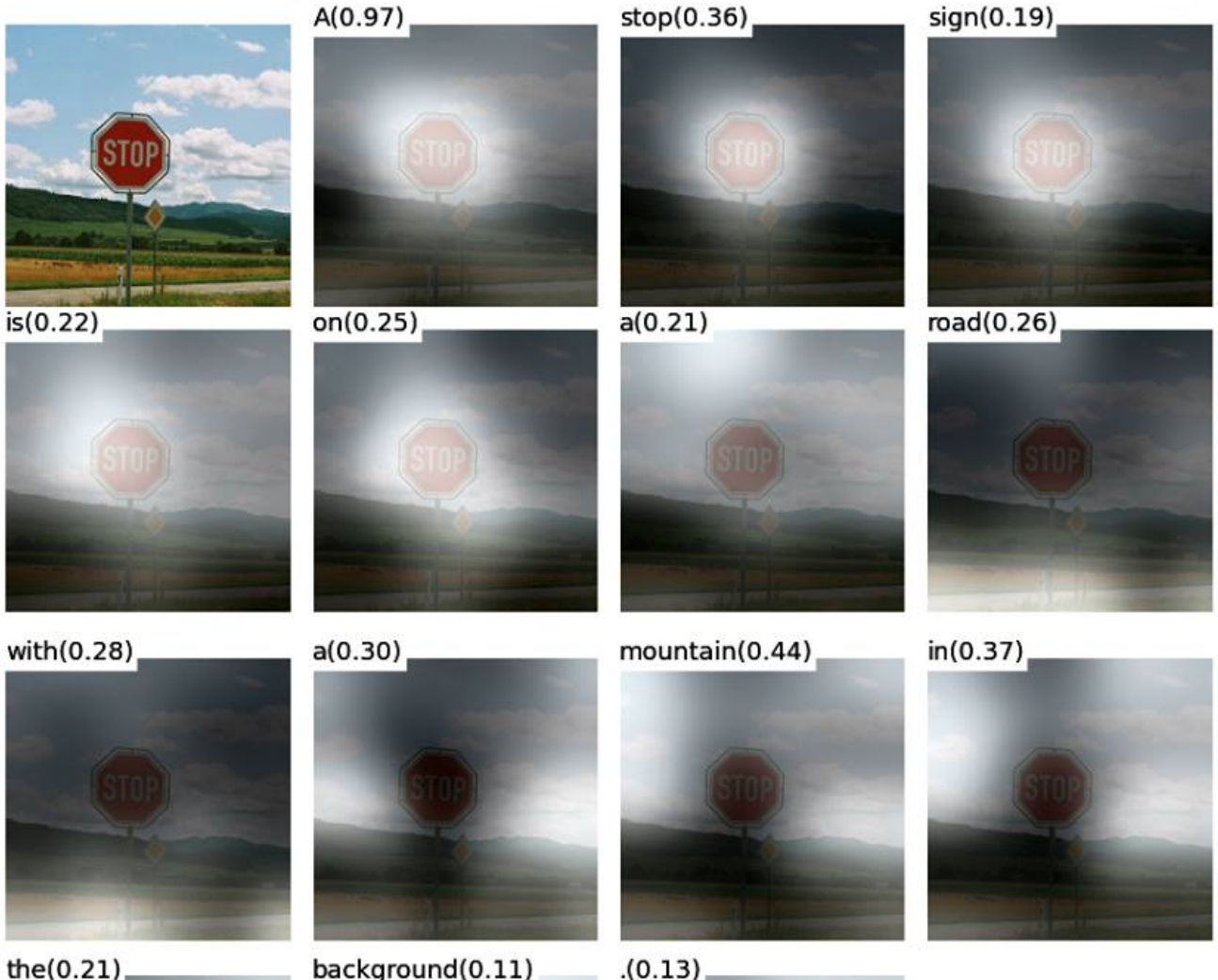
How does the continuous space look like?

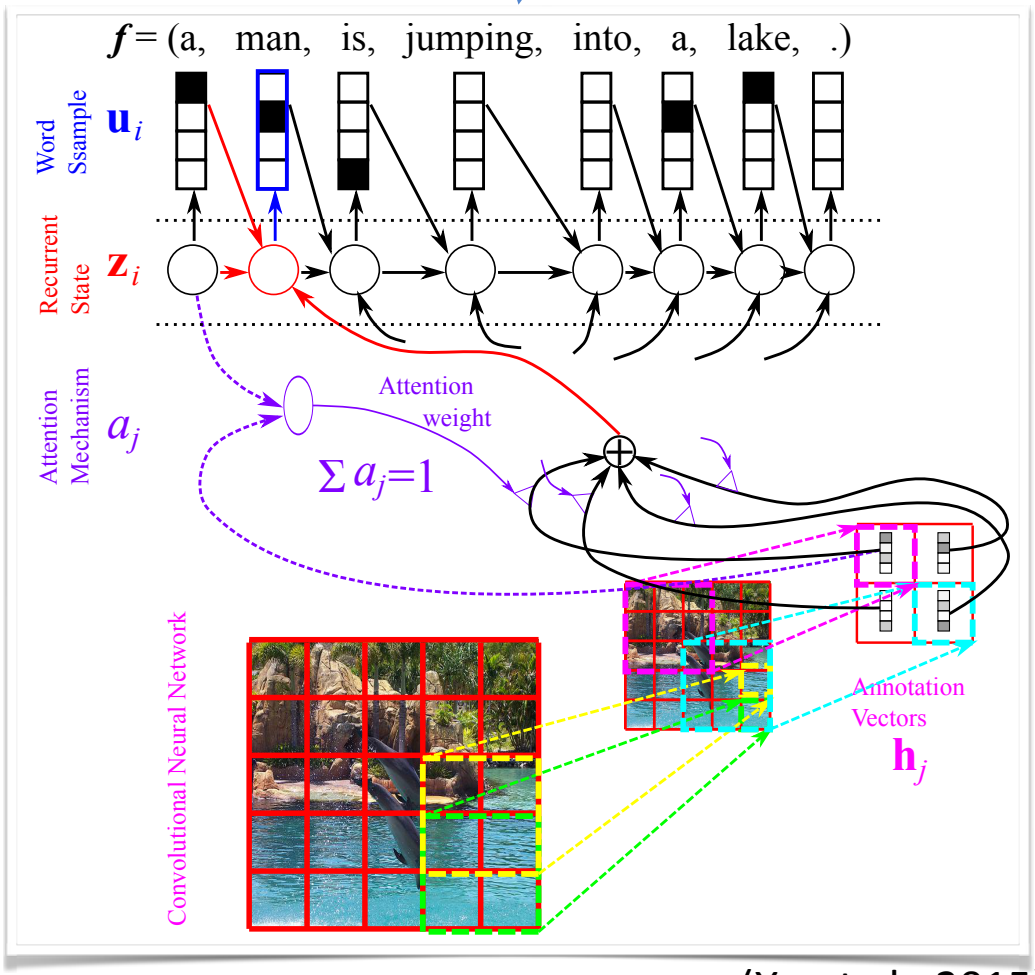


(Johnson et al., 2016)

Apparently not so cool anymore!

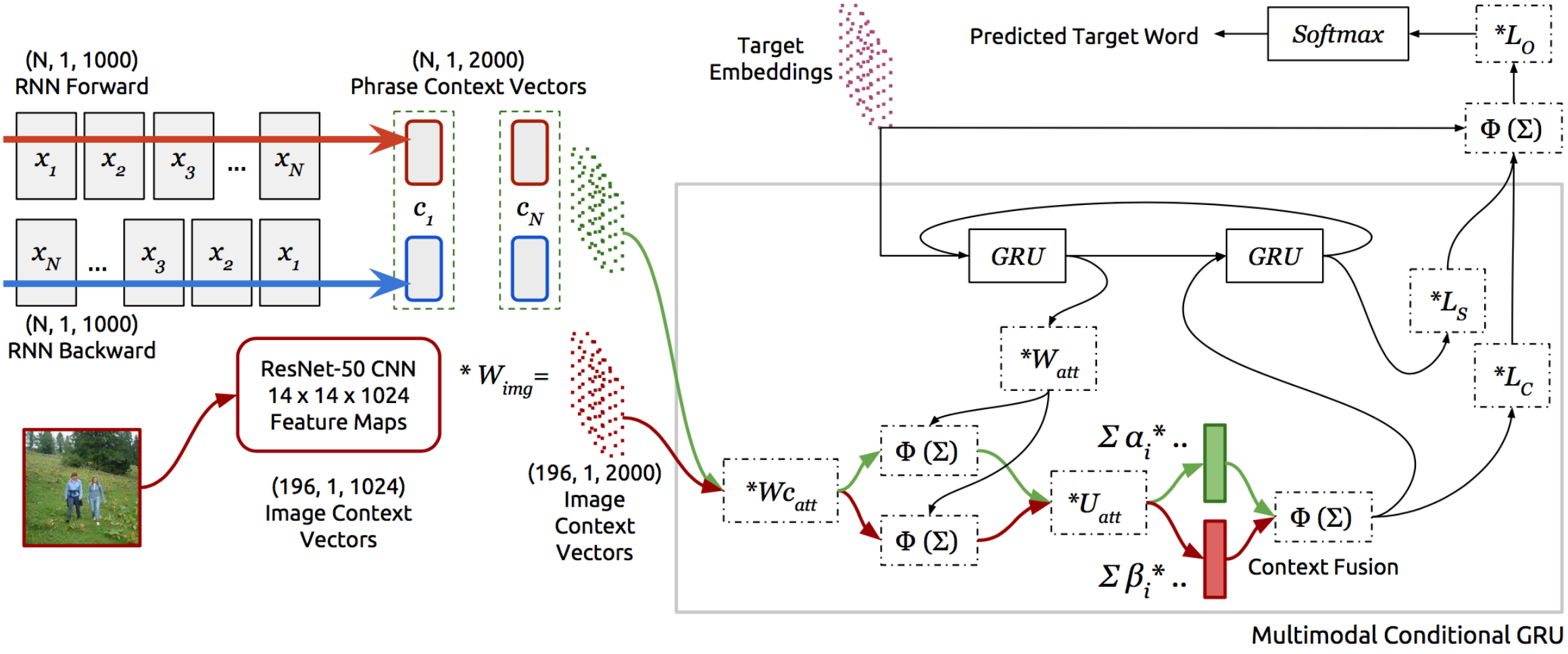
Beyond languages: multimodal translation





(Xu et al., 2015)

Beyond languages: multimodal translation



(Caglayan et al., 2016; Elliott & Kadar, 2017)

What is a sentence?

Is a sentence a sequence of phrases, words, morphemes or characters?

What is a sentence to a neural net?

- Each word/symbol: one-hot vector
- Prior-less encoding
- Permutation invariant

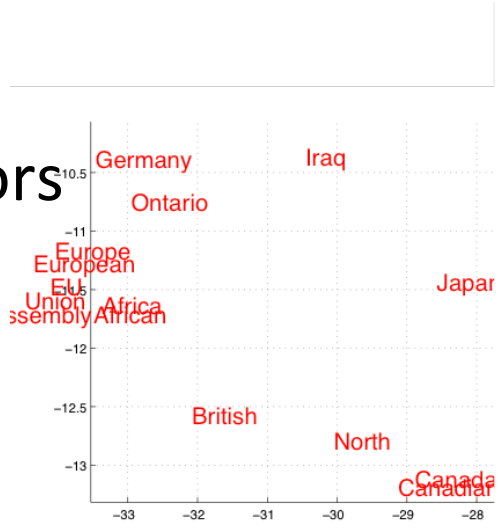
ID	Word
1	the
2	a
2093	**cat**

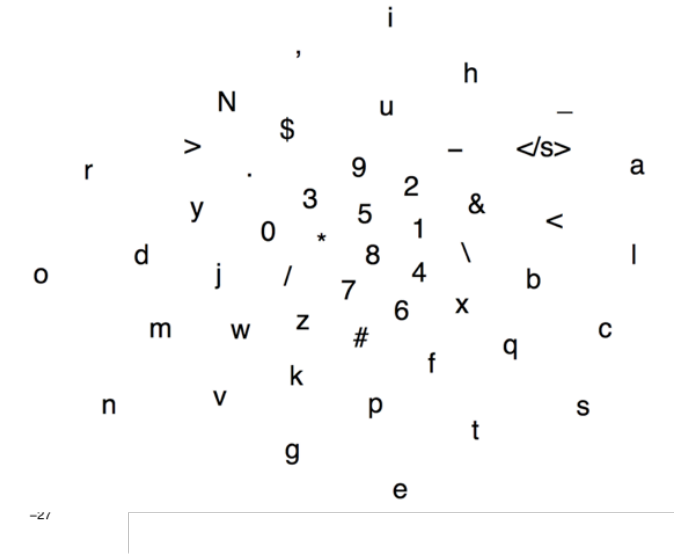
$$e_{\text{cat}} = \begin{bmatrix} 0, \\ \vdots, \\ 0, \\ 1, \\ 0, \\ \vdots, \\ 0 \end{bmatrix}^\top$$

2093-th element

- Sentence
  - To us: a sequence of words
  - To NN: a sequence of one-hot vectors

- *What does it mean?*





Why not words?

