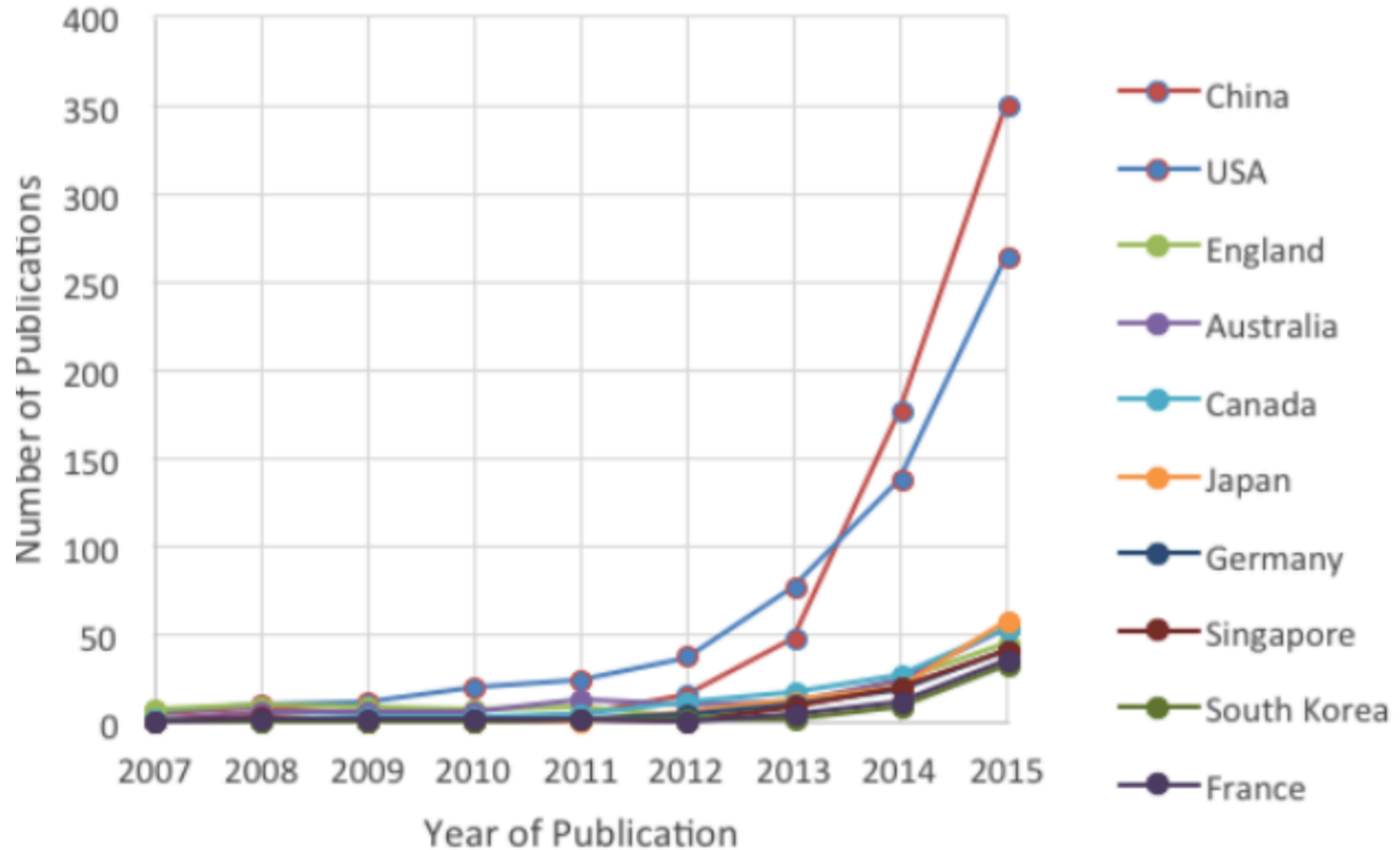


Deep Learning – Neural Machine Translation

Hassan Sajjad

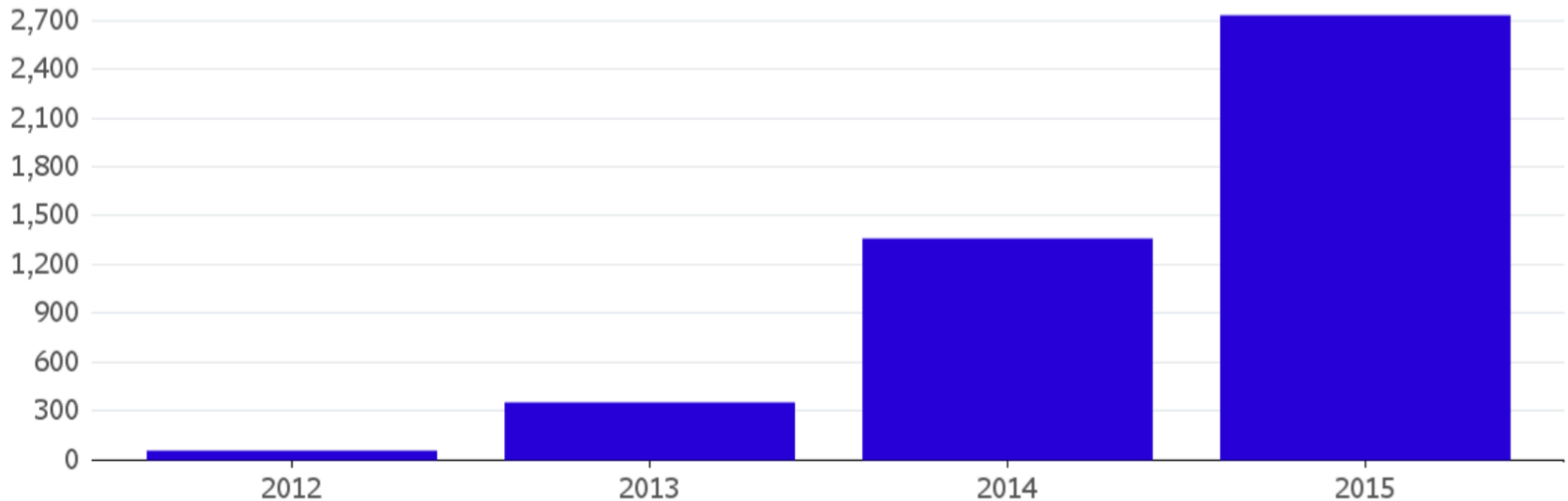
Qatar Computing Research Institute,
Hamad Bin Khalifa University, Qatar

Research Community on Deep Learning



Artificial Intelligence at Google

Number of software projects within Google that uses a key AI technology, called Deep Learning.



