Human Language Technology:
Application to Information Access

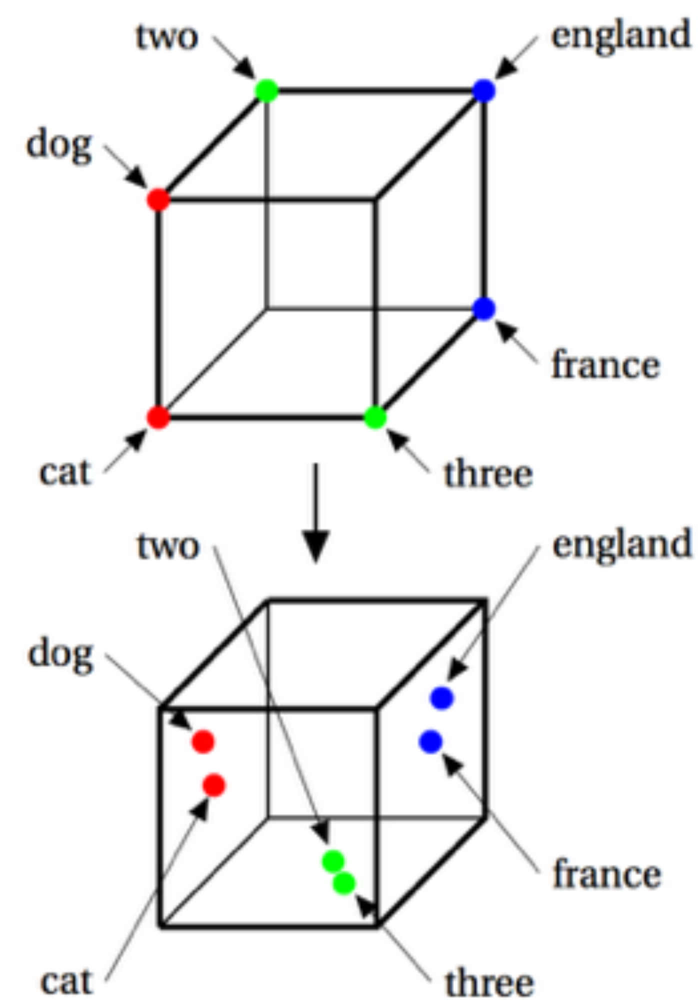
Lesson 10
Deep learning for NLP:
Multilingual Word Sequence Modeling

December 15, 2016

EPFL Doctoral Course EE-724
Nikolaos Pappas
Idiap Research Institute, Martigny

Outline of the talk

1. Recap: Word Representation Learning
2. Multilingual Word Representations
   - Alignment models
   - Evaluation tasks
3. Multilingual Word Sequence Modeling
   - Essentials: RNN, LSTM, GRU
   - Machine Translation
   - Document Classification
4. Summary



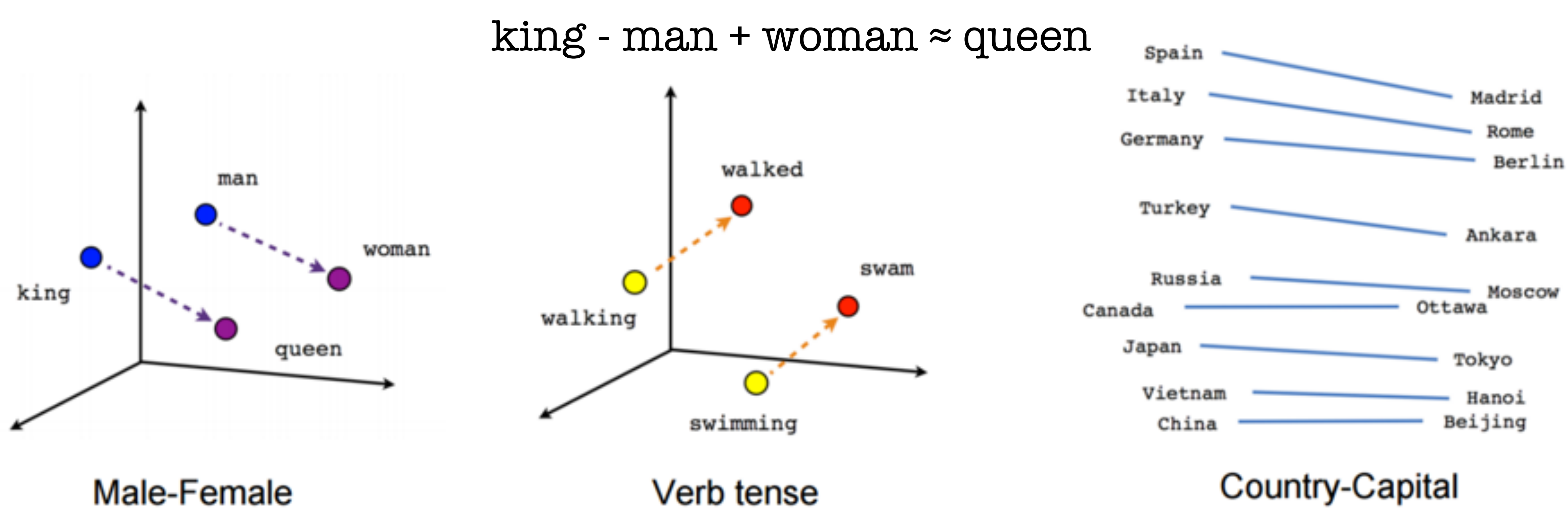
* Figure from Lebret's thesis, EPFL, 2016