- Inefficient handling of various morphological variants
  - Sub-optimal segmentation/tokenization
  - “Etxaberria”, “Etxazarra”, “Etxaguren”, “Etxarren”: four independent vectors
- Lack of generalization to novel/rare morphological variants
  - For instance, ولمركبته in Arabic => “and to his vehicle”
- One vector for compound words?
  - “kolmi/vaihe/kilo/watti/tunti/mittari” => one vector?
  - “kolme” => one vector?
- Spelling issues
  - Social media: full of misspelt and non-trivial words
  - Historical documents: spelling normalization
  - Good segmentation/tokenization needed for each language

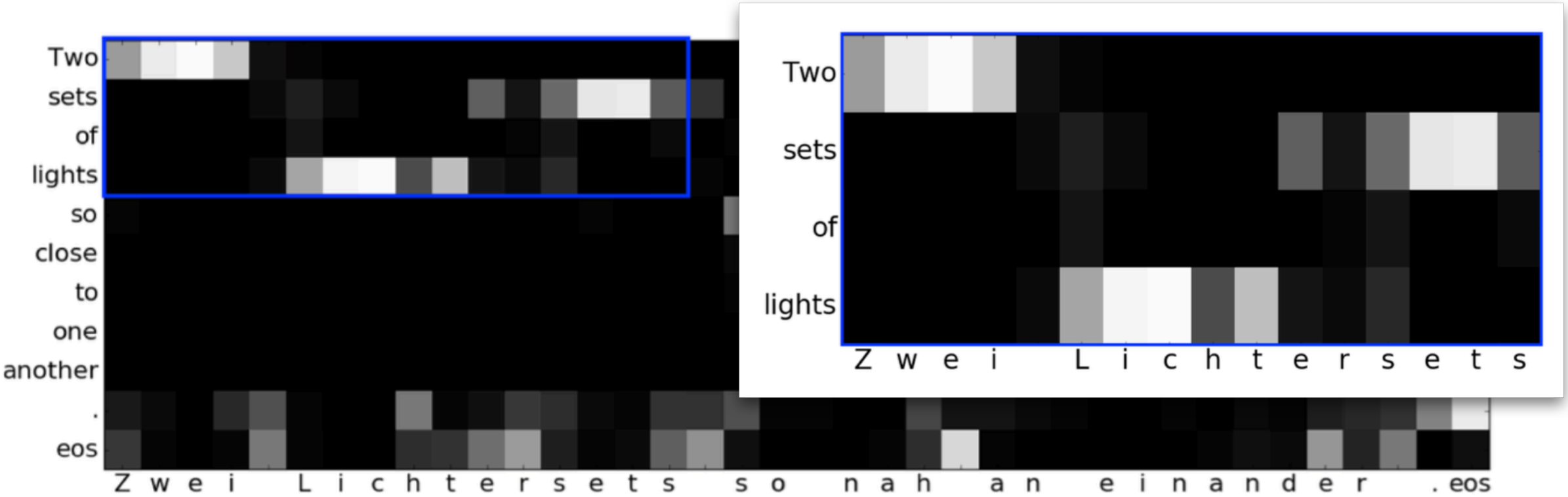
- *So, no, words don’t look like the units we want to work with…*

Then, what should we do…?

- Original: 고양이가 침대 위에 누워있습니다
- Word-level modelling:
  (고양이가, 침대, 위에, 누워있습니다)
- Subword-level modelling (Sennrich et al., 2015; Wu et al., 2016)
  (고양이, 가, 침대, 위, 에, 누워, 있습니, 다)
- Character-level modelling with segmentation
  (Wang et al., 2015; Luong & Manning, 2016; Costa-Jussa & Fonollosa, 2016)
  ((ㄱ,ㅗ,ㅇ,ㅑ,ㅇ,ㅣ,ㄱ,ㅏ), (ㅊ,ㅣ,ㅁ,ㄷ,ㅐ), (ㅇ,ㅟ,ㅇ,ㅔ), (ㄴ,ㅜ,ㅇ,ㅝ,ㅇ,ㅣ,ㅆ,ㅅ,ㅡ,ㅂ,ㄴ,ㅣ,ㄷ,ㅏ))
- Fully character-level modelling (Chung et al., 2016; Lee et al., 2017)
  (ㄱ,ㅗ,ㅇ,ㅑ,ㅇ,ㅣ,ㄱ,ㅏ,_,ㅊ,ㅣ,ㅁ,ㄷ,ㅐ,_,ㅇ,ㅟ,ㅇ,ㅔ,_,ㄴ,ㅜ,ㅇ,ㅝ,ㅇ,ㅣ,ㅆ,ㅅ,ㅡ,ㅂ,ㄴ,ㅣ,ㄷ,ㅏ))
- Visual modelling of a character (Liu et al., 2017)

Character-level translation

- Source: subword-level representation
- Target: character-level representation
- *The decoder implicitly learned word-like units automatically!*

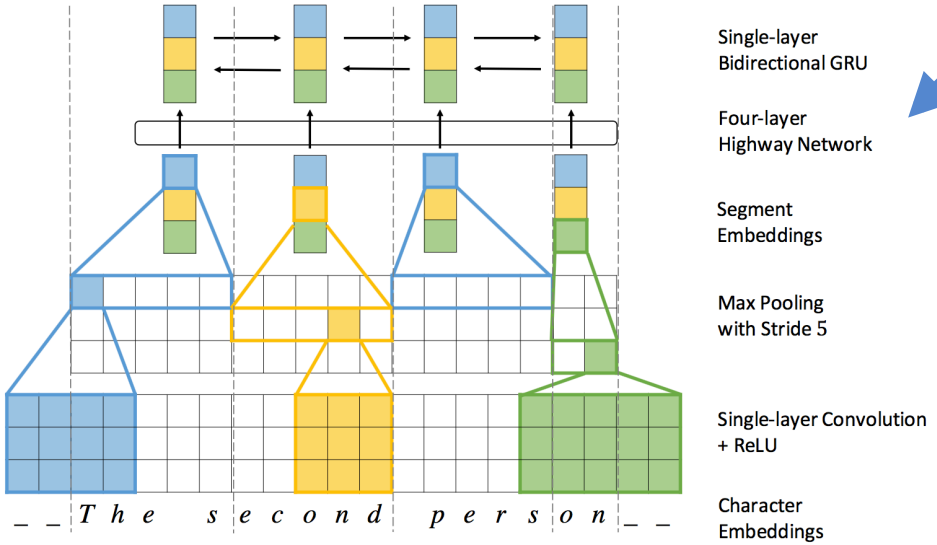


(Chung et al., 2017)

(Lee et al., 2017)

Fully Character-level translation

- Source: character-level representation
- Target: character-level representation
- Efficient modelling with a convolutional-recurrent encoder
- Works as well as, or *better than*, subword-level translation



Depth matters more when the representation is more raw

		Setting	Src	Trg	Dev	Test1	Test2
DE-EN	(a)*	bi	bpe	bpe	24.13		24.00
	(b)	bi	bpe	char	25.64	24.59	25.27
	(c)	bi	char	char	**26.30**	**25.77**	**25.83**
	(d)	multi	bpe	char	24.92	24.54	25.23
	(e)	multi	char	char	25.67	25.13	25.79
CS-EN	(f)*	bi	bpe	bpe	21.24		20.32
	(g)	bi	bpe	char	22.95	23.78	22.40
	(h)	bi	char	char	23.38	24.08	22.46
	(i)	multi	bpe	char	23.27	24.27	22.42
	(j)	multi	char	char	**24.09**	**25.01**	**23.24**
FI-EN	(k)*	bi	bpe	bpe	13.15		12.24
	(l)	bi	bpe	char	14.54		13.98
	(m)	bi	char	char	14.18		13.10
	(n)	multi	bpe	char	14.70		14.40
	(o)	multi	char	char	**15.96**		**15.74**
RU-EN	(p)*	bi	bpe	bpe	21.04		22.44
	(q)	bi	bpe	char	21.68	26.21	22.83
	(r)	bi	char	char	21.75	**26.80**	22.73
	(s)	multi	bpe	char	21.75	26.31	22.81
	(t)	multi	char	char	**22.20**	26.33	**23.33**

(a) Spelling mistakes

DE ori	Warum sollten wir nicht Freunde sei ?
DE src	Warum solltne wir nich Freunde sei ?
EN ref	Why should not we be friends ?
bpe2char	Why are we to be friends ?
char2char	Why should we not be friends ?

(b) Rare words

DE src	Siebentausendzweihundertvierundfünfzig .
EN ref	Seven thousand two hundred fifty four .
bpe2char	Fifty-five Decline of the Seventy .
char2char	Seven thousand hundred thousand fifties .

- More robust to errors
- Better handles rare tokens
  - *Rare tokens are not necessary rare!*

(Lee et al., 2017)

(d) Nonce words

DE src	Der Test ist nun über , aber ich habe keine gute Note . Es ist wie eine Verschlimmbesserung .
EN ref	The test is now over , but i don't have any good grade . it is like a worsened improvement .
bpe2char	The test is now over , but i do not have a good note .
char2char	The test is now , but i have no good note , it is like a worsening improvement .

Character-level Multilingual Translation

- When symbols are shared across multiple languages, why not share a single encoder/decoder for them?
  1. Language transfer at all levels: letters, words, phrases, sentences, …
  2. Intra-sentence code-switching *without* any specific data

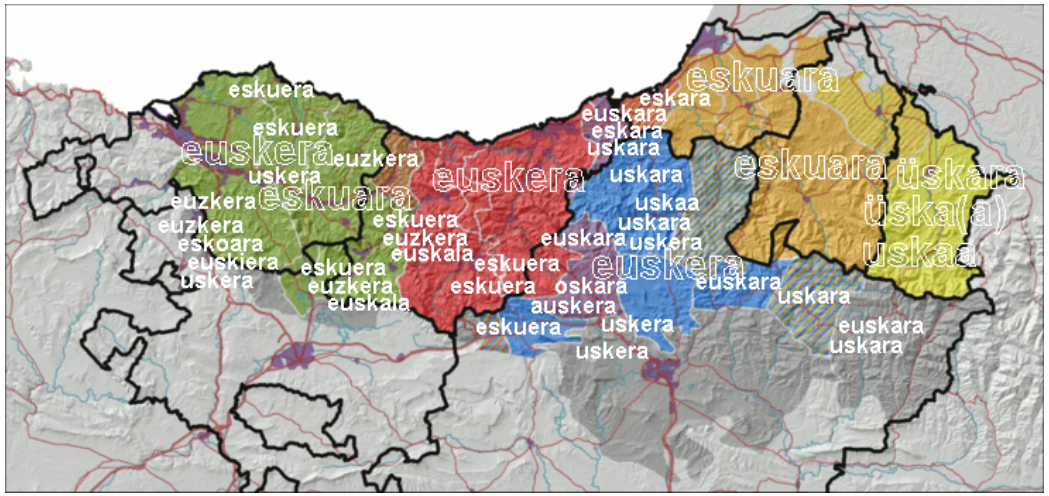
(e) Multilingual

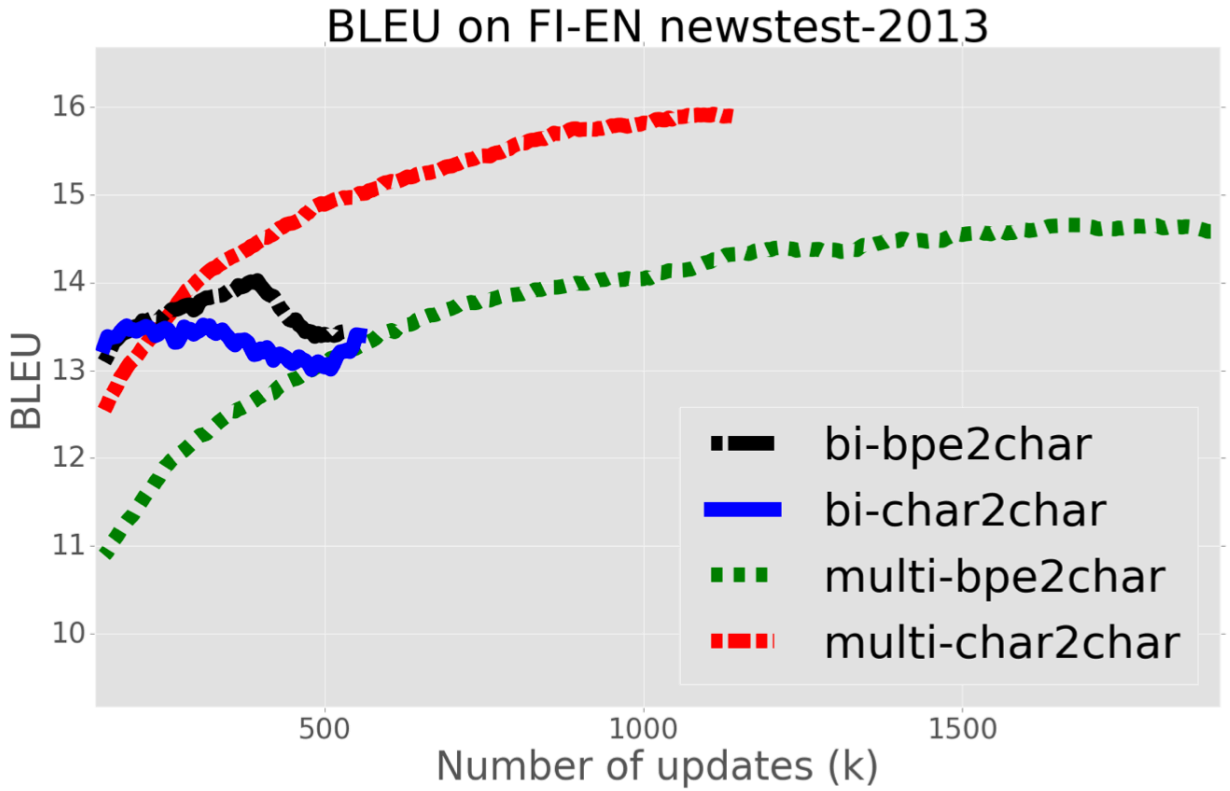
Multi src	Bei der Metropolitního výboru pro dopravu für das Gebiet der San Francisco Bay erklärten Beamte , der Kongress könne das Problem банкротство доверительного Фонда строительства шоссейных дорог einfach durch Erhöhung der Kraftstoffsteuer lösen .
EN ref	At the Metropolitan Transportation Commission in the San Francisco Bay Area , officials say Congress could very simply deal with the bankrupt Highway Trust Fund by raising gas taxes .
bpe2char	During the Metropolitan Committee on Transport for San Francisco Bay , officials declared that Congress could solve the problem of bankruptcy by increasing the fuel tax bankrupt .
char2char	At the Metropolitan Committee on Transport for the territory of San Francisco Bay , officials explained that the Congress could simply solve the problem of the bankruptcy of the Road Construction Fund by increasing the fuel tax .