Source: Google

Note: 2015 data does not incorporate data from Q4

Bloomberg 

A Few Applications of Deep Learning

Game of Go: 4:1



Image Colorization



<http://richzhang.github.io/colorization/>

Image Synthesis



Input style



Input content

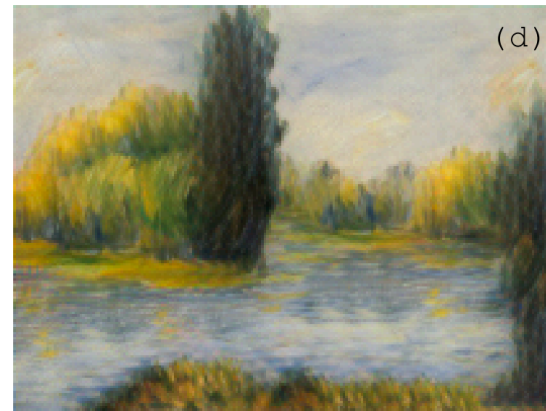


Gatys et al



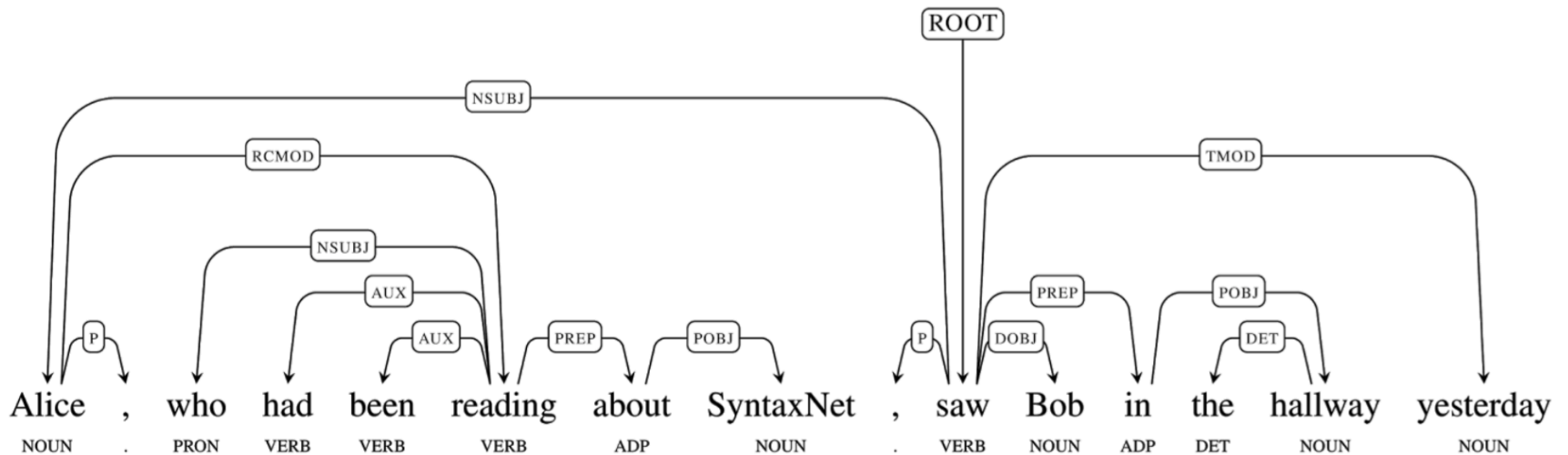
Ours

Synthesizing Painting



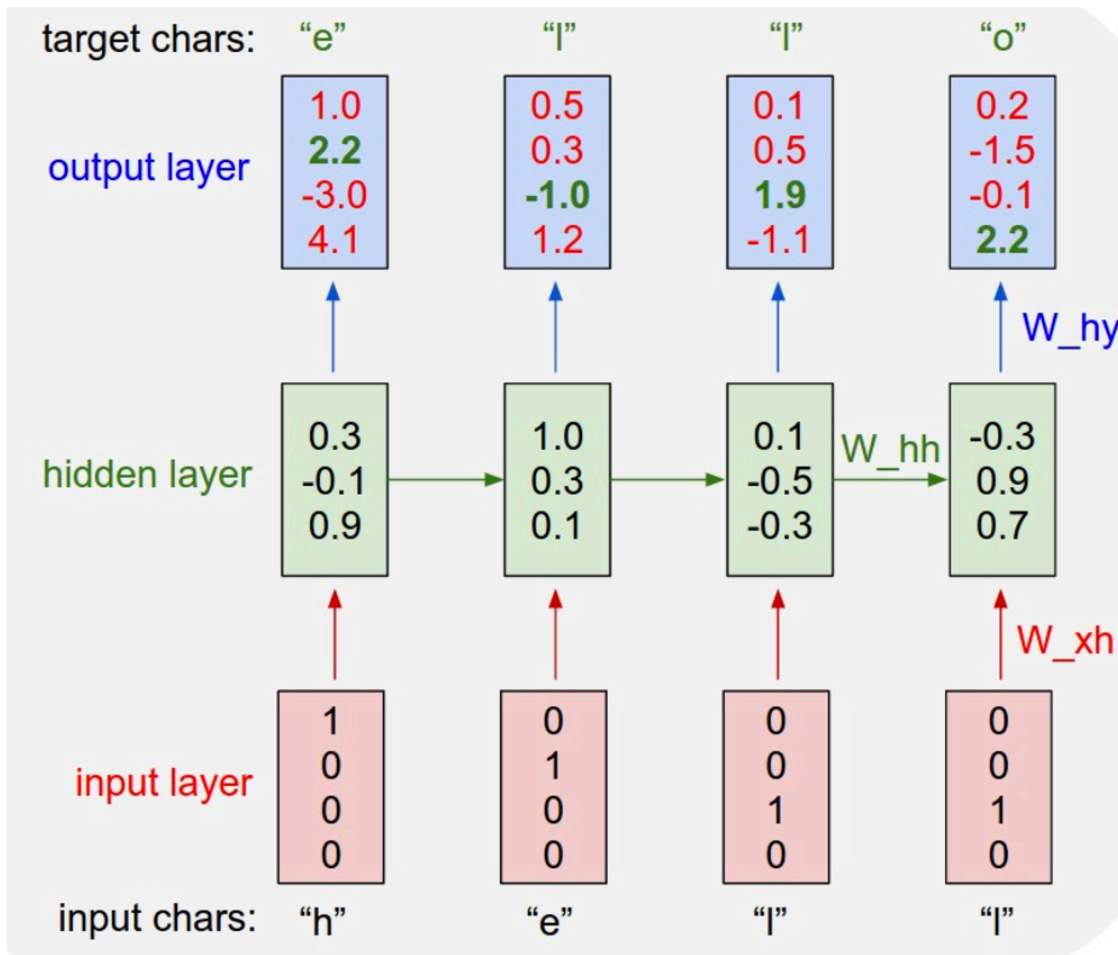
a) Original painting, b) Semantic annotation, c) Desired painting, d) Output

NLP: Syntax Parsing



SyntaxNet (Parsey McParseface) tags each word with a part-of-speech tag, and it determines the syntactic relationships between words in the sentence with an **94% accuracy** compared to a human performance at 96%.

Text Generation



Learn to spell words
using Recurrent
Neural Network

Text Generation

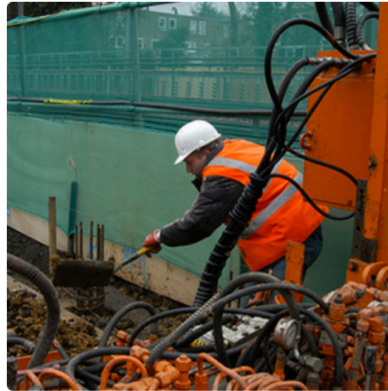
Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict. Copyright was the succession of independence in the slop of Syrian influence that was a famous German movement based on a more popular servicious, non-doctrinal and sexual power post. Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]] (PJS)[<http://www.humah.yahoo.com/guardian.cfm/7754800786d17551963s89.htm> Official economics Adjoint for the Nazism, Montgomery was swear to advance to the resources for those Socialism's rule, was starting to signing a major tripad of aid exile.]]

Trained on structured Wikipedia markdown. Network learns correct syntactic structure

Image Captioning



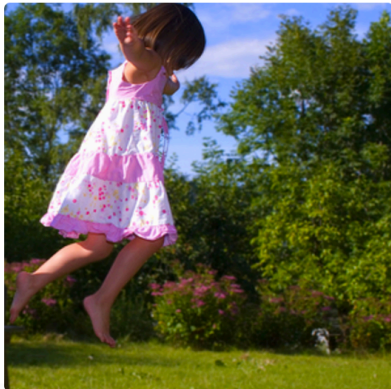
"man in black shirt is playing guitar."



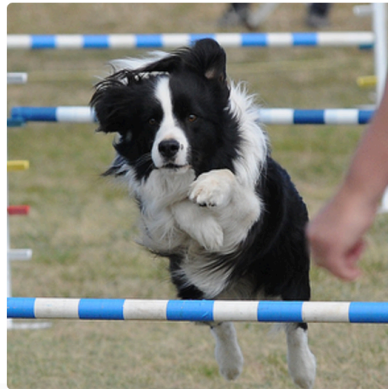
"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



"girl in pink dress is jumping in



"black and white dog jumps over



"young girl in pink shirt is

Deep Learning

- A multi-layer neural network
- It learns from experience
- It maps input to a continuous space representation
- It's effective in learning relationships and patterns

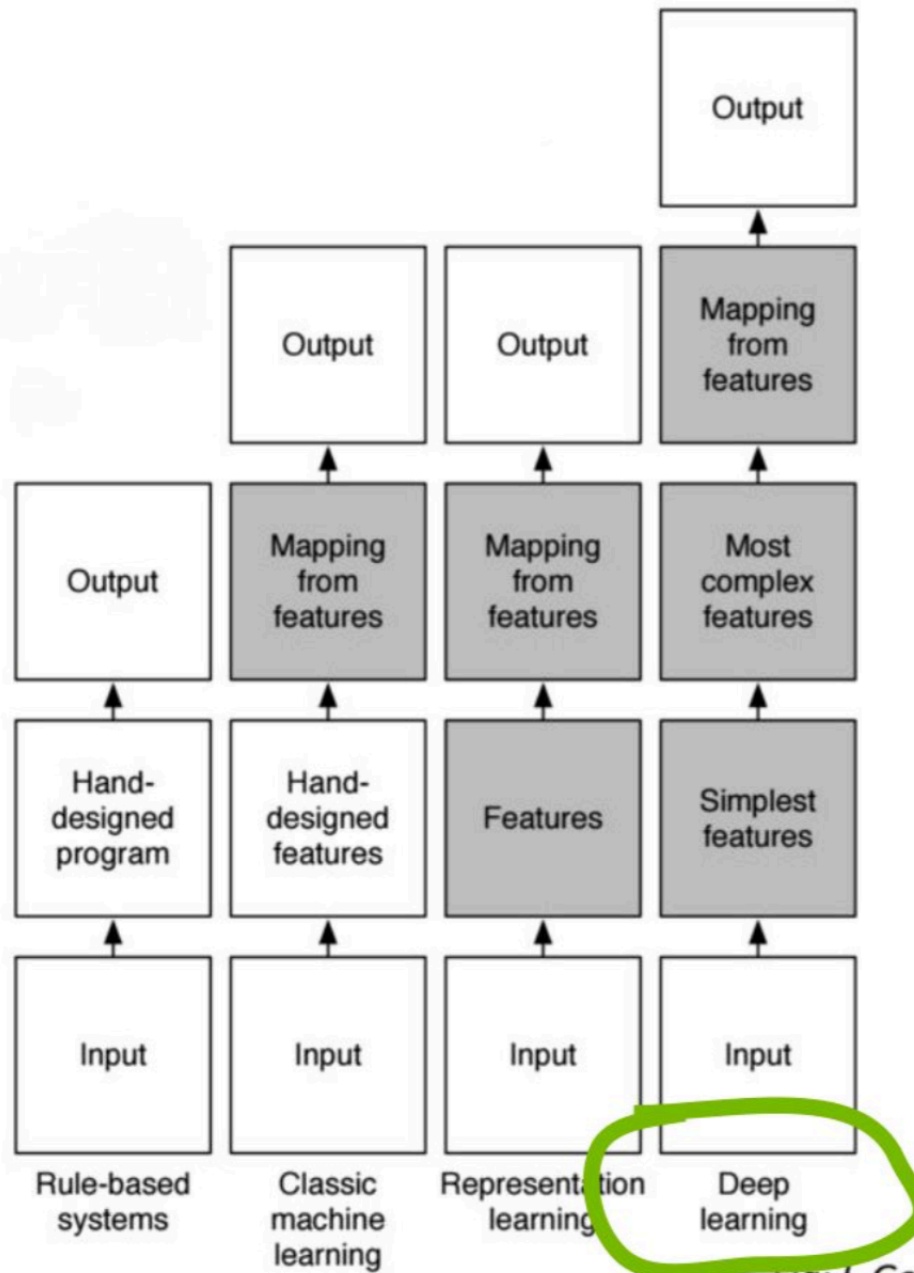


Fig: 1. Goodfellow

Brief History



1958 Perceptron

1974 Backpropagation



Convolution Neural Networks for Handwritten Recognition

1998

Google Brain Project on 16k Cores



2012

awkward silence (AI Winter)