Disclaimer

- Research highlights rather than in-depth analysis
  - By no means exhaustive (progress too fast!)
  - Tried to keep most representatives
- Focus on feature learning and two major NLP tasks
- Not enough time to cover other exciting tasks:
  - Question answering
  - Relation classification
  - Paraphrase detection
  - Summarization

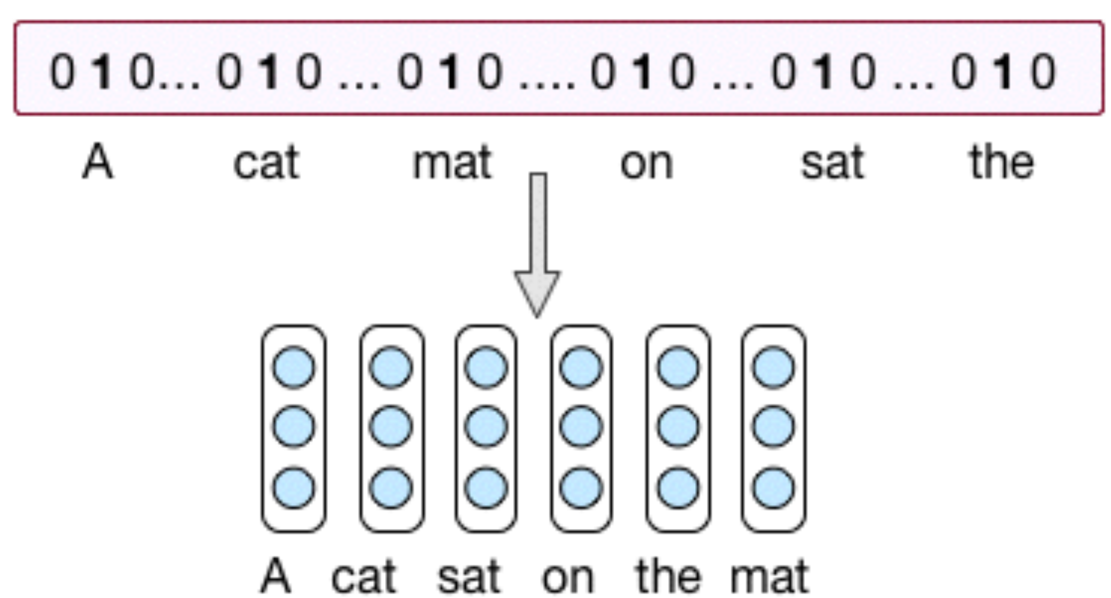
Recap: Learning word representations from text

- **Why should we care about them?**
  - tackles curse of dimensionality
  - captures semantic and analogy relations of words
  - captures general knowledge in an unsupervised way