(Lee et al., 2017; Johnson et al., 2016; Ha et al., 2016)

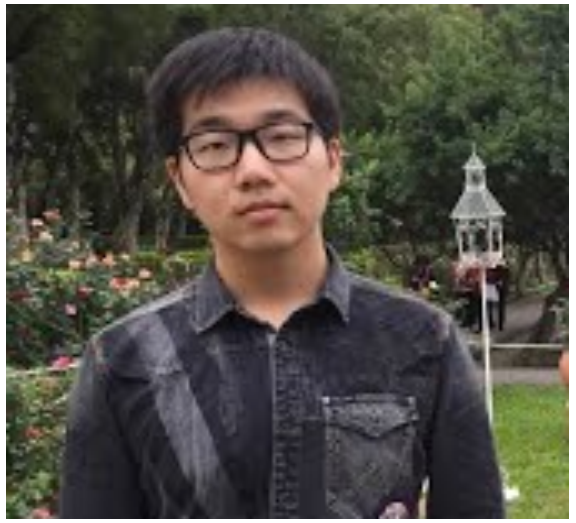
Character-level Multilingual Translation

- Prevents overfitting with low-resource language pairs
- Perhaps, a way to build a MT system for Basque languages?
  - Many dialects: Biscayan, Gipuzkoan, Navarrese, Navarro-Lapurdian, Souletin





(Lee et al., 2017)



Trainable Decoding of Neural Machine Translation

Jiatao Gu, Graham Neubig, K Cho and Victor Li. Learning to Translate in Real-time with Neural Machine Translation. EACL 2017.

Jiatao Gu, K Cho and Victor Li. Trainable Greedy Decoding for Neural Machine Translation. EMNLP 2017.

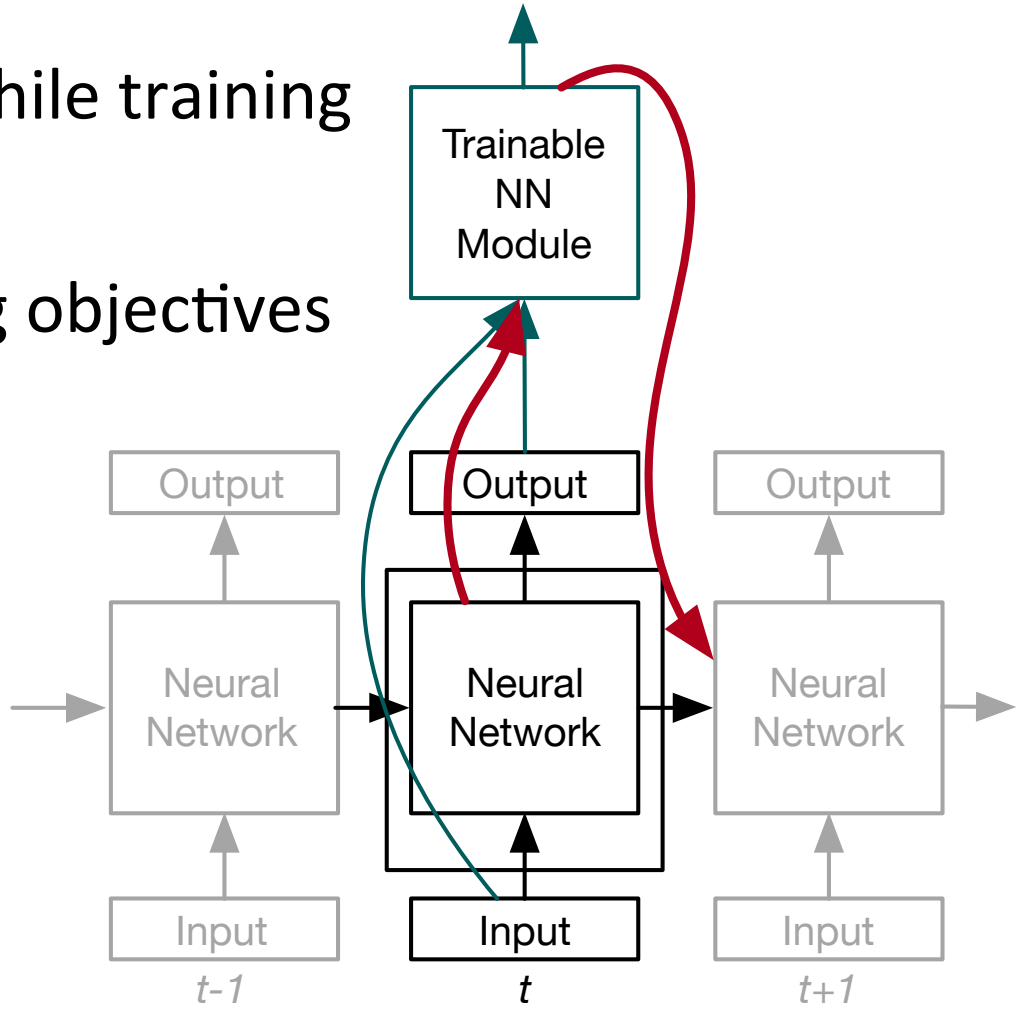
Trainable Decoding

Motivation

- Many decoding objectives unknown while training
- Lack of target training examples
- Arbitrary (non-differentiable) decoding objectives

Our Approach

- Train NMT with supervised learning
- Train a decoding module on top



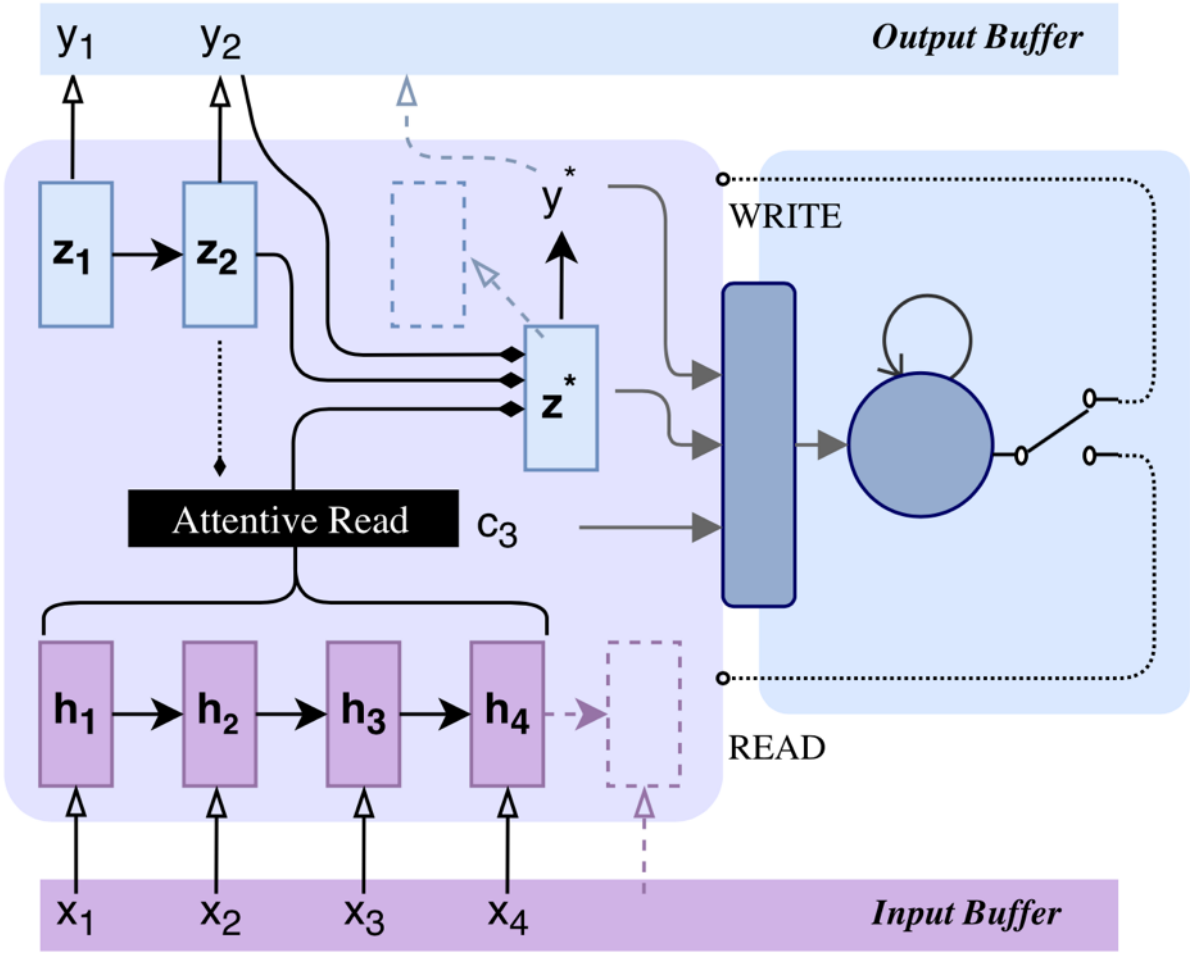
(1) Real-Time Translation

Decoding

1. Start with a pretrained NMT
2. A simultaenous decoder intercepts and interprets the incoming signal
3. The simultaneous decoder forces the pretrained model to either
   1. output a target symbol, or
   2. wait for a next source symbol

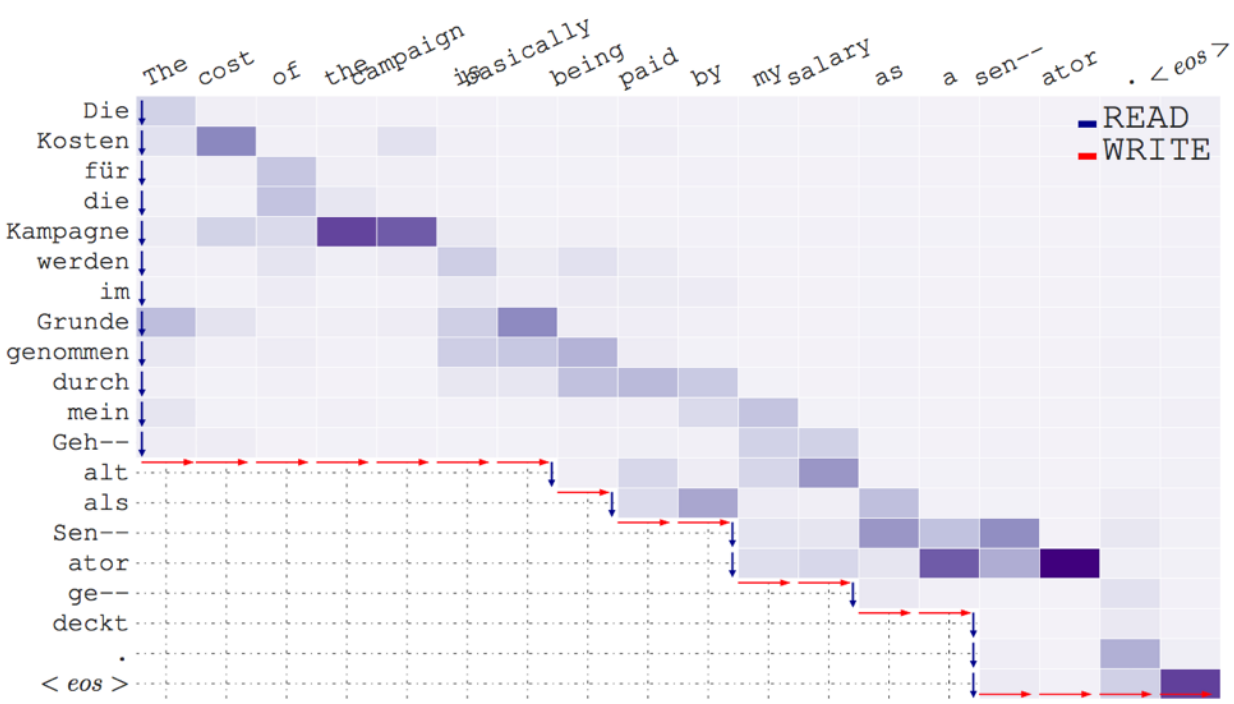
Learning

1. Trade-off between delay and quality
2. Stochastic policy gradient (REINFORCE)

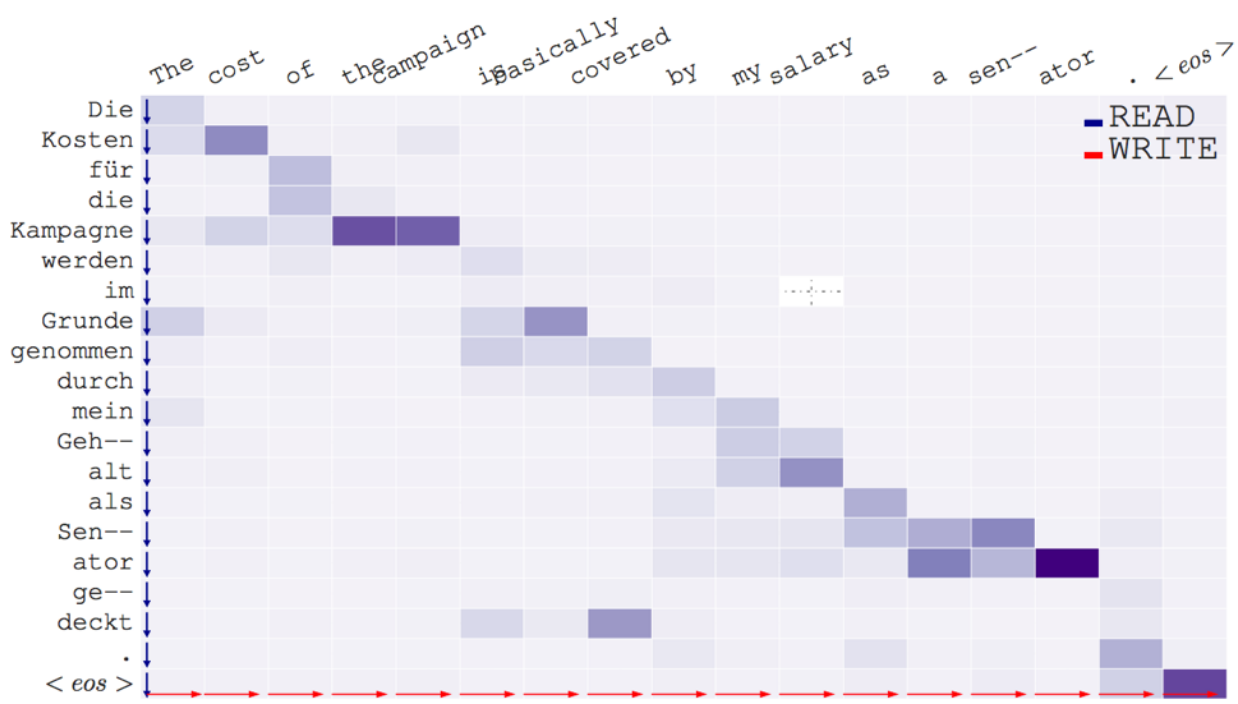


(Gu, Neubig, Cho & Li, EACL 2017)