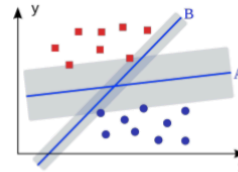
1969

Perceptron criticized



1995

SVM reigns



2006

Restricted Boltzmann Machine



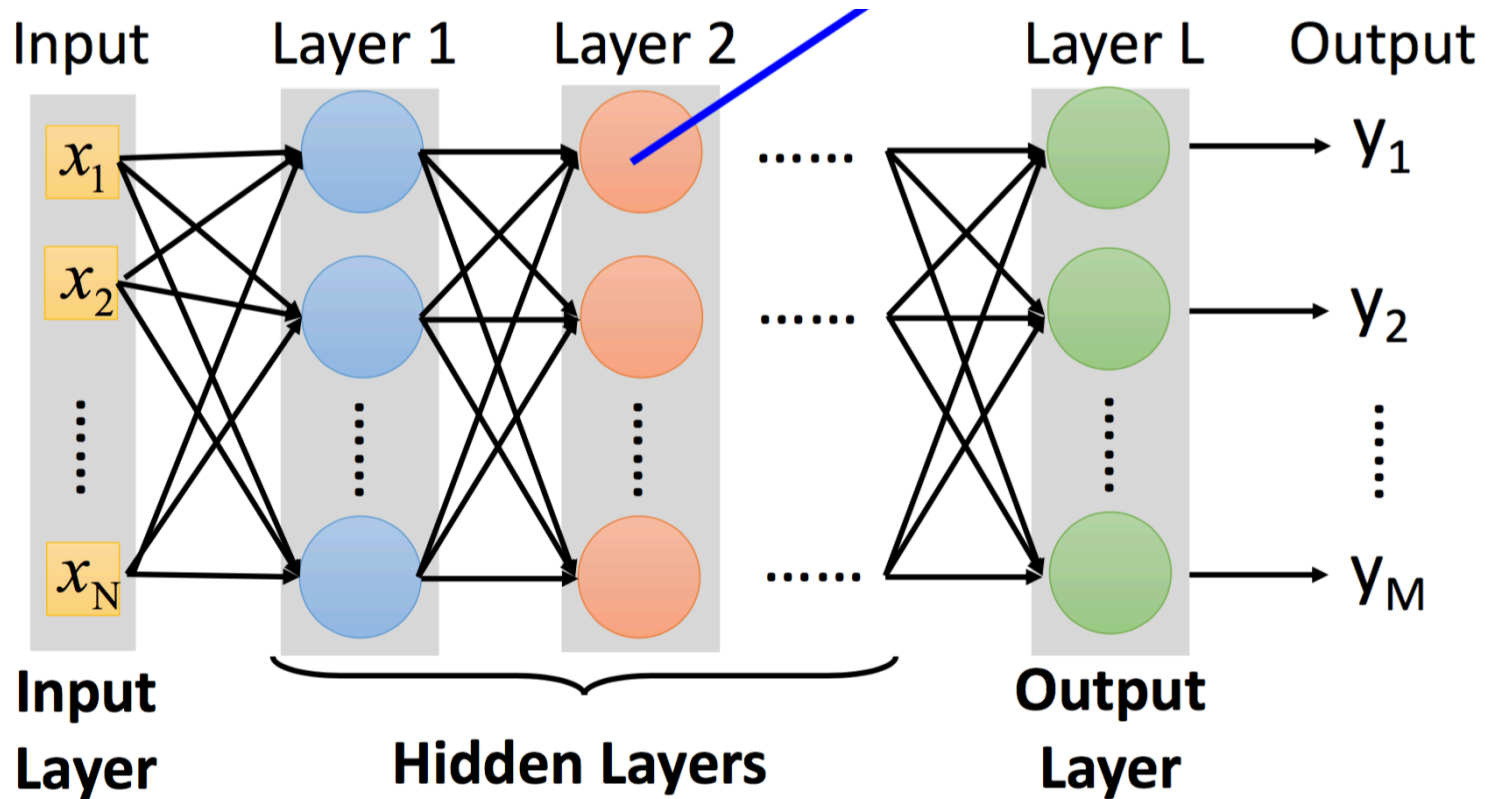
2012

AlexNet wins ImageNet
IMAGENET

What has Changed?