Recap: Learning word representations from text

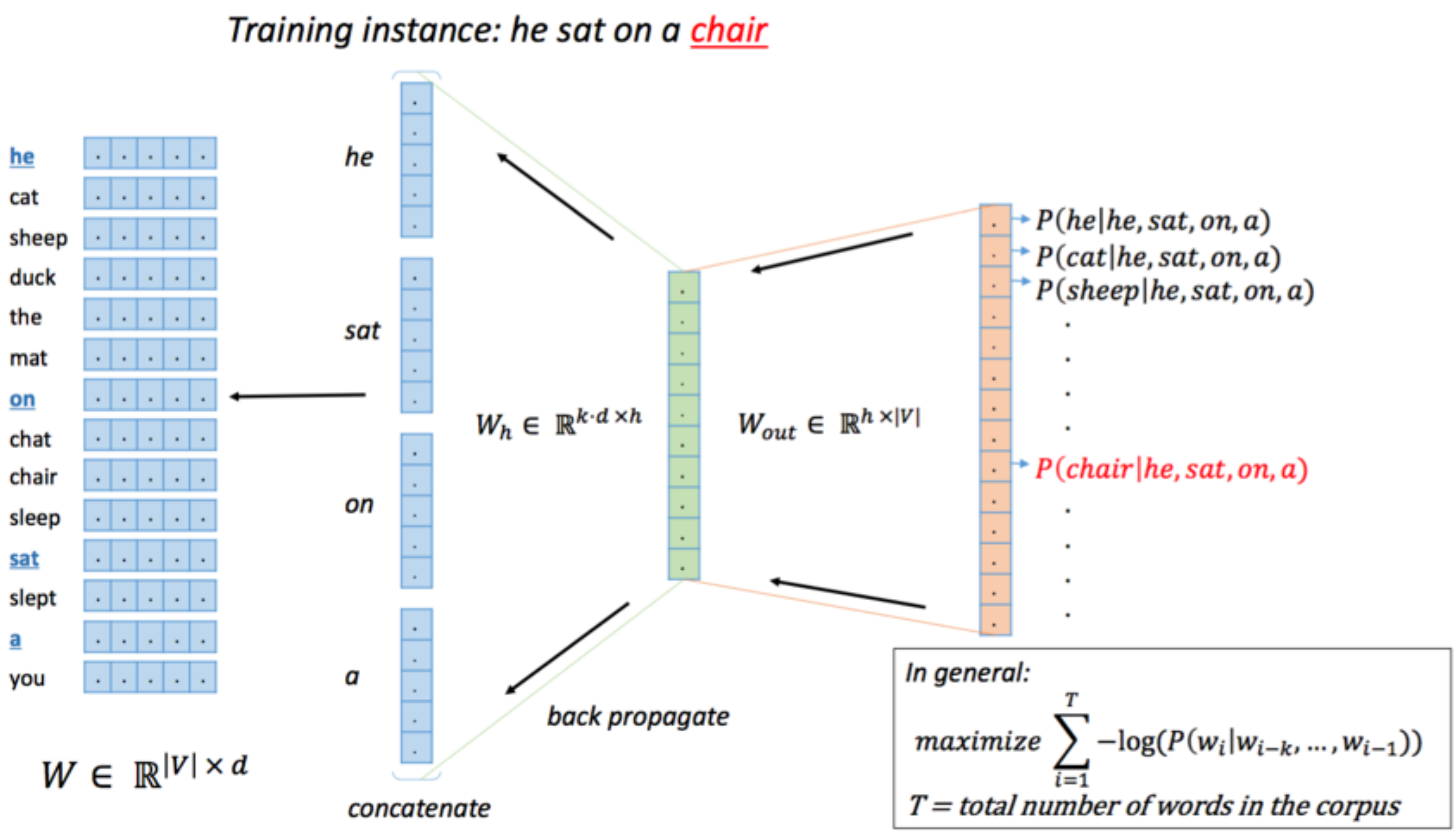
- **How can we benefit from them?**
  - study linguistic properties of words
  - inject general knowledge on downstream tasks
  - transfer knowledge across languages or modalities
  - compose representations of word sequences