(1) Real-Time Translation



(a) Simultaneous Neural Machine Translation



(b) Neural Machine Translation

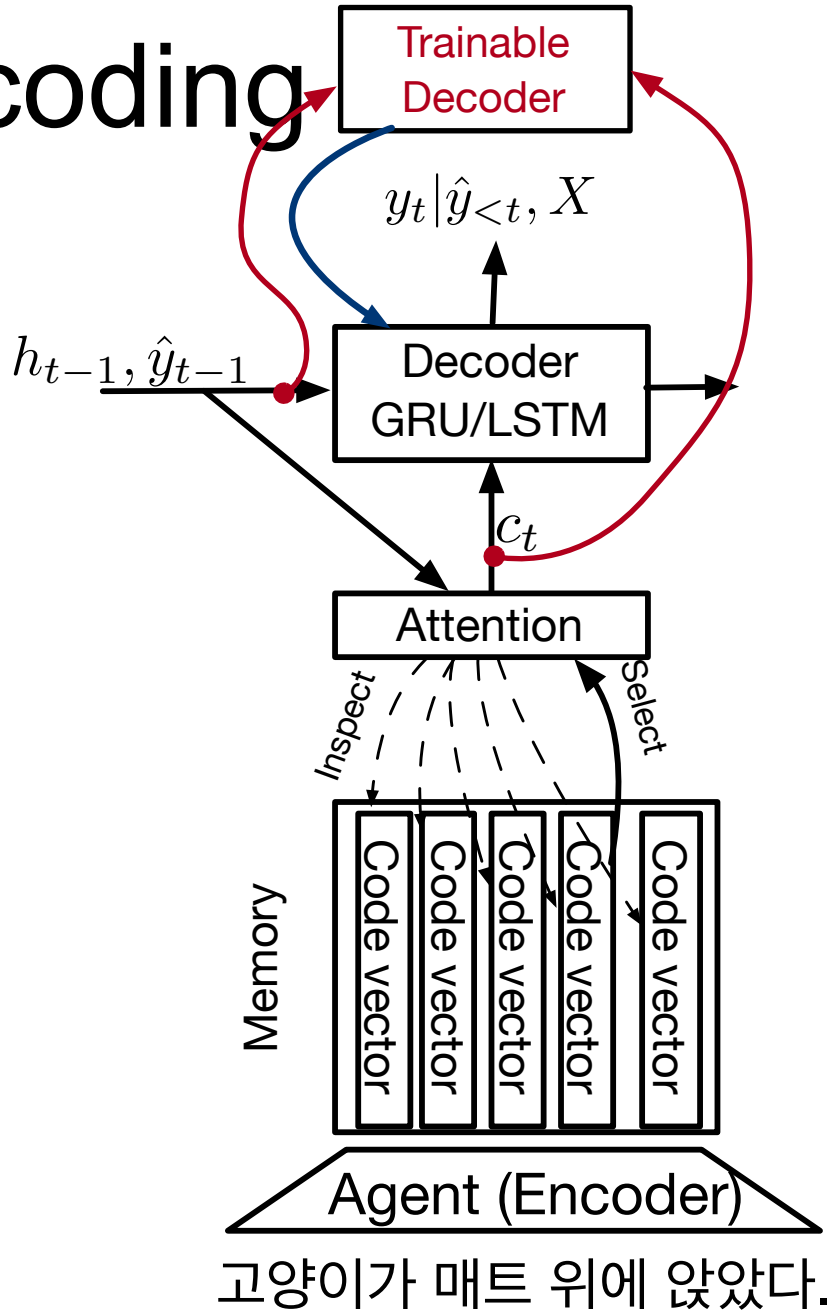
(2) Trainable Greedy Decoding

Decoding

1. Start with a pretrained NMT
2. A Trainable decoder intercepts and interprets the incoming signal
3. The trainable decoder sends out the altering signal back to the pretrained model

Learning

1. Deterministic policy gradient
2. Maximize any arbitrary objective

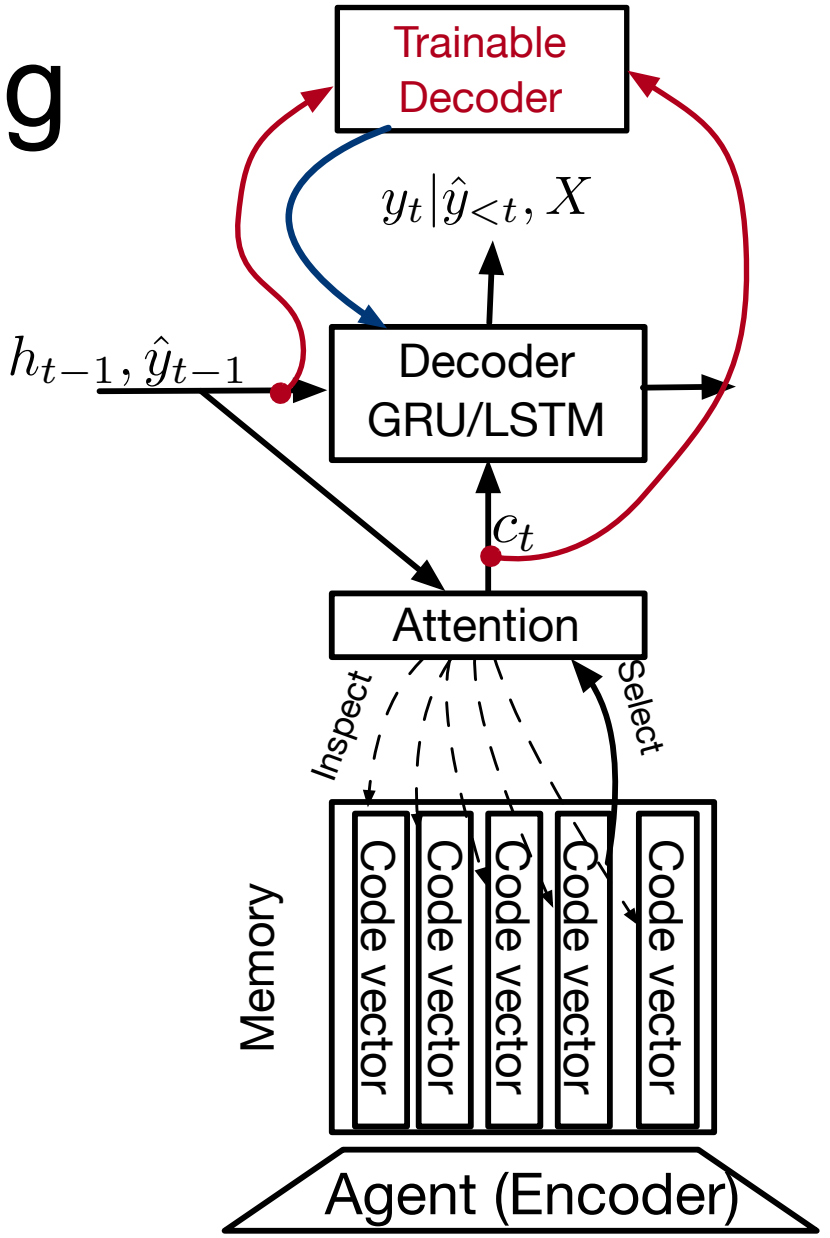


(Gu, Cho & Li, 2017)

(2) Trainable Greedy Decoding

Models

1. Actor $\pi : \mathbb{R}^{3d} \rightarrow \mathbb{R}^d$
   - Input: prev. hid. state h_{t-1}, prev. symbol $\hat{y}_{t-1}$, and context c_t from the attention model
   - Output: additive bias for hid. state z_t
   - Example:
     $$z_t = U\sigma\left(W\left[h_{t-1}; E(\hat{y}); c_t\right] + b\right) + c$$
2. Critic $R^c : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \rightarrow \mathbb{R}$
   - Input: a sequence of the hidden states from the decoder
   - Output: a predicted return
   - In our case, the critic estimates the full return rather than Q at each time step



(Gu, Cho & Li, 2017)

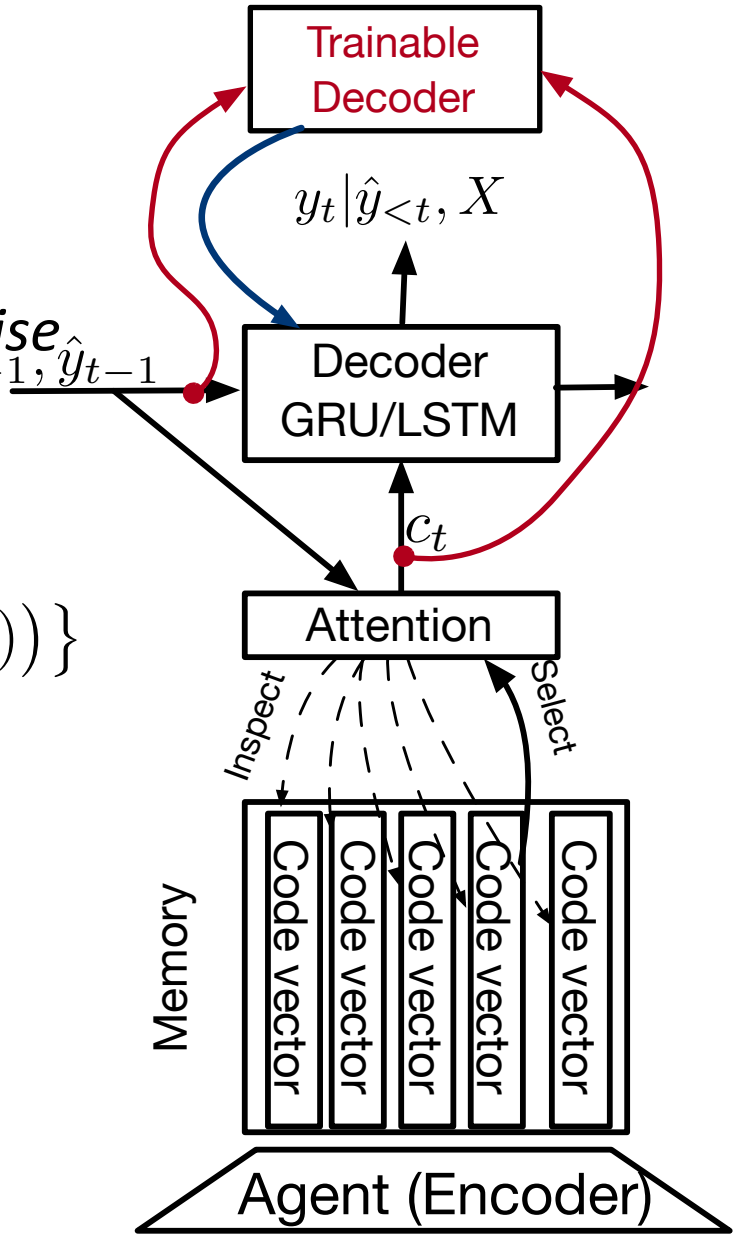
(2) Trainable Greedy Decoding

Learning

1) Generate translation given a source sentence with *noise* $((h_1, z_1), \ldots, (h_T, z_T))$ and R
2) Train the critic to minimize $(R^c(h_1, \ldots, h_T) - R)^2$
3) Generate multiple translations with *noise* $\left\{ \left((h_1^1, z_1^1), \ldots, (h_T^1, z_T^1)\right), \ldots, \left((h_1^M, z_1^M), \ldots, (h_T^M, z_T^M)\right) \right\}$
4) Critic-aware actor learning: *newly proposed*