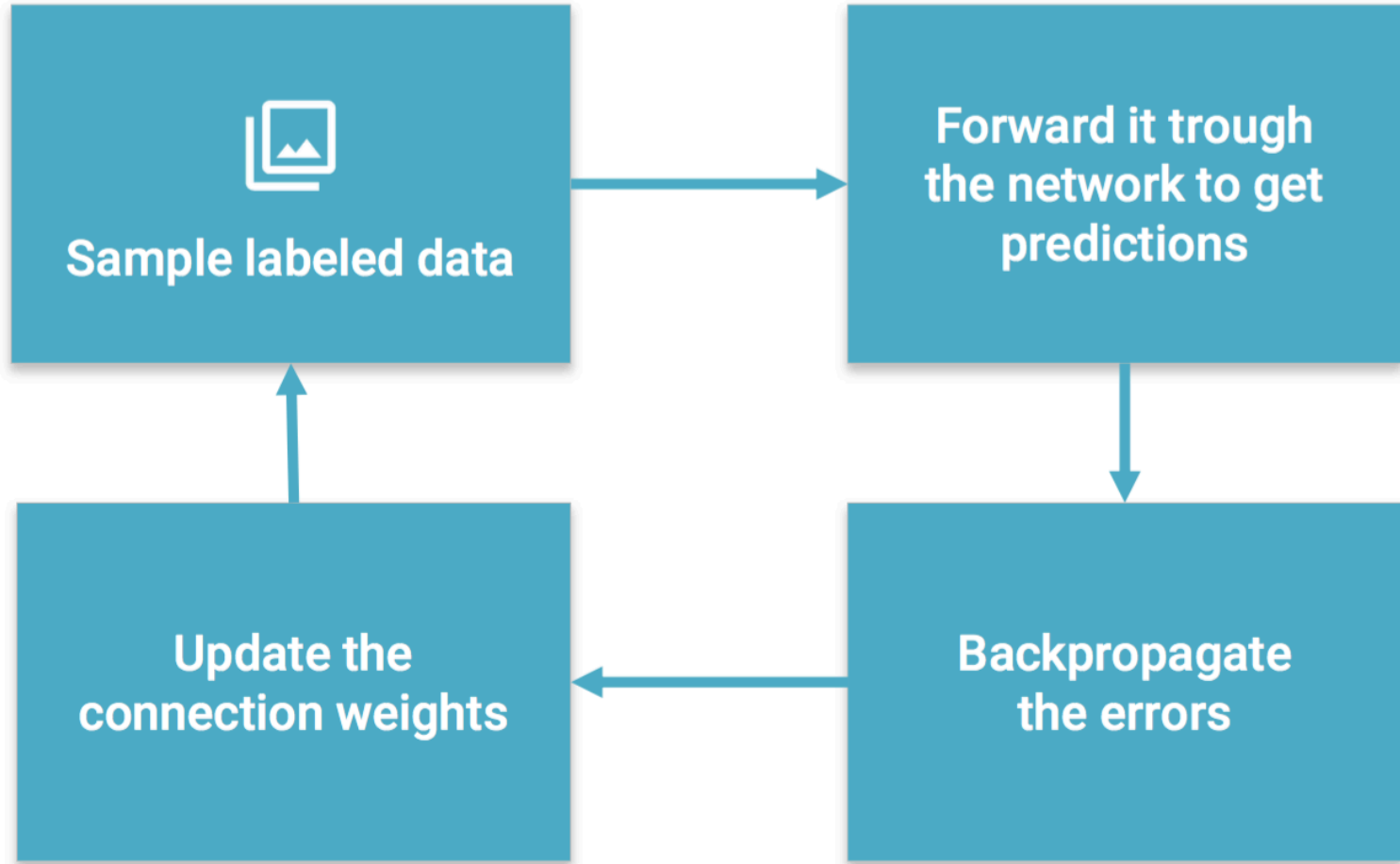
- Computational power
- Big Data

Fully Connected Neural Network



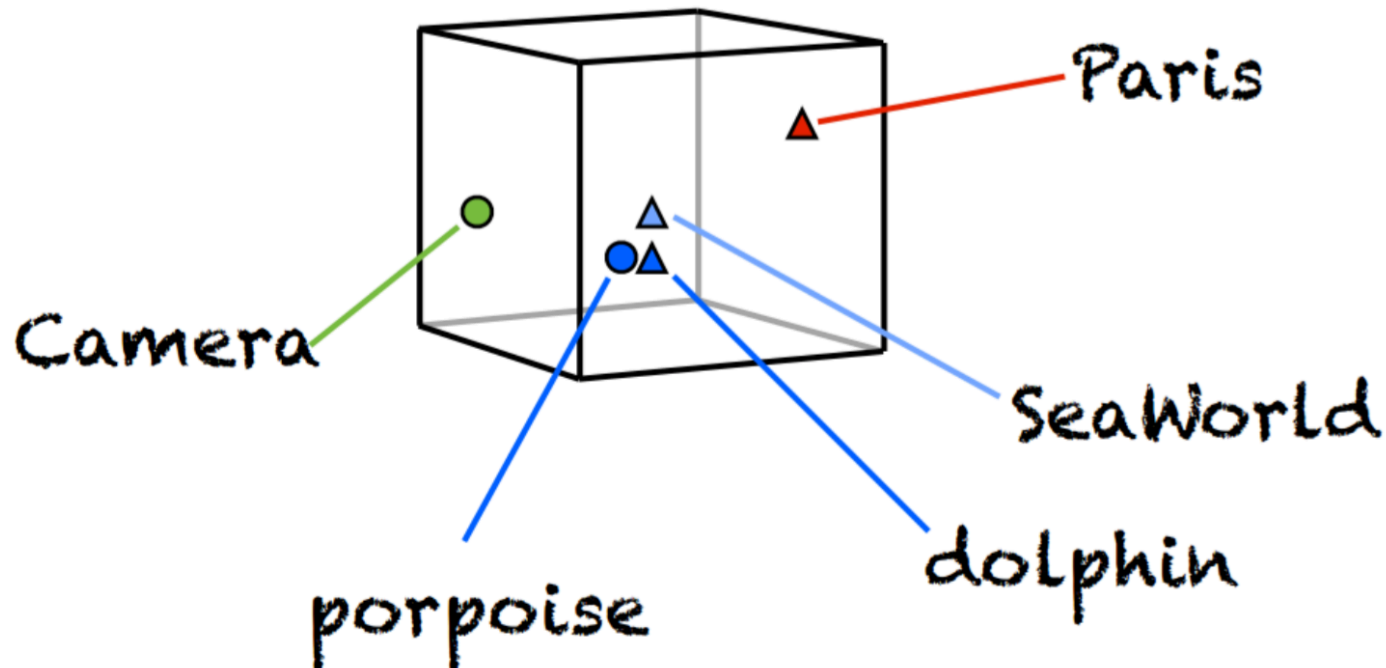
Deep means many hidden layers

Neural Network Training



Word Embedding

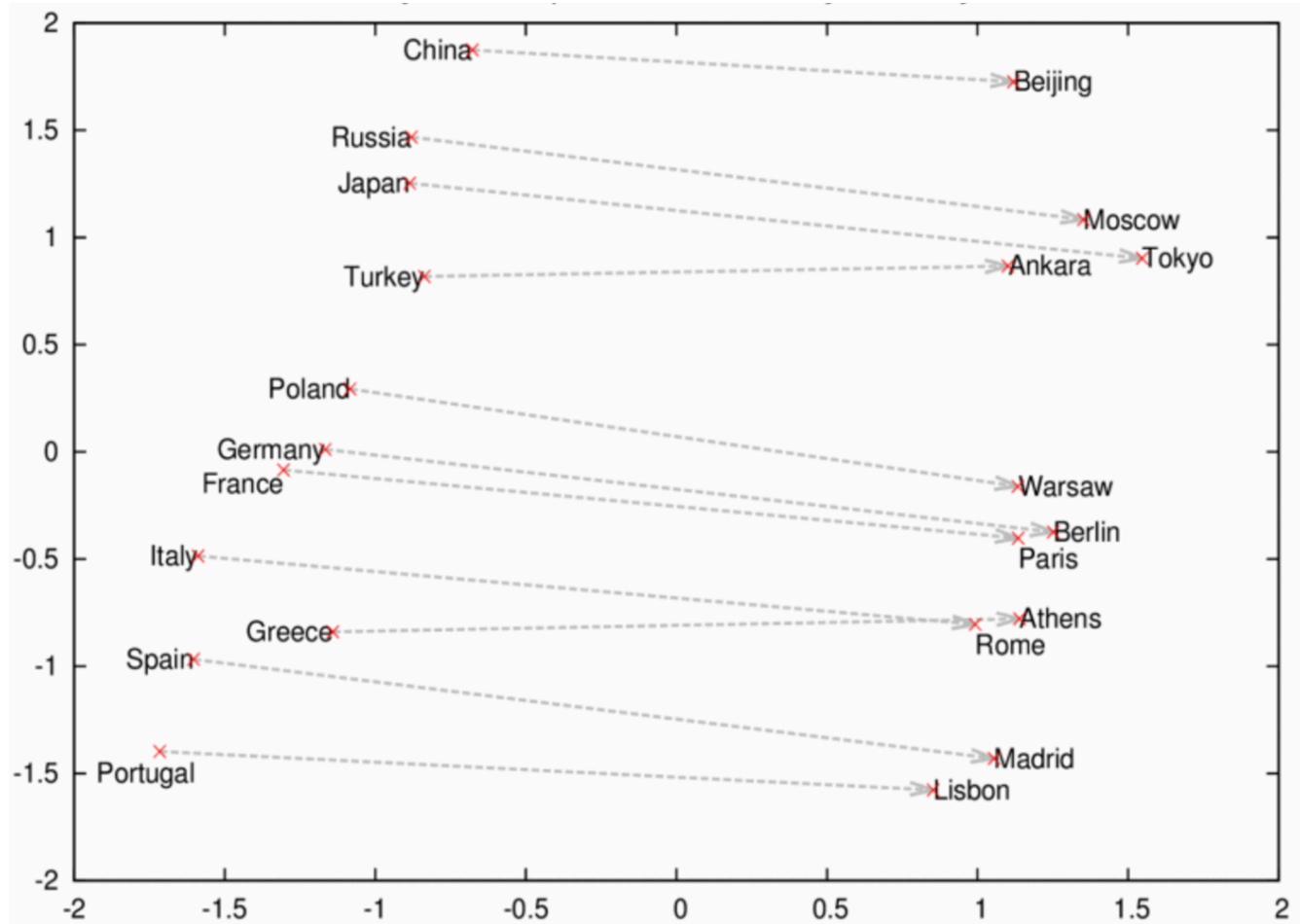
- Turn textual data into high dimensional vector representation
- Group semantically similar data in a vector space



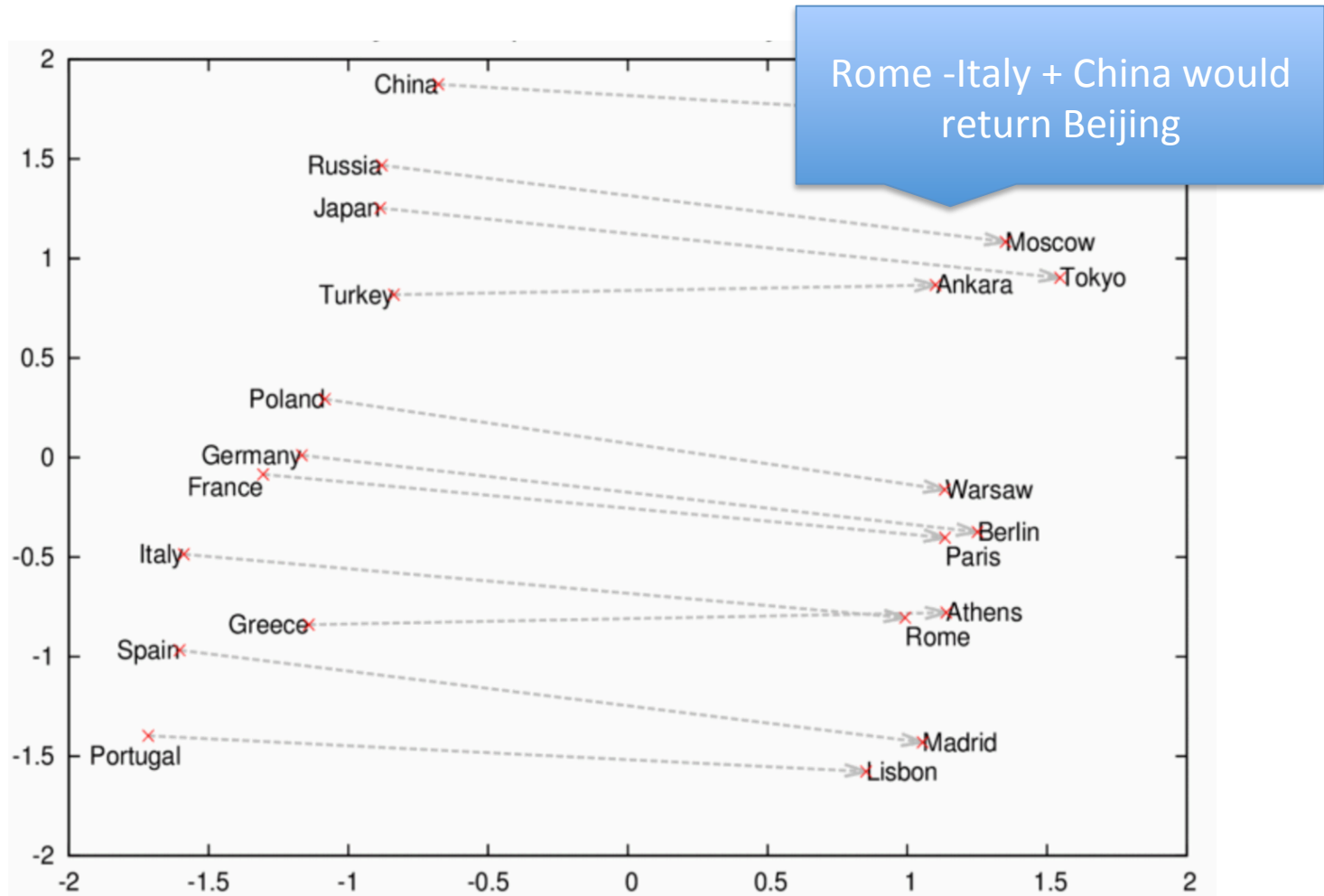
Word2vec

- A two-layer neural net
- Group vectors of similar words together into a vector space