Recap: Learning word representations from text

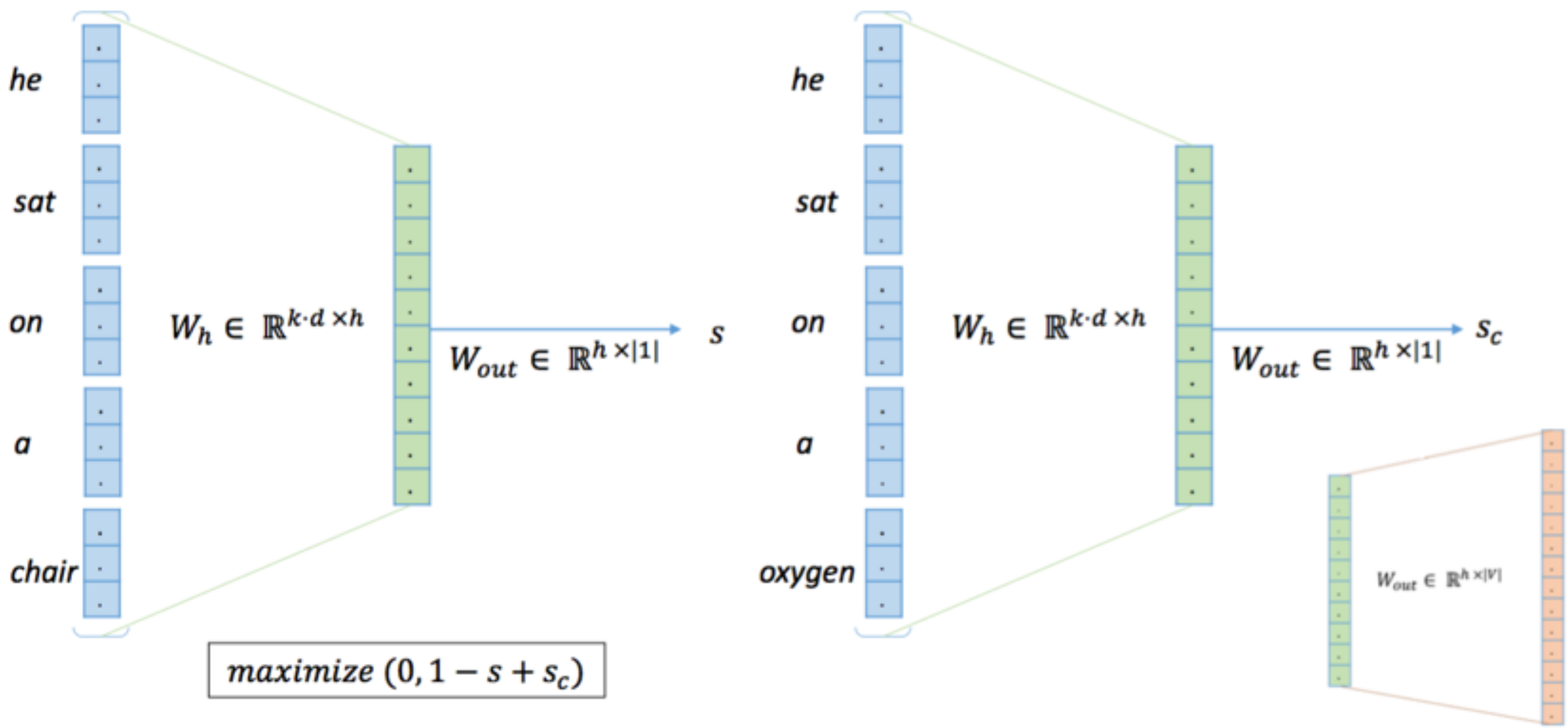
- **Which method to use for learning them?**
  - neural versus count-based methods
    - ➡ neural ones implicitly do SVD over a PMI matrix
    - ➡ similar to count-based when using the same tricks
  - neural methods appear to have the edge (word2vec)
    - ➡ efficient and scalable objective + toolkit
    - ➡ intuitive formulation (=predict words in context)

Recap: Continuous Bag-of-Words (CBOW)



Recap: Continuous Bag-of-Words (CBOW)