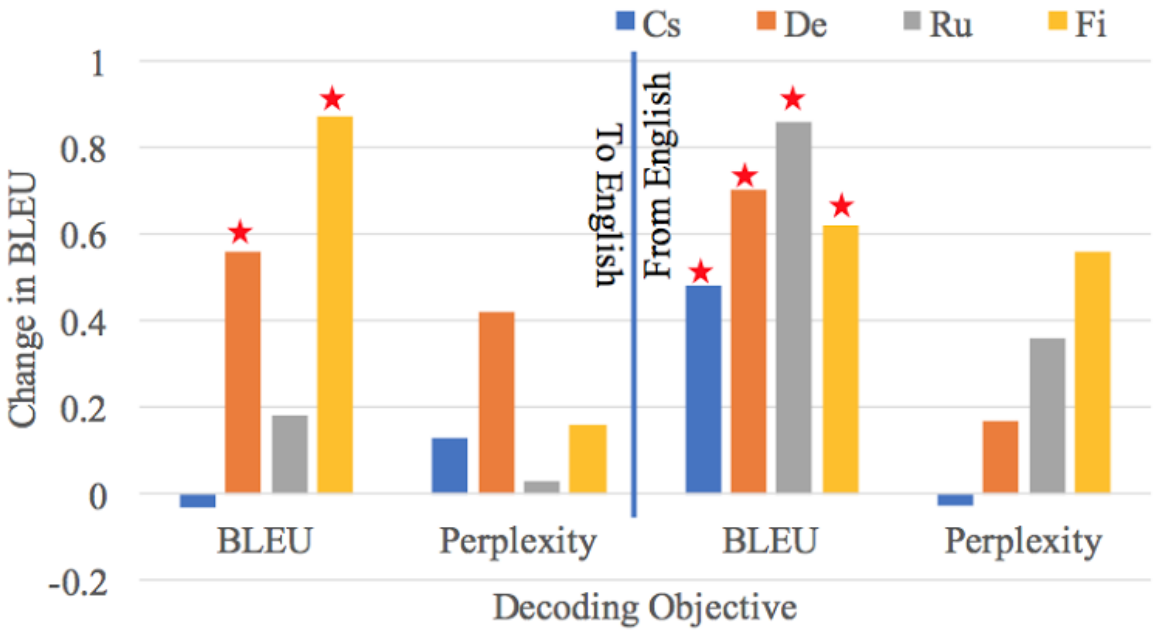
$$\mathbb{E}_Q \left[\frac{\partial R_\psi^c}{\partial \phi} \right], \textit{ where } Q(\epsilon) \propto \exp(\underbrace{-(R_\psi^c - R)^2/\tau}_{\text{Critic-awareness}}) \exp(\underbrace{-\frac{\epsilon^2}{2\sigma^2}}_{\text{Locality}})$$

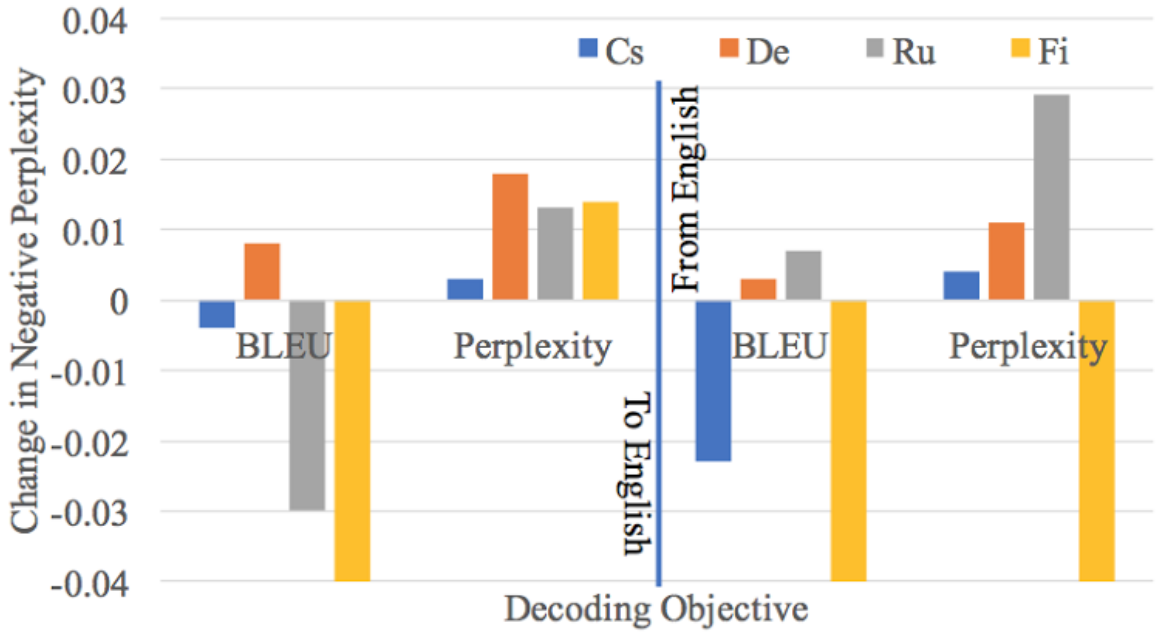
Inference: simply throw away the critic and use the actor



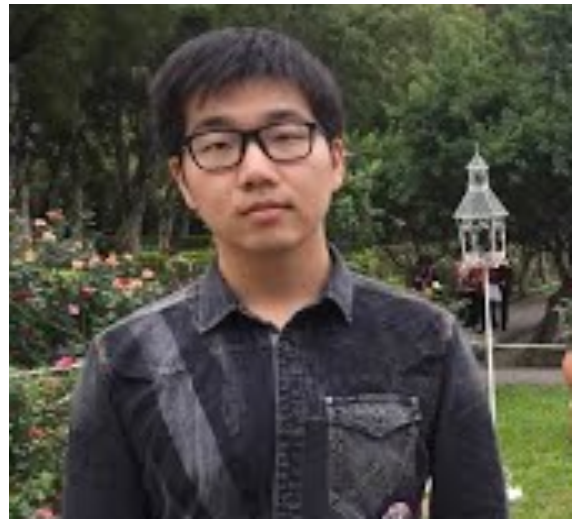
(Gu, Cho & Li, EMNLP 2017)

(2) Trainable Greedy Decoding





- The trainable decoder does improve the target decoding objective
- Training is quite unstable without the critic-aware actor learning algorithm
- More work is definitely needed for further improvement

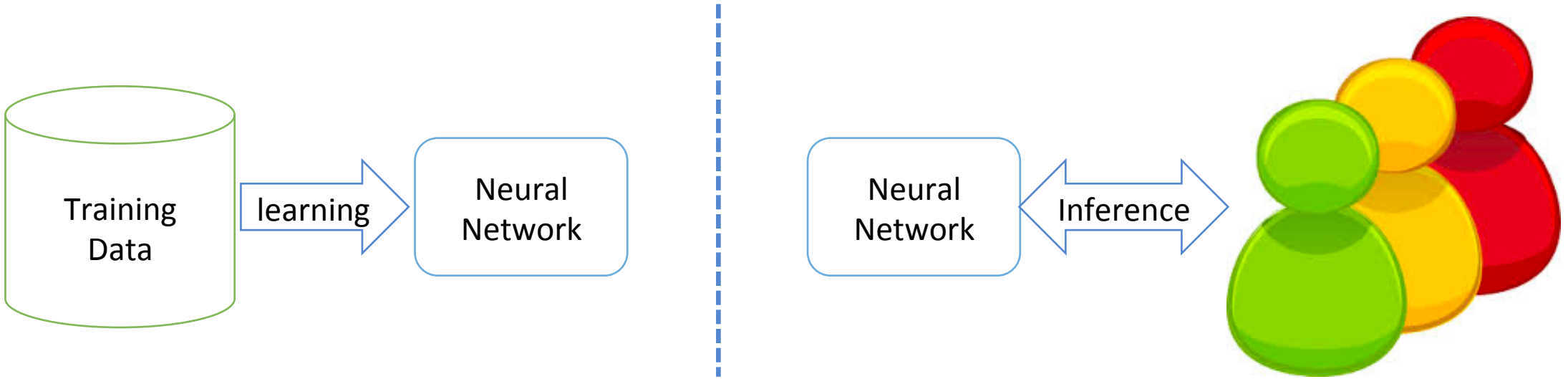


Non-parametric neural machine translation

Gu, Wang, Cho, Li. Search Engine Guided Non-Parametric Neural Machine Translation. About to be rejected from NIPS'17.

Parametric ML: Learning as Compression

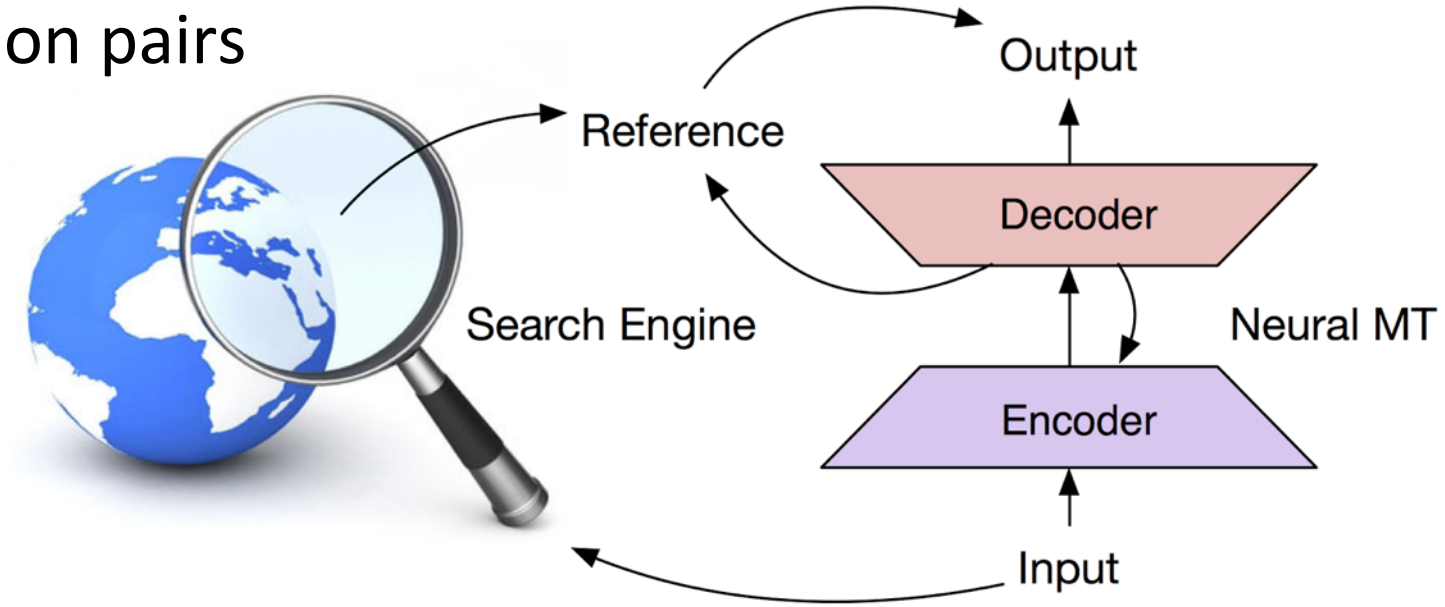
- What does learning do?



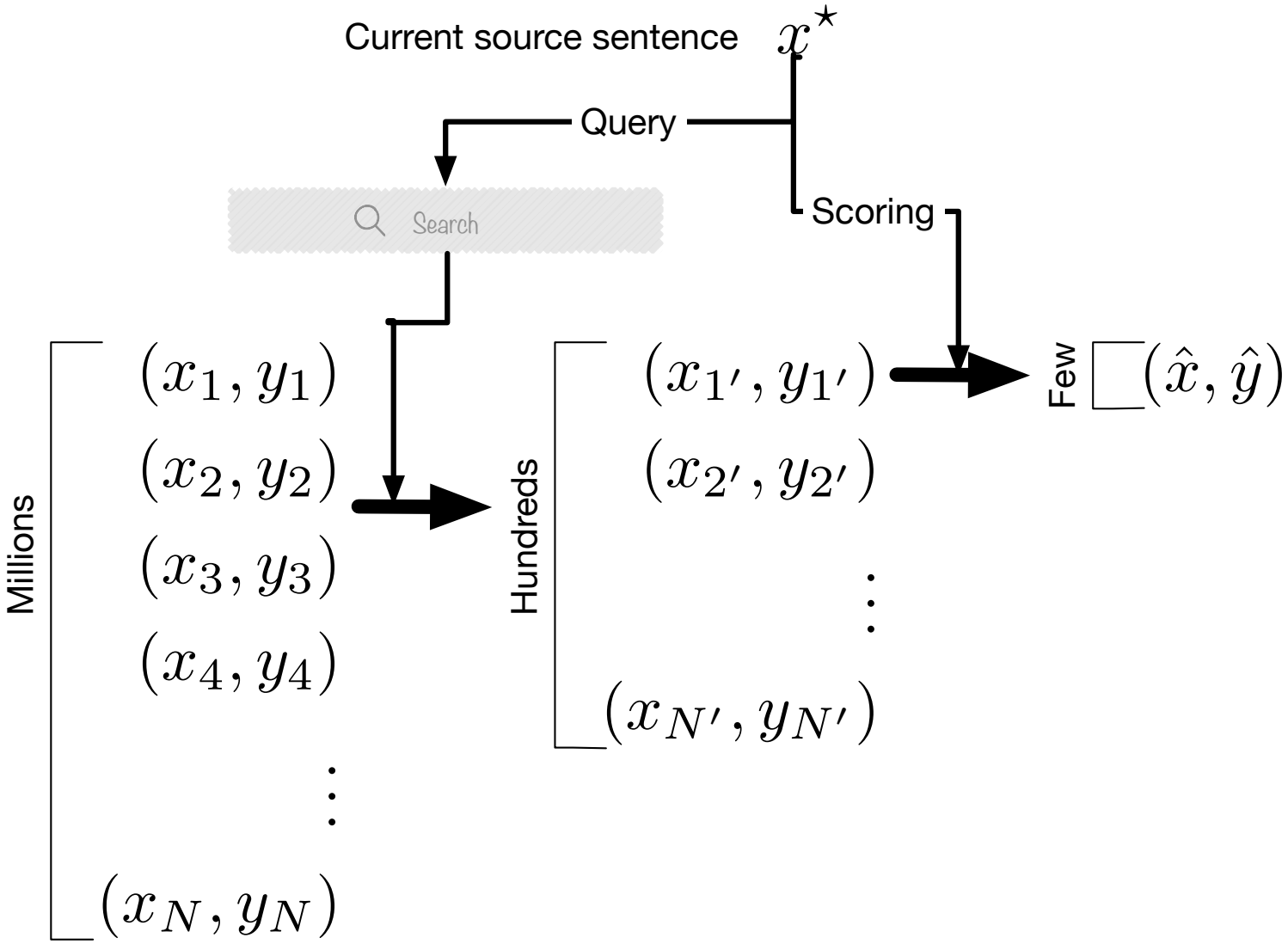
- Parametric machine learning: **data compression** + pattern matching