Word2vec



Word2vec



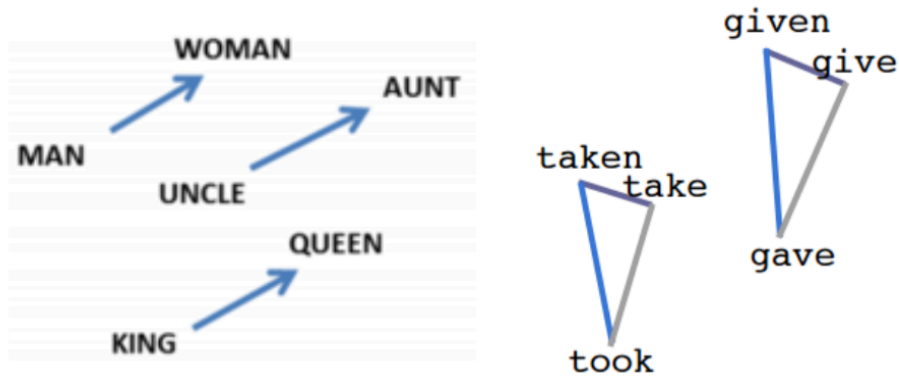
Word2vec

Semantically-related clusters

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

Word2vec

Learning relationships from data



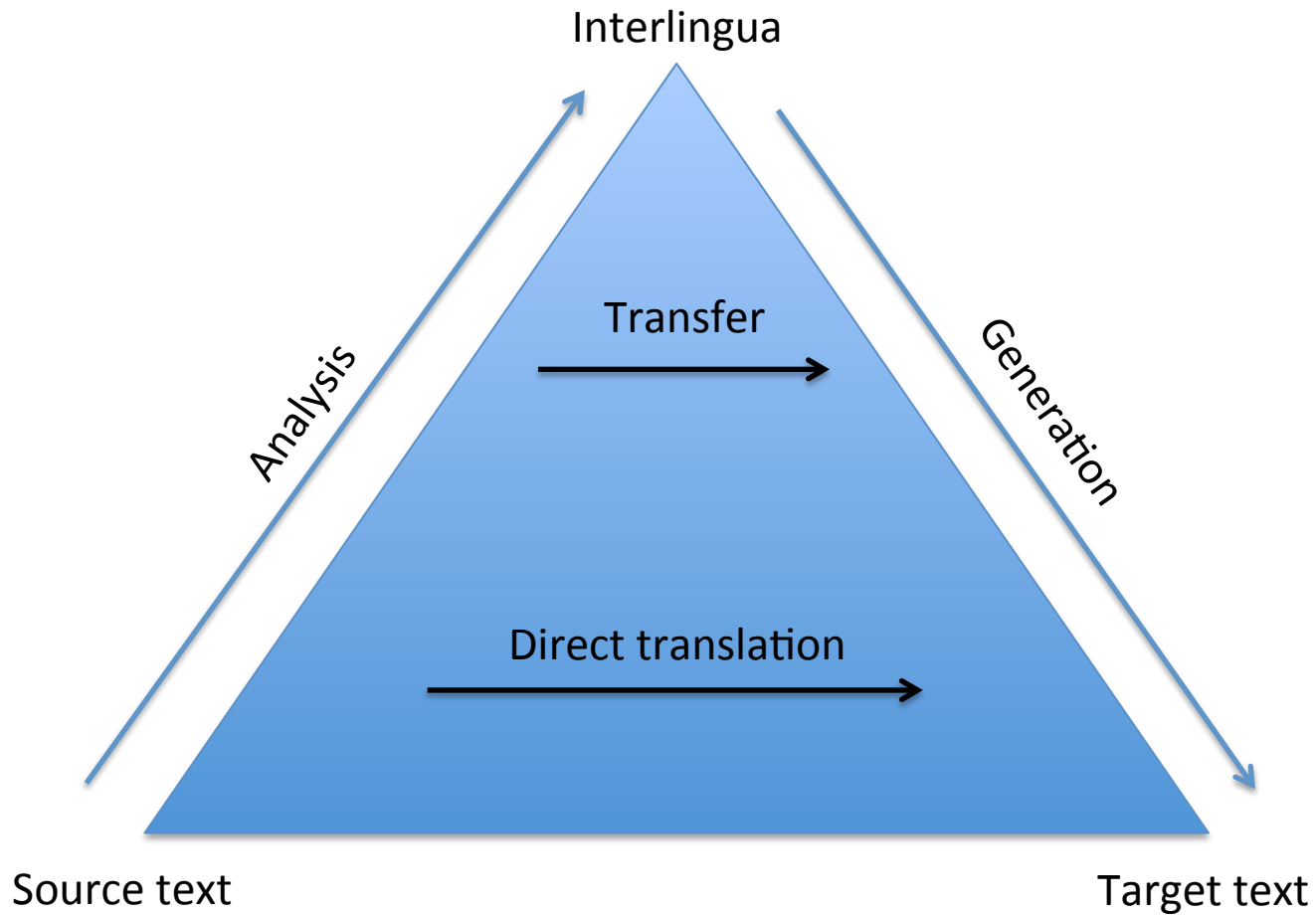
Woman - Man \approx Aunt - *Uncle*
King - Male + Female \approx *Queen*
Human - Animal \approx *Ethics*

Variants of Neural Networks

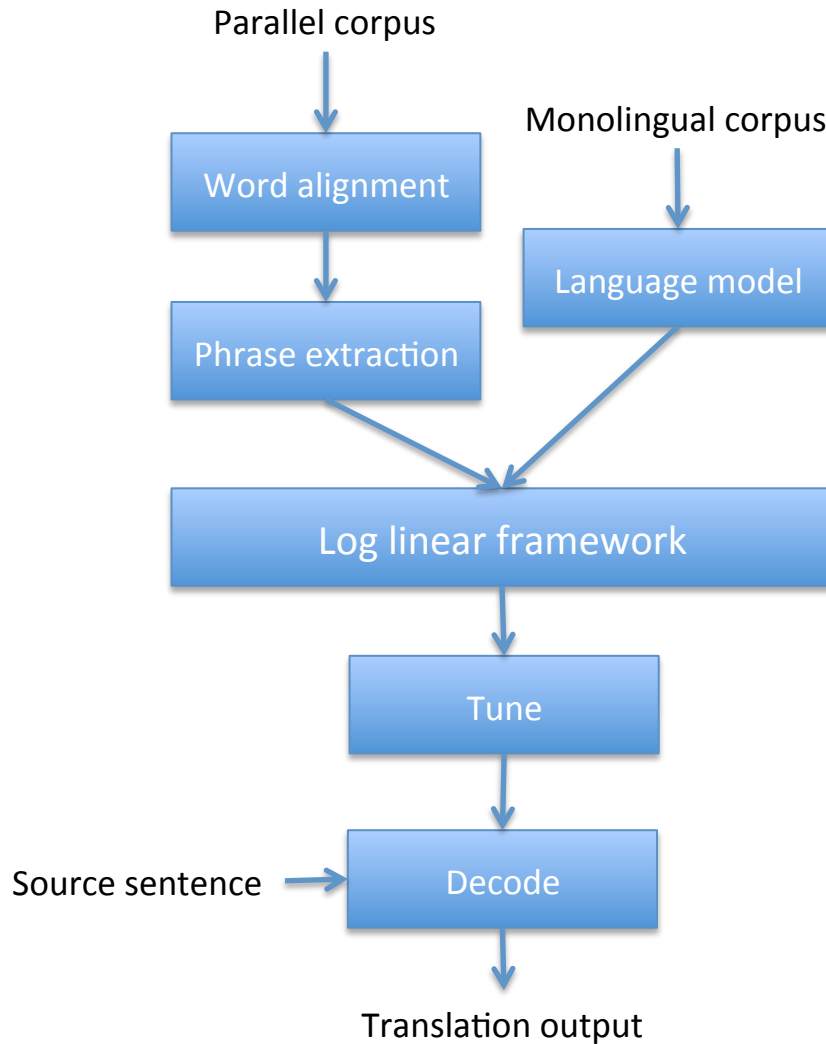
- Convolutional Neural Network (CNN)
 - mainly used in image processing
- Recurrent Neural Network (RNN)
 - sequence to sequence learning
 - language modeling