Positive: he sat on a chair — *Negative: he sat on a oxygen*

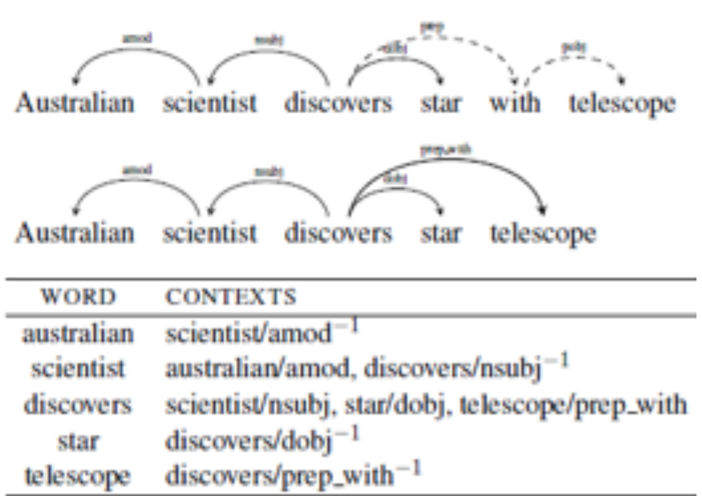


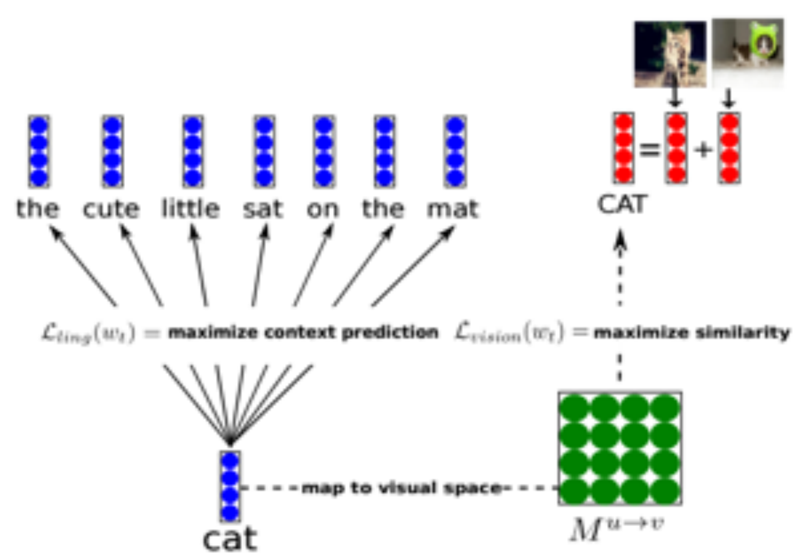
back propagate and update word representations

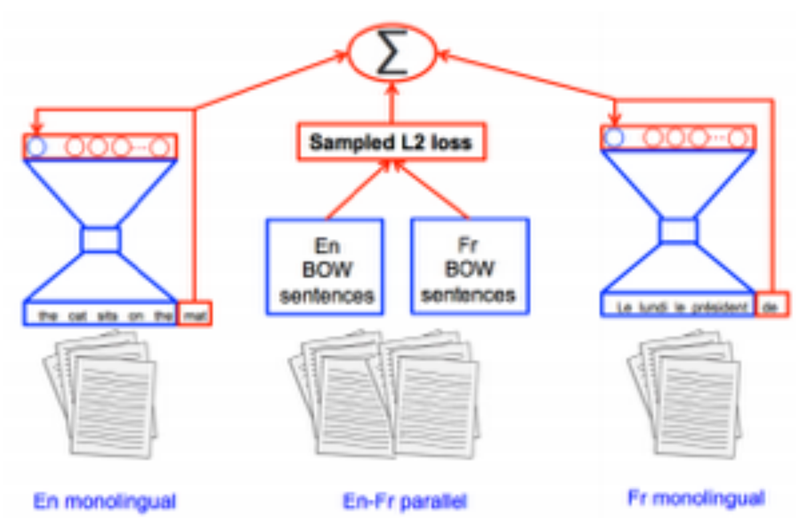
Advantage: does not require this expensive matrix multiplication

Recap: Learning word representations from text

- **What else can we do with word embeddings?**
  - dependency-based embeddings: Levy and Goldberg 2014
  - retrofitted-to-lexicons embeddings: Faruqui et al. 2014
  - sense-aware embeddings: Li and Jurafsky 2015
  - visually-grounded embeddings: Lazaridou et al. 2015
  - multilingual embeddings: Gouws et al 2015