Non-Parametric NMT

- Bring the whole training corpus together with a model
- Retrieved a small subset of examples using a fast search engine
- Let NMT figure out how to fuse
    1. the current sentence, and
    2. the retrieved translation pairs



Retrieval Stage



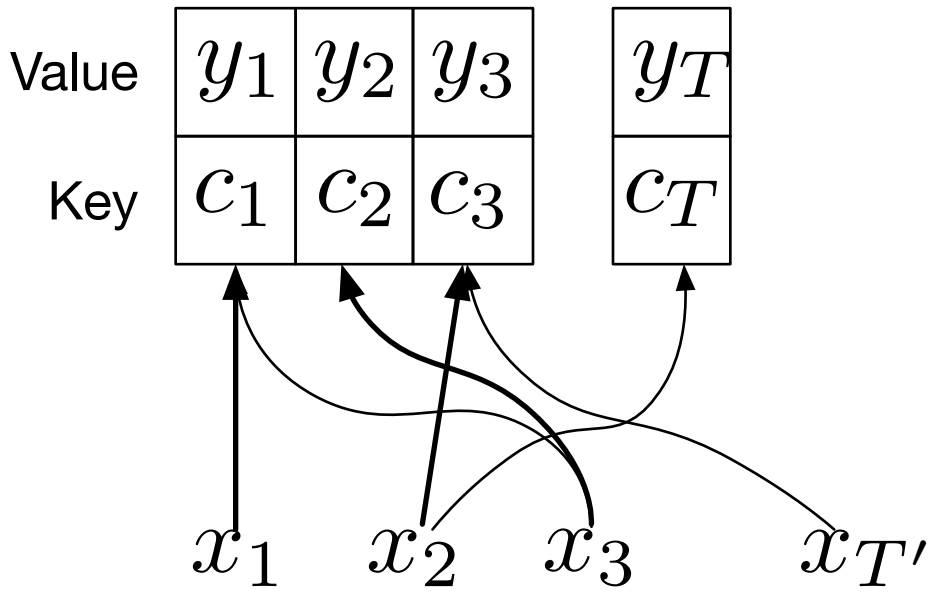
- Off-the-shelf search engine indexes the whole training set
- The engine is queried with a current source sentence
- Only top-few translation pairs are selected at the end

- Extremely efficient!

(Gu et al., 2017)

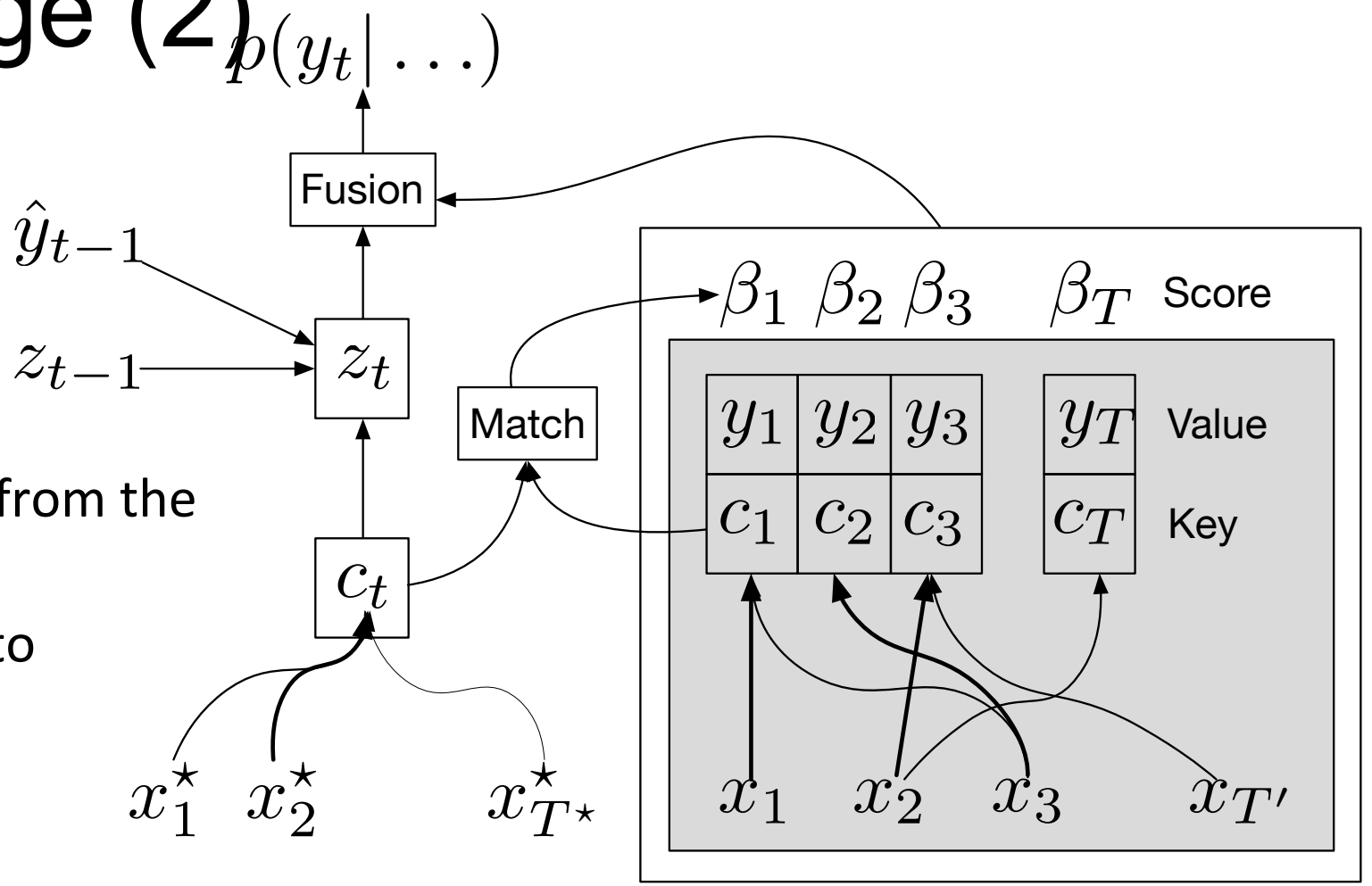
Translation Stage (1)

- Store the retrieved pair in a key-value memory (Gulcehre et al., 2016; Henaff et al., 2017)
- Use the attention-based neural machine translation
  - Key: the context vector
  - Value: the target symbol



(Gu et al., 2017)

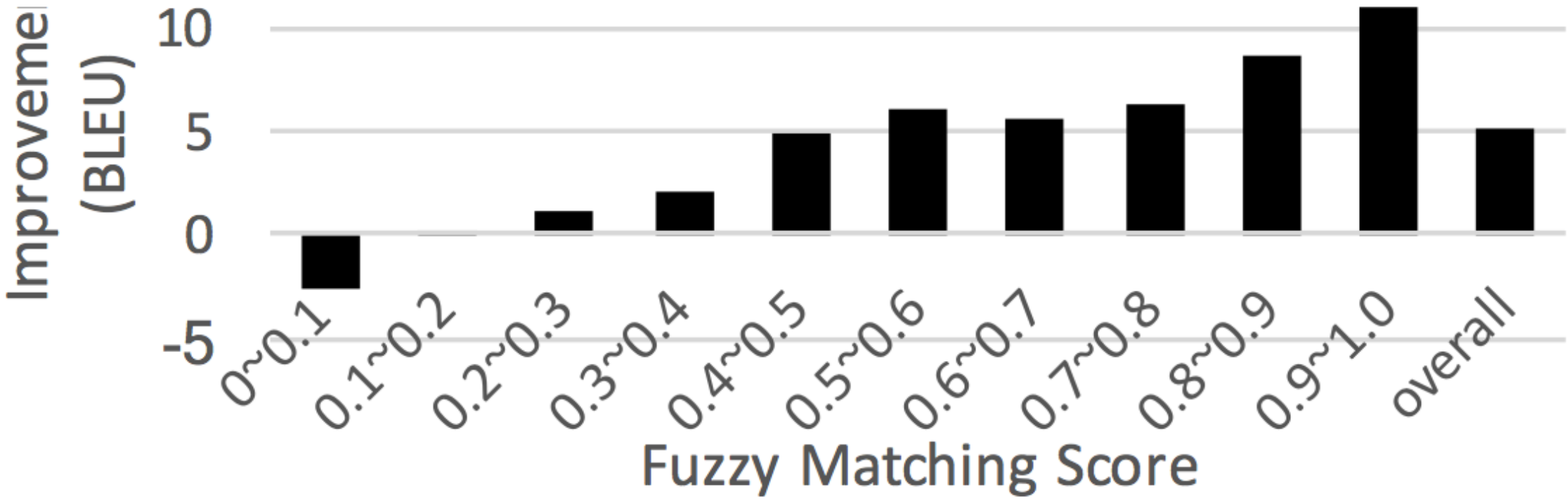
Translation Stage (2)



- Retrieves relevant target symbols from the memory
- Incorporate the retrieved value into computing the target distribution
- Similar to larger-context NMT
  - [Wang et al., 2017; Jean et al., 2017]
- Similar to NMT with external knowledge
  - [Ahn et al., 2016; Bahdanau et al., 2017]

(Gu et al., 2017)

Better retrieval, better translation



More consistent translation

S:	La Commission adopte une décision sur les demandes de révision des programmes opérationnels dans les plus brefs délais à compter de la soumission formelle de la demande par l " État membre .	Il y aurait lieu de remplacer " dans les plus brefs délais " par un délai précis .		RS:	5 .La Commission adopte chaque programme opérationnel dans les plus brefs délais après sa soumission formelle par l " État membre .	Il y aurait lieu de remplacer " dans les plus brefs délais " par un délai précis (actuellement le délai est de cinq mois) .	
A:	The Commission shall adopt a decision on the requests for revision of operational programmes as soon as possible after the formal submission of the request by the Member State .	The phrase " as soon as possible " should be replaced by a exact deadline .		RT:	5. The Commission shall adopt each operational programme as soon as possible after its formal submission by the Member State .	The phrase " as soon as possible " should be replaced with an exact deadline (the deadline is currently five months) .	
B:	The Commission shall adopt a decision on applications for revision of operational programmes as quickly as possible from the formal submission of the application by the Member State .	The Commission should be replaced as soon as possible " by a precise period .		T:	The Commission shall adopt a decision on the requests for revision of operational programmes as soon as possible after formal submission of the request by the Member State .	The phrase " as soon as possible " should be replaced with an exact deadline .	

Fuzzy matching score: 0.49, Edit distance (TM-NMT=3, NMT=17)

- Translation with consistent style and word choice
- Personalized translation

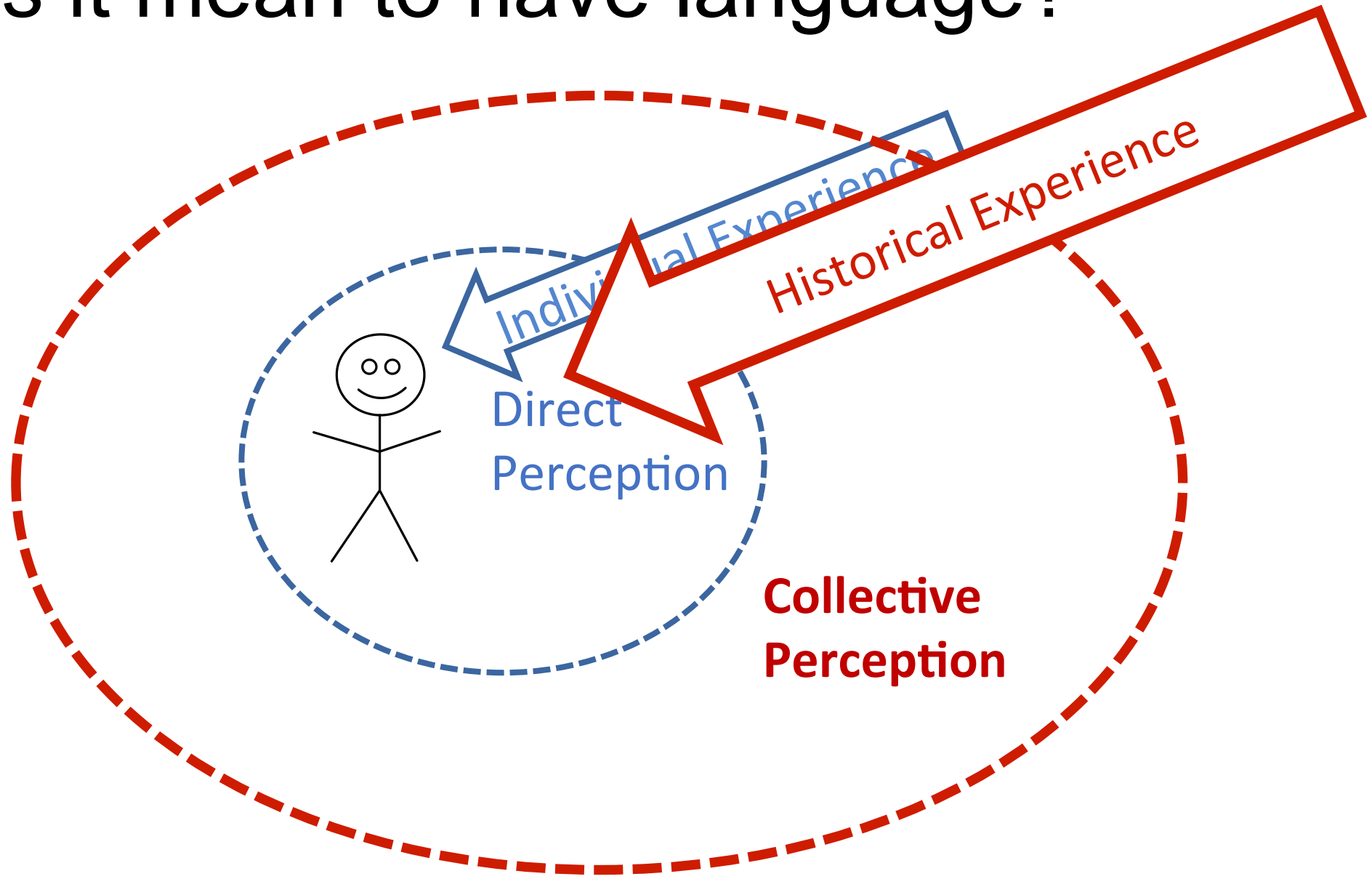
Language Modelling with External Knowledge

Hill, Cho, Korhonen, Bengio. Learning to understand phrases by embedding the dictionary. TACL 2015.