MACHINE TRANSLATION PARADIGM SHIFT

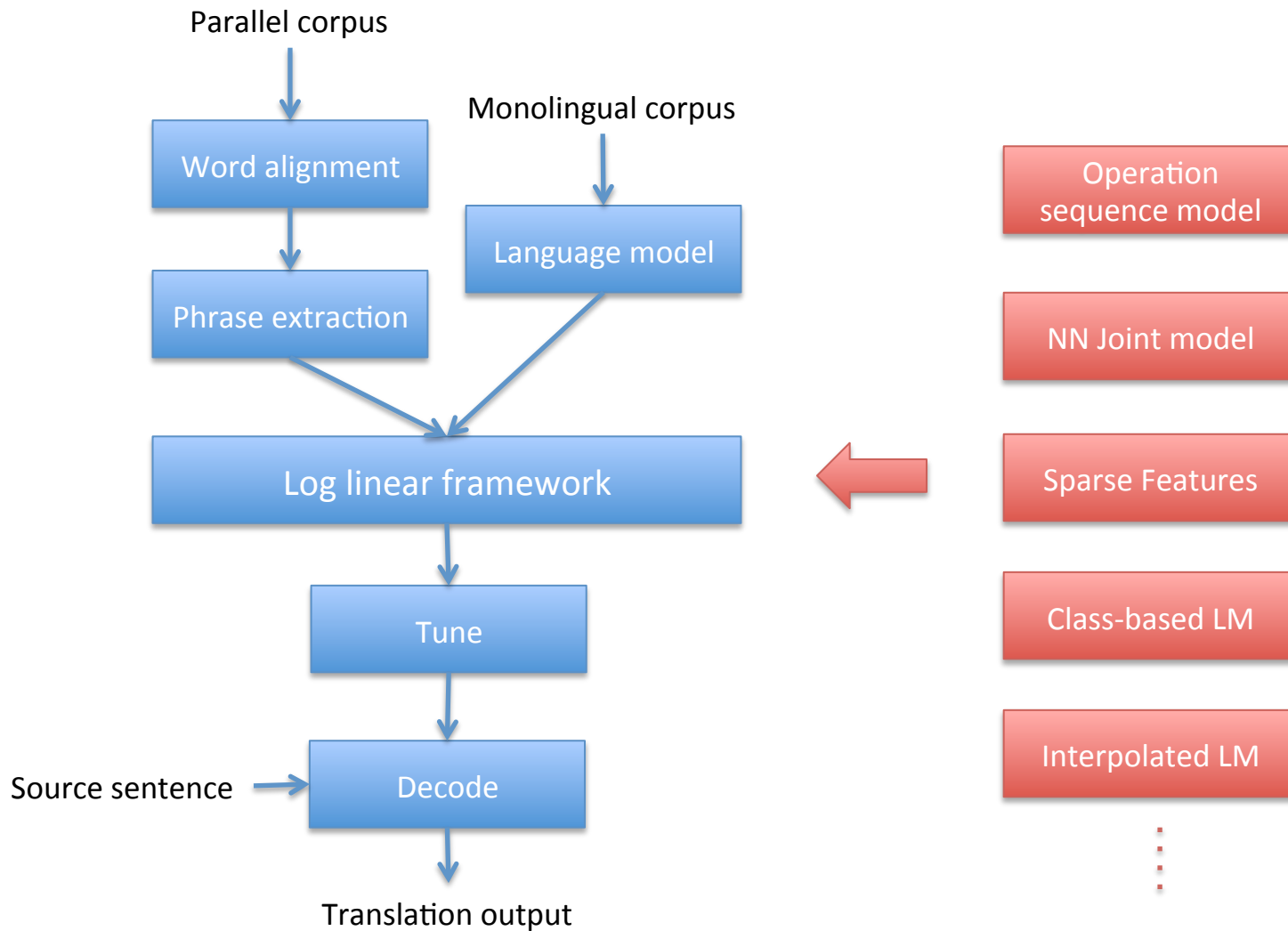
Machine Translation Pyramid



Statistical Machine Translation



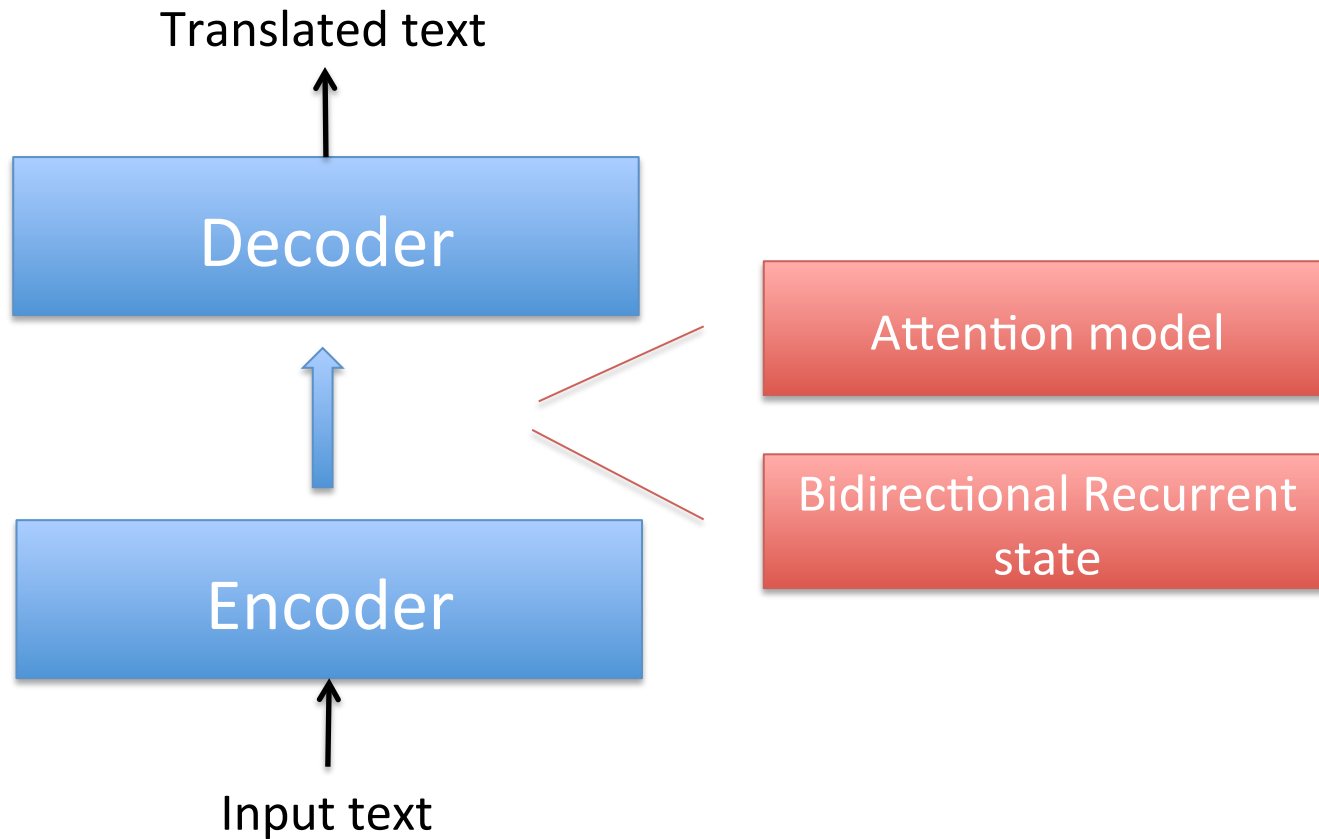
Statistical Machine Translation



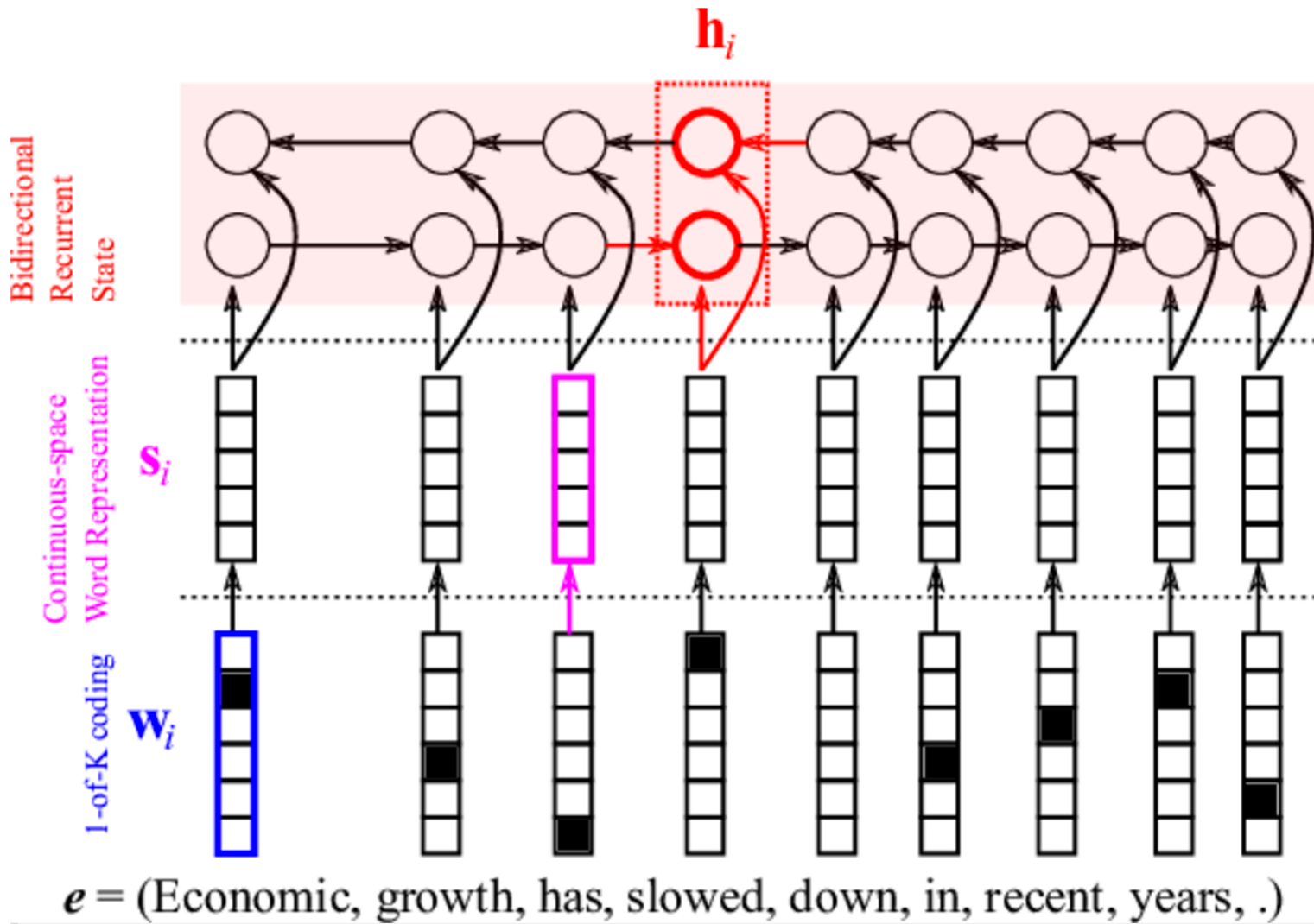
Other Variations

- Syntax-based machine translation
- Language dependent processing
 - pre-reordering of German
 - Arabic morphological segmentation
 - Urdu word segmentation

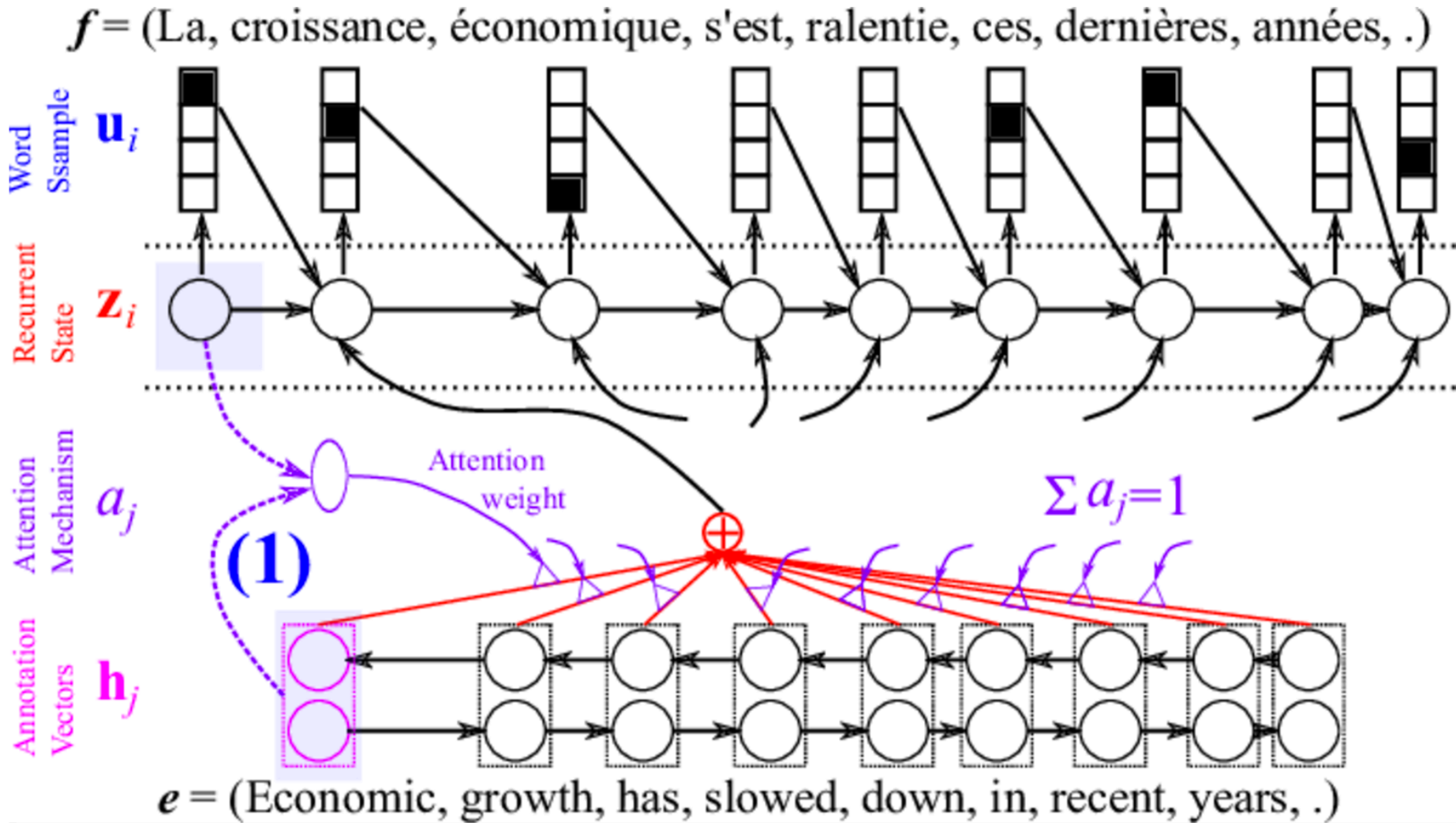
Neural Machine Translation



Encoder



Decoder



Task Overview

- Multi-domain parallel corpus

Parallel Corpus	Token (en)
TED (Cettolo et al. 2014)	4.7M
QED (Guzmán et al. 2013)	1.6M
UN (Ziems et al. 2016)	489M
OPUS (Lison et al. 2016)	184M

- Monolingual corpus
 - 287M sentences (WMT 2013)
 - target side of available parallel corpus

Task Overview

- Evaluation
 - avg. BLEU score on IWSLT test11-14

Results: Phrase-based MT

Train	Avg. BLEU	Description
TED (baseline)	28.6	
TED + QED + UN	27.3 (-1.3)	Concatenation
TED + Back-off PT(QED,UN)	29.1 (+0.5)	
TED + MML (QED,UN)	29.2 (+0.6)	
TED + MML (QED,UN) + OPUS	30.4 (+1.8)	
Interpolated LM (80GB)	30.9 (+2.3)	
Interpolated OSM	31.5 (+2.9)	
NNJM	32.1 (+3.5)	Train on concatenation
NNJM-UN	31.9 (+3.3)	Train on UN, fine tune on TED