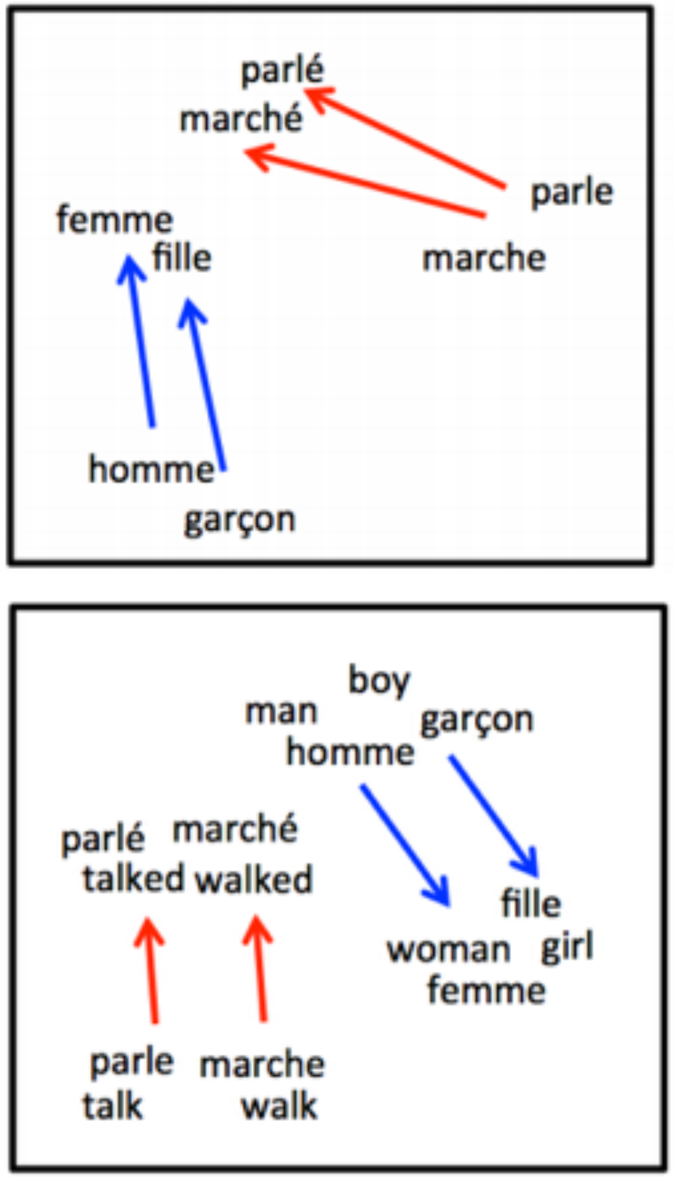
Outline of the talk

1. Recap: Word Representation Learning

2. Multilingual Word Representations
   - Alignment models
   - Evaluation tasks

3. Multilingual Word Sequence Modeling
   - Essentials: RNN, LSTM, GRU
   - Machine Translation
   - Document Classification

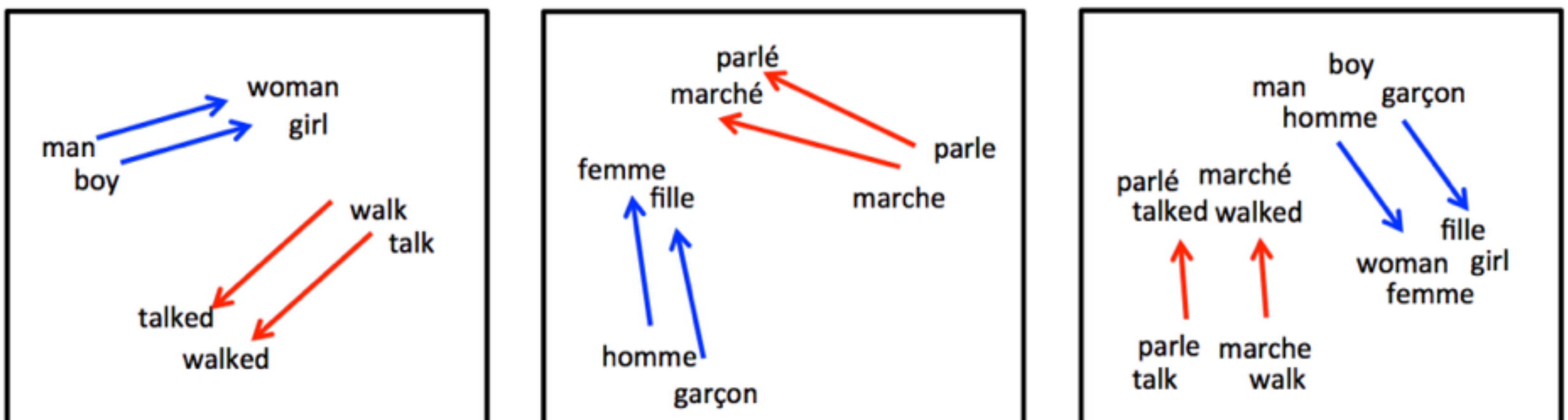
4. Summary



* Figure from Gouts et al., 2015.