Ahn, Choi, Pämamaa, Bengio. A Neural Knowledge Language Model. arXiv 2016.

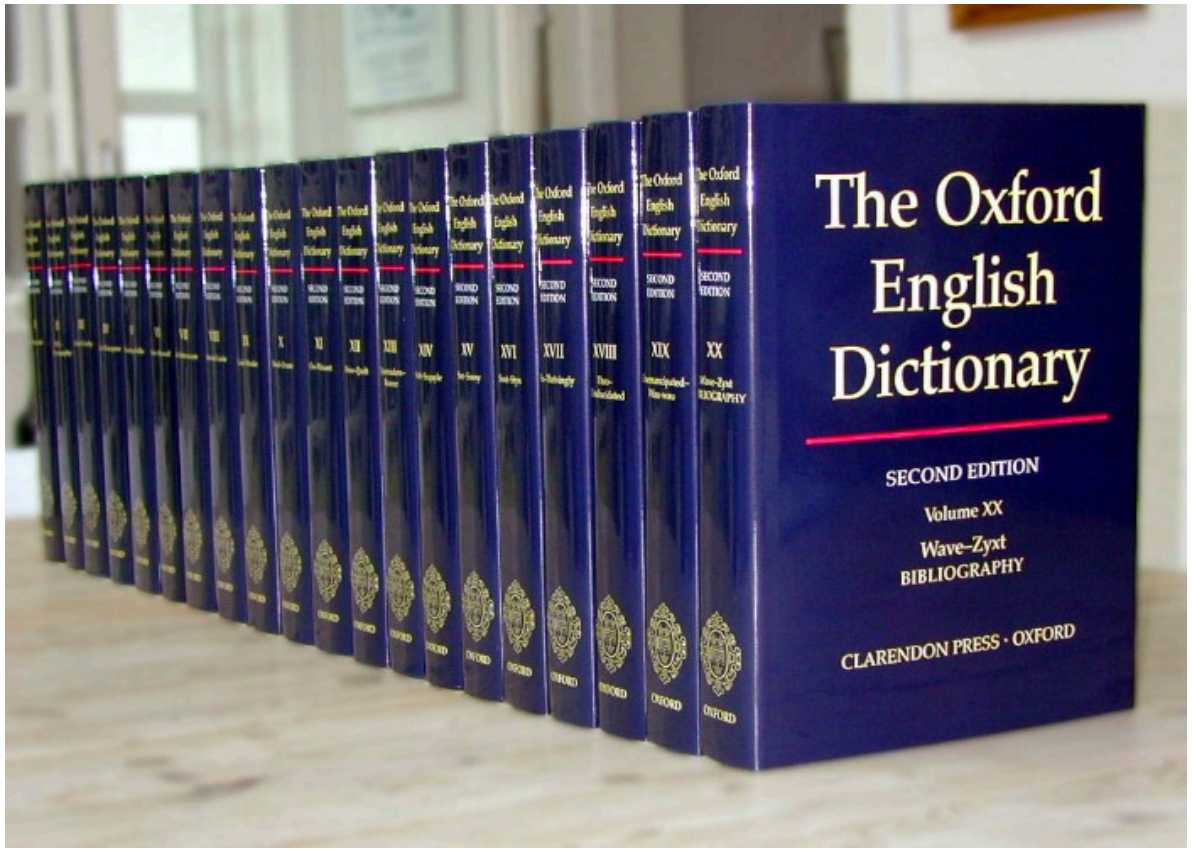
Bahdanau et al. Learning to Compute Word Embeddings on the Fly. arXiv 2017.

What does it mean to have language?



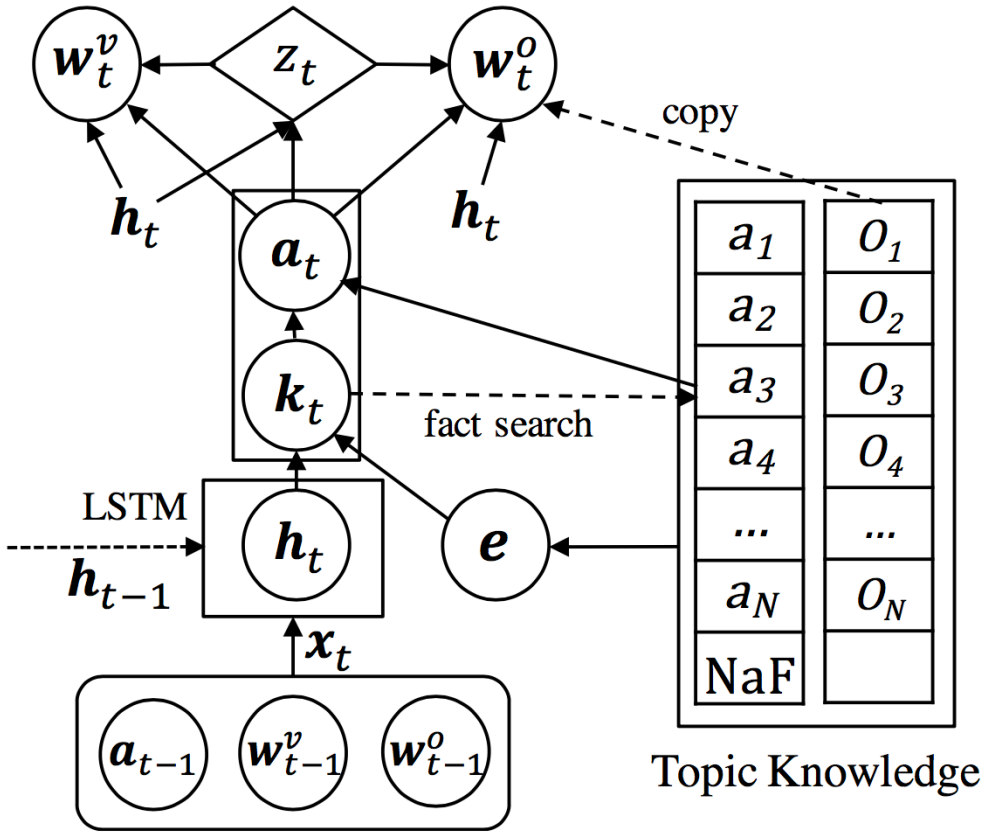
Sources of collective, historical knowledge

- Some structured, some unstructured external knowledge sources



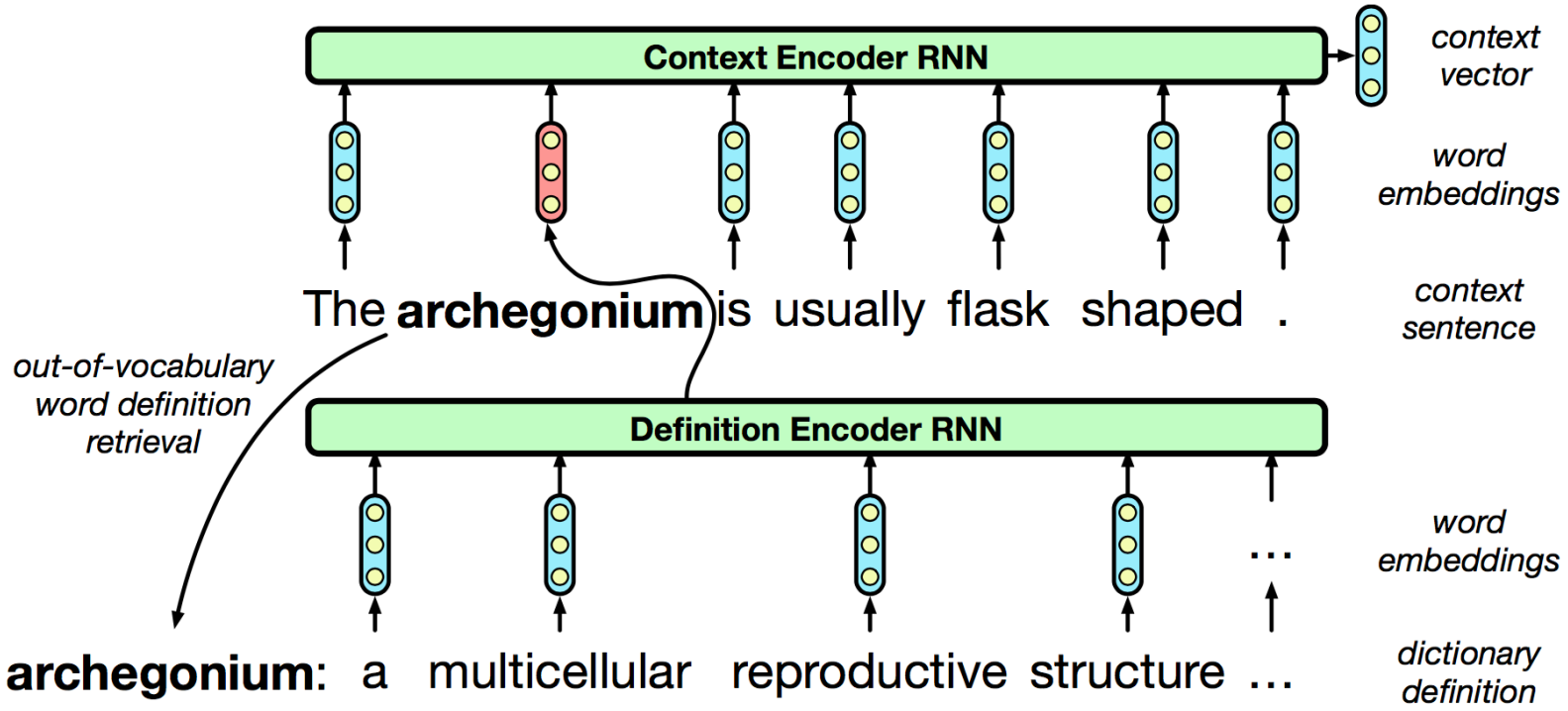
Neural Knowledge Language Modelling

- (Symbol, Fact) pairs from knowledge base
- Neural language model searches for a relevant fact
- If deemed appropriate, use the symbol associated with the retrieved fact



(Ahn et al., 2016)

Dictionary-based word embeddings



- Difficult to learn about rare words + the vocabulary explodes
- The def's of rare words are often available in a dictionary
- The dictionary definition => rare word embedding vector

(Bahdanau et al., 2017)

Other advances in neural machine translation

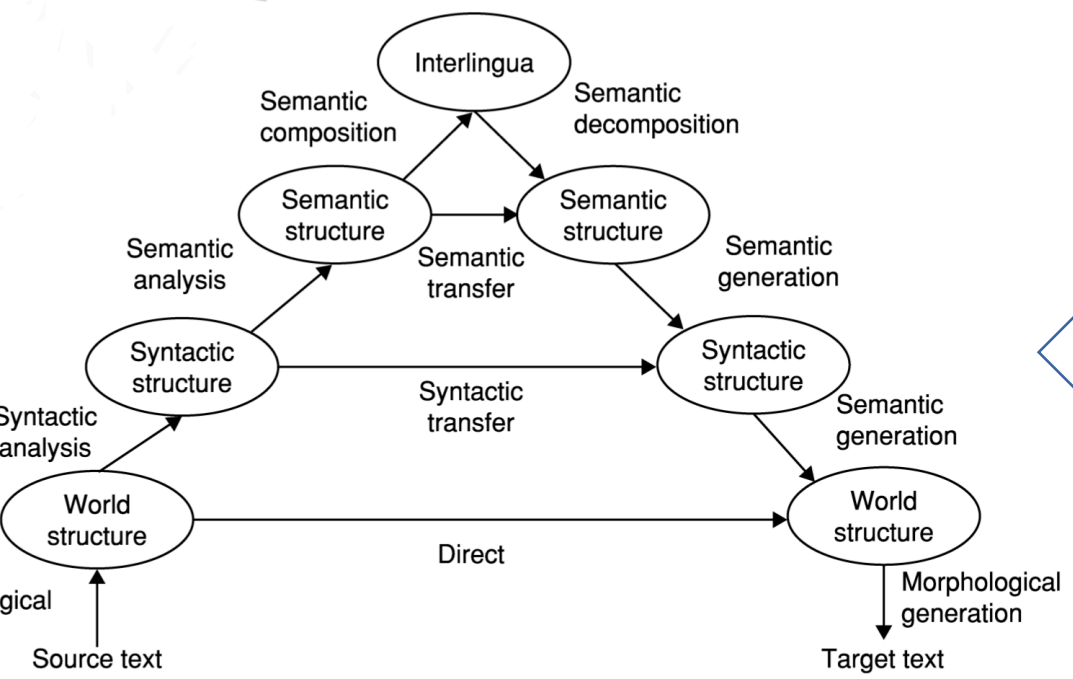
- Discourse-level machine translation
  - [Jean et al., 2017; Wang et al., 2017]
- Better decoding strategies
  - Learning-to-search [Wiseman & Rush, 2016]
  - Reinforcement learning [Shen et al., 2016; Ranzato et al., 2015; Bahdanau et al., 2015]
  - Trainable decoding [Gu et al., 2017ab]
  - Alternative decoding cost [Li et al., 2016; Li et al., 2017]
  - Continuous Relaxation [Cohn et al., 2016]
- Linguistics-guided neural machine translation
  - Learning to parse and translate
    [Eriguchi et al., 2017; Rohee & Goldberg, 2017; Luong et al., 2016]
  - Syntax-aware neural machine translation [Nadejde et al., 2017; **Bastings** et al., 2017]
- See **Chris Dyer**'s lecture slides from this morning