NNJM-Opus	32.3 (+3.7)	Train on OPUS, fine tune on TED
Class-based LM	32.4 (+3.8)	
Drop-OOV	32.6 (+4.0)	
Transliteration	32.5 (+3.9)	

Neural MT

- Generalize well
- More data is better
 - be patient with training
- Online training
 - simple adaptation
 - commercially viable
- Small model size

NMT – Morphological Segmentation

- Can NMT resolve the bottleneck of language dependent pre-processing?
 - For Arabic, morphology aware segmentation
 - important to achieve good quality translations

Arabic Segmentation

- Alternates?
 - byte pair encoding
 - frequency-based model
 - iterative bottom-up character merging
 - single hyper-parameter to control vocabulary size
 - CNN-based word embedding
 - character-based model

Segmentation Results: Arabic-to-English

Segmentation Type	Avg. BLEU	Description
No segmentation	25.55	word-to-word training

Segmentation Results: Arabic-to-English

Segmentation Type	Avg. BLEU	Description
No segmentation	25.55	word-to-word training
Morphological segmentation	30.65	MADA segmented

Segmentation Results: Arabic-to-English

Segmentation Type	Avg. BLEU	Description
No segmentation	25.55	word-to-word training
Morphological segmentation	30.65	MADA segmented
BPE-to-word model	29.12	15000 merge operations
BPE-to-BPE model	29.41	15000 merge operations