Learning cross-lingual word representations

- Monolingual embeddings capture semantic, syntactic and analogy relations between words
- **Goal**: capture this relationships two or more languages



* Figure from Gouts et al., 2015.

Supervision of cross-lingual alignment methods

- **Parallel sentences for MT:** **Guo et al., 2015**
  Sentence by sentence and word alignments
- **Parallel sentences:** **Gouws et al., 2015**
  Sentence by sentence alignments
- **Parallel documents:** **Søgaard et al., 2015**
  Documents with topic or label alignments
- **Bilingual dictionary:** **Ammar et al., 2016**
  Word by word translations
- **No parallel data:** **Faruqui and Dyer, 2014**
  Really!

high

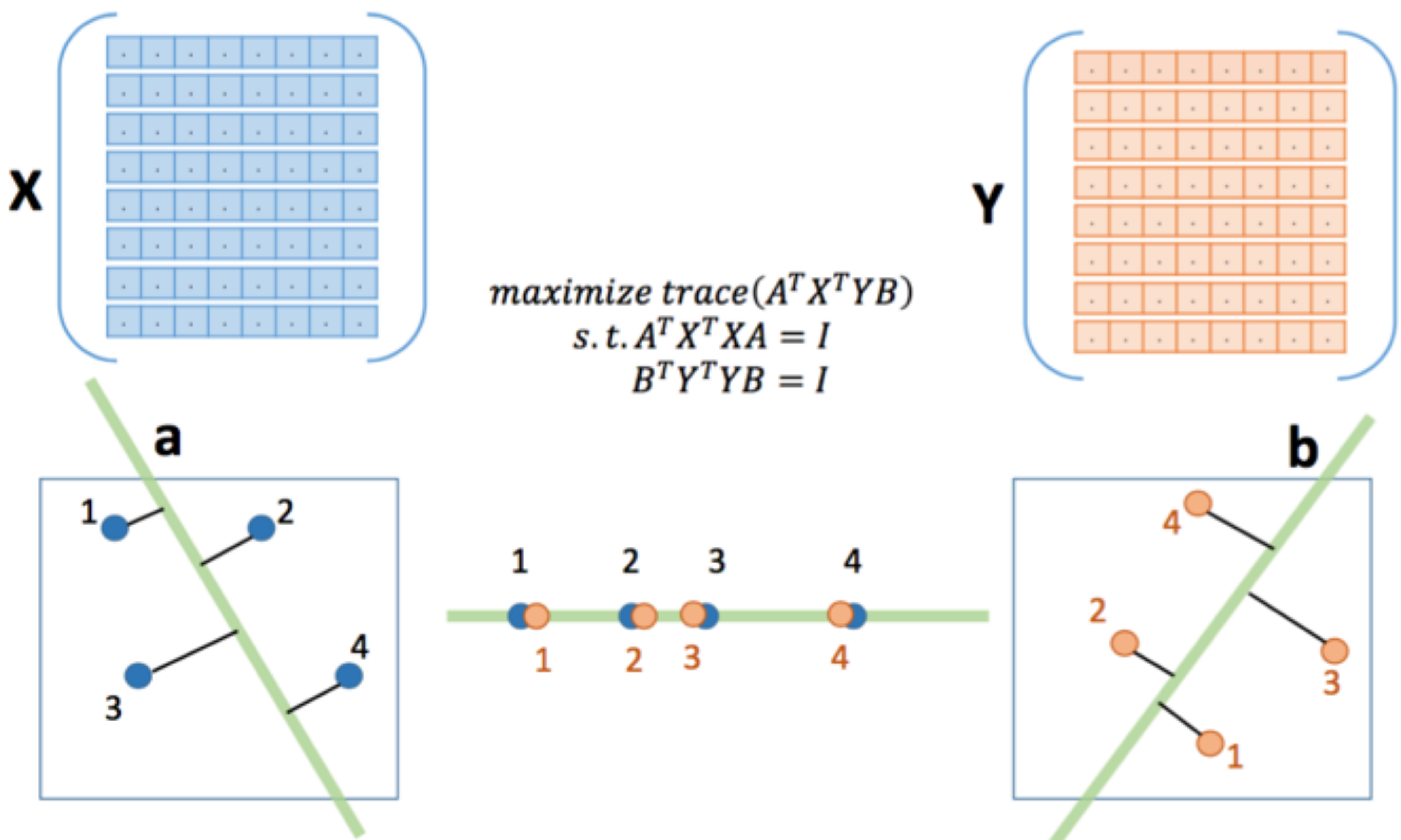
Annotation cost

low

Cross-lingual alignment with no parallel data

After Stage 1: *X = representations of English words*
Y = representations of French words

Goal *: transform X and Y such that the transformed representations of (cat, chat), (you, toi), etc. are close to each other*