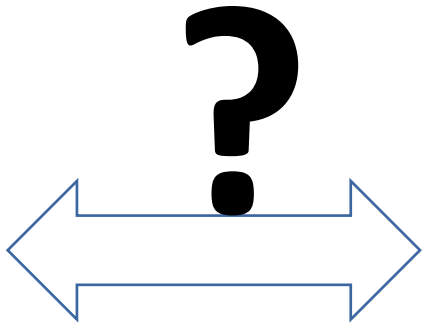
A monitor at the summer school

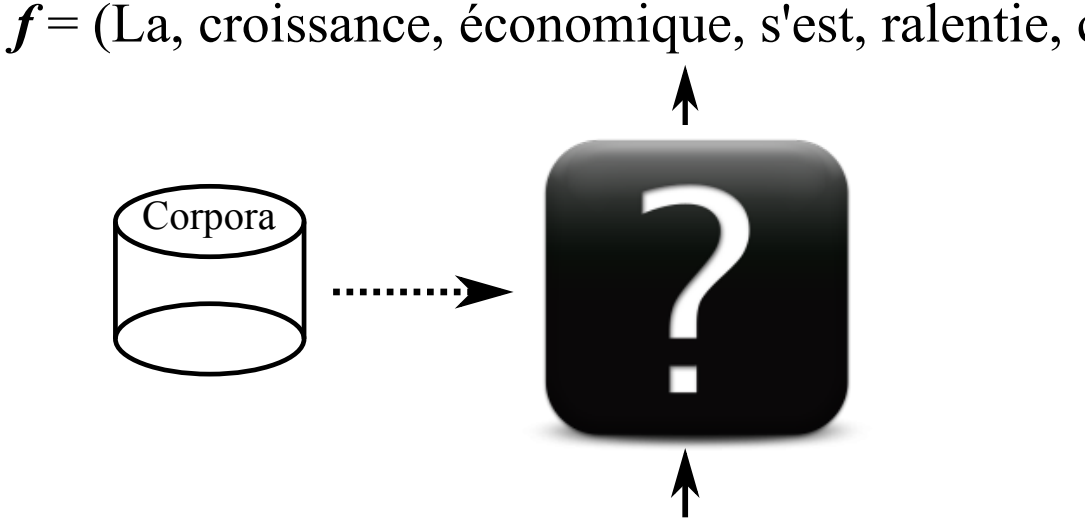
Task-driven Linguistics

Understanding what neural networks have learned about languages

Task-driven Linguistics (1)



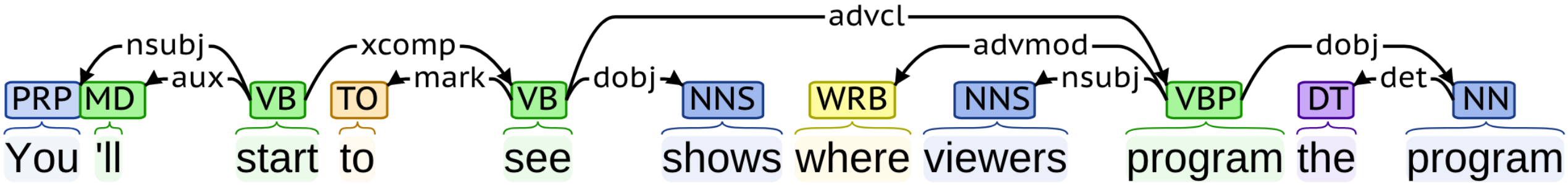




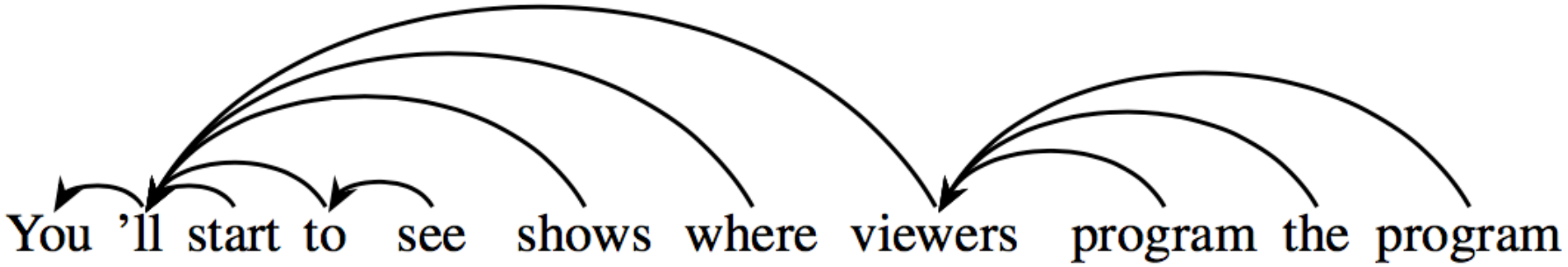
- What has the neural net learned about natural languages?
- Can it tell us one or two things about language that we didn't know?

Task-driven Linguistics (2)

- Dependency parser (Stanford parser)



- Unsupervised recurrent language model (Lee, Levy & Zettlemoyer, 2017)



Task-driven Linguistics (3)

The Germanic Languages (Harbert, 2006)

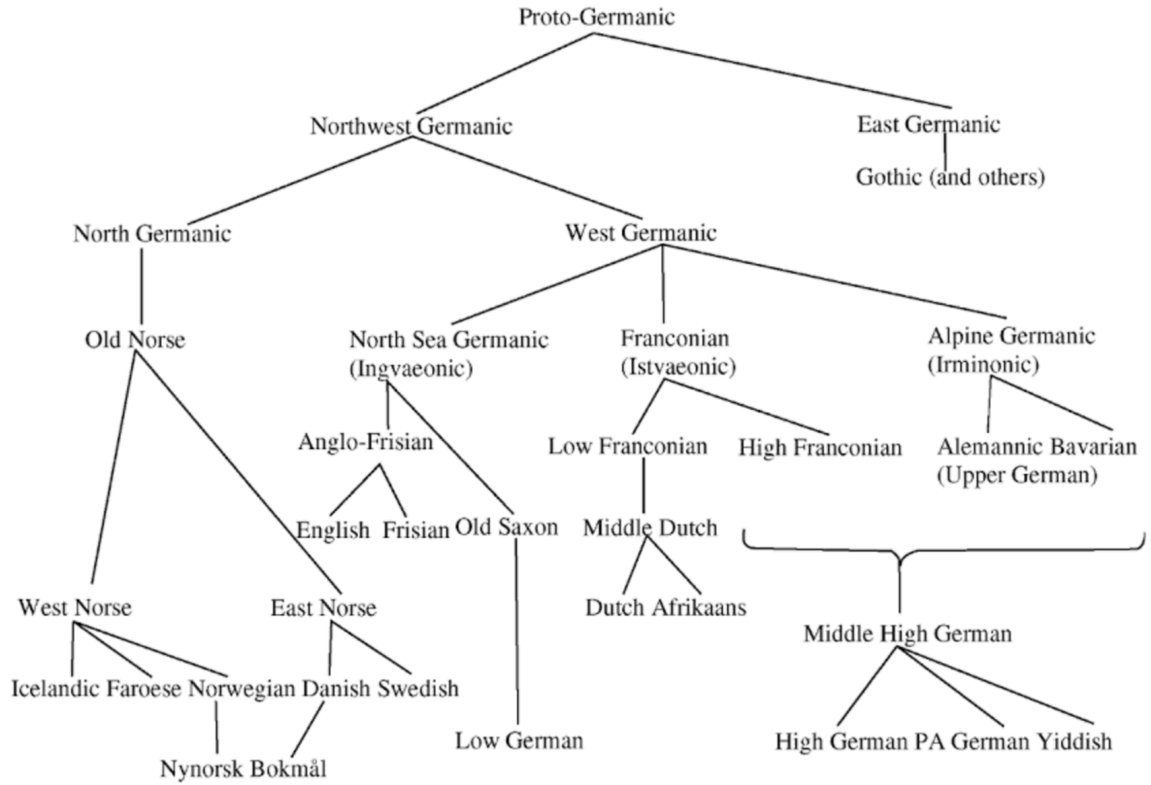
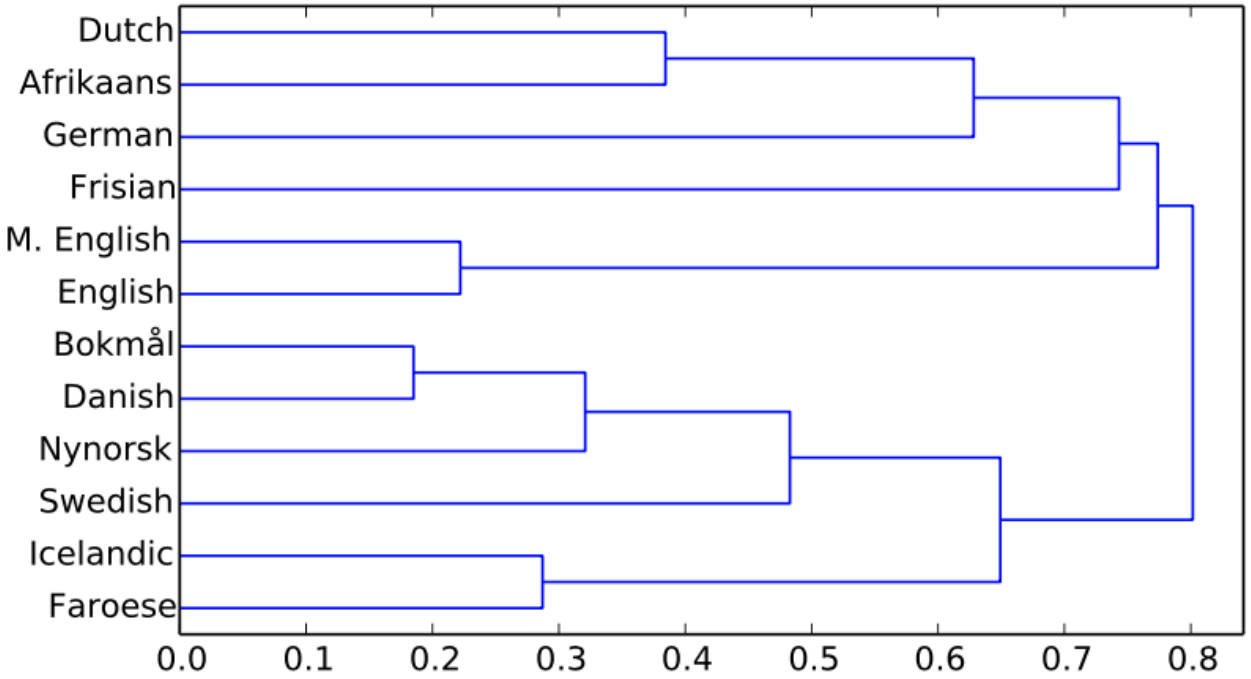


Figure 1.1 *The Germanic Family Tree*

Multilingual recurrent language model

(Östling & Tiedemann, 2016)

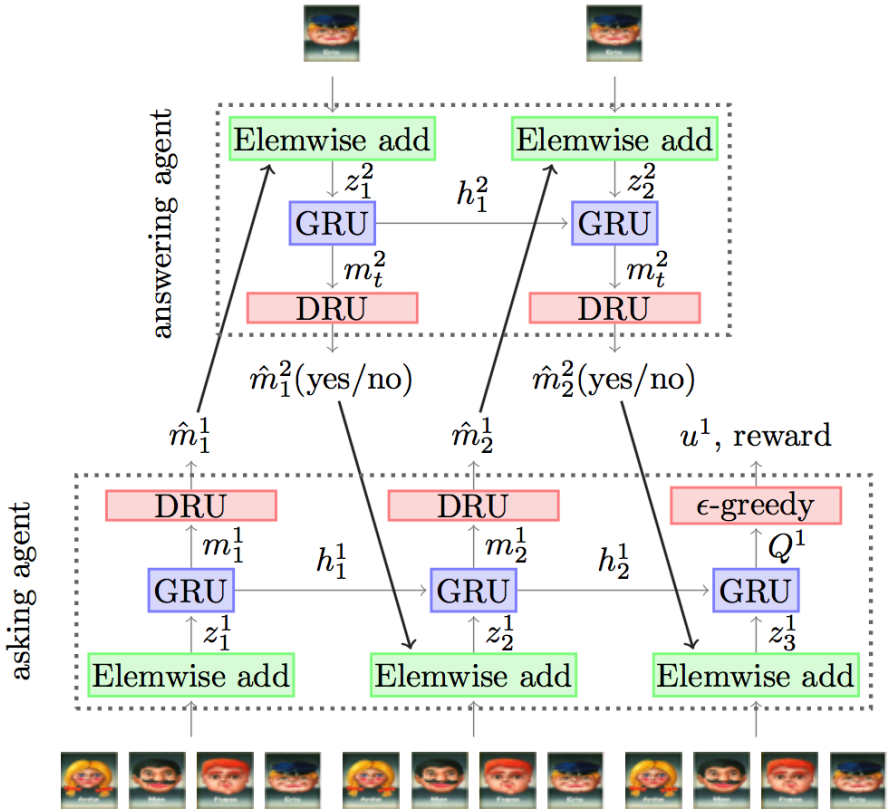


Task-driven Linguistics (4)

Pragmatics and **Grounding**

Communicating neural networks

(Lazaridou et al., 2017; Jorge et al., 2017; Evtimova et al.; and others)



Task-driven Linguistics – Proposal

1. Build/update a end-to-end trainable system for a task
2. Train the system for the task using (large) data
3. Analyze the trained system with respect to known facts/properties
4. Extract new knowledge out of the trained system.
5. Go back to 1

Thank you!