Segmentation Results: Arabic-to-English

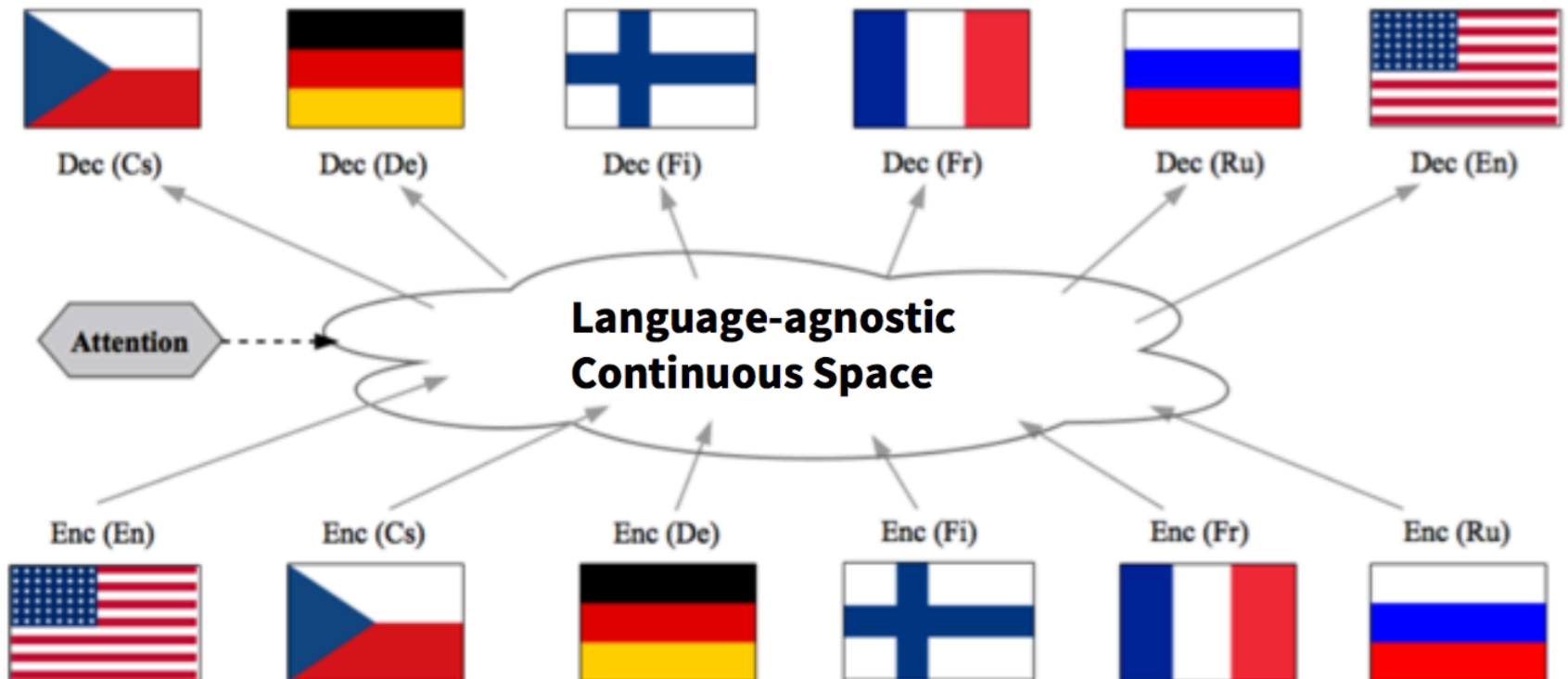
Segmentation Type	Avg. BLEU	Description
No segmentation	25.55	word-to-word training
Morphological segmentation	30.65	MADA segmented
BPE-to-word model	29.12	15000 merge operations
BPE-to-BPE model	29.41	15000 merge operations
charCNN-to-word model	29.74	source character CNN to target words
charCNN-to-char model	30.28	source character CNN to target characters
character-to-word model	30.41	

Training Time

No.	Segmentation Type	Avg. BLEU	Training Time/ epoch
1	No segmentation	25.55	2h12m
2	Morphological segmentation	30.65	2h47m
3	BPE-to-word model	29.12	2h12m
4	BPE-to-BPE model	29.41	1h16m
5	charCNN-to-word model	29.74	2h48m
6	charCNN-to-char model	30.28	3h54m
7	character-to-word model	30.41	4h53m

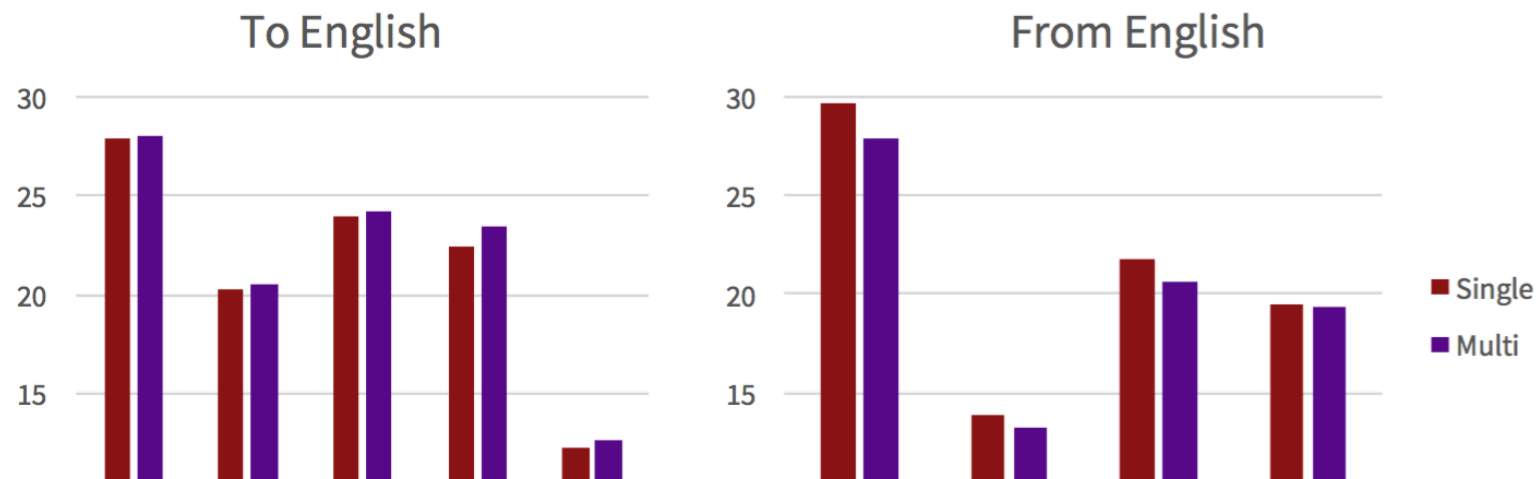
Morphological segmentation \approx charCNN-to-word
but losing in BLEU by 0.9 points

Multi-lingual Translation






Multi-lingual Translation

- 10 language pair-directions
 - $\text{En} \rightarrow \{\text{Fr, Cs, De, Ru, Fi}\} + \{\text{Fr, Cs, De, Ru, Fi}\} \rightarrow \text{En}$
- 60+ million bilingual sentence pairs
- *Comparable to 10 single-pair models*



Low Resource Translation

- Low-resource translation
 - Positive language transfer from high-resource to low-resource language pair-directions

		# Symbols		# Sentence		
		# En	Other	Train	Dev	Test
	En-Uz	1.361m	1.186m	73.66k	948	882
	En-Es	908.1m	924.9m	34.71m	3003	3000
	En-Fr	1.837b	1.911b	65.77m	3003	3000

Low Resource Translation

Uz-En: 6.45

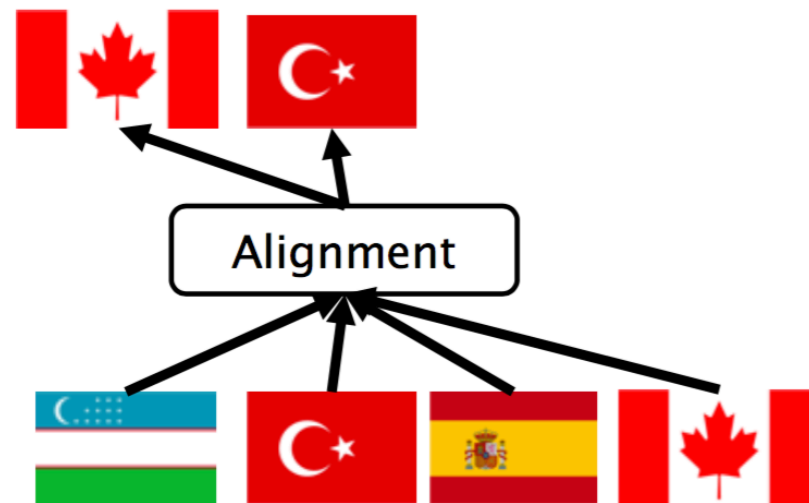
Uz-En + Tr-En: 9.34

Uz-En + Tr-En + Es-En: 10.34

Uz-En + Tr-En + Es-En + En-Tr: 9.41

Ensemble: 12.99

- 3x Uz-En + Tr-En + Es-En
- 3x Uz-En + Tr-En + Es-En + En-Tr



Short Comings

- Problem of interpretability
 - not easy to explain what is happening
- Hard to incorporate linguistic information
- Fewer control over the model
- Special hardware requirement (GPU)

Useful links

- NMT implementations
 - Nematus (<https://github.com/rsennrich/nematus>)
 - Theano-based
 - backed by Edinburgh University
 - seq2seq-attn (<https://github.com/harvardnlp/seq2seq-attn>)
 - Torch-based
 - backed by Harvard plus Systran is supporting
 - Dynet (<https://github.com/clab/dynet>)
 - C++
 - backed by CMU
 - Tensor flow (sequence to sequence)

Useful Links

- [NMT Tutorial - https://devblogs.nvidia.com/paralleforall/introduction-neural-machine-translation-gpus-part-1/](https://devblogs.nvidia.com/paralleforall/introduction-neural-machine-translation-gpus-part-1/)
- RNN character-level examples
 - <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

THANKS