$$\text{maximize } trace(A^T X^T Y B)$$
$$s.t.\ A^T X^T X A = I$$
$$B^T Y^T Y B = I$$

(Faruqui and Dyer, 2014)

82

Cross-lingual alignment with parallel sentences

Training data: Parallel sentences

$a = English\ sentence$
$b = parallel\ French\ sentence$
$n = random\ French\ sentence$

minimize
$E(a,b) = ||f(a) - g(b)||^2$

minimize
$max(0, m + E(a,b) - E(a,n))$

degenerate solution is to make $f(a) = g(b) = 0$

To avoid this use max-margin training

Backpropagate & update w_i's in both languages

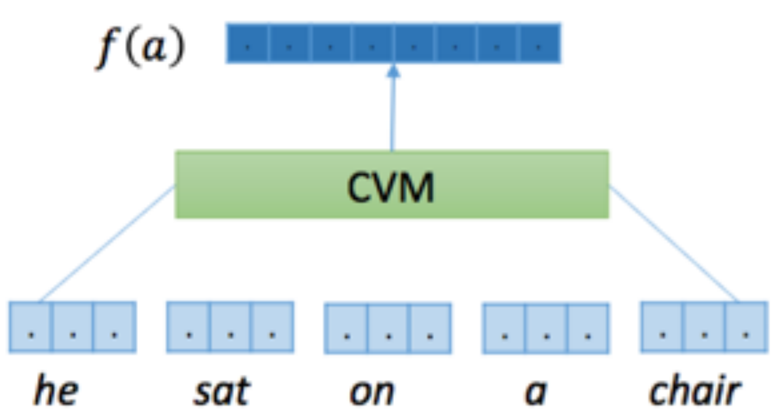
Compose word representations to get a sentence representation using a Compositional Vector Model(CVM)

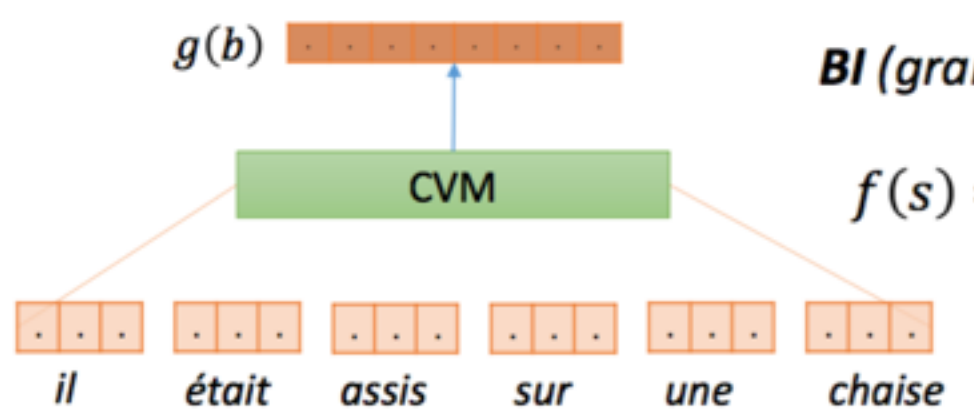
Two options considered:
***ADD**: (simply add word vectors)*
$s = sentence$
$w_i = representation\ of\ word\ i\ in\ the\ sentence$

$$f(s) = \sum_{i=1}^{n} w_i$$

***BI** (gram):*

$$f(s) = \sum_{i=1}^{n} \tanh(w_{i-1} + w_i)$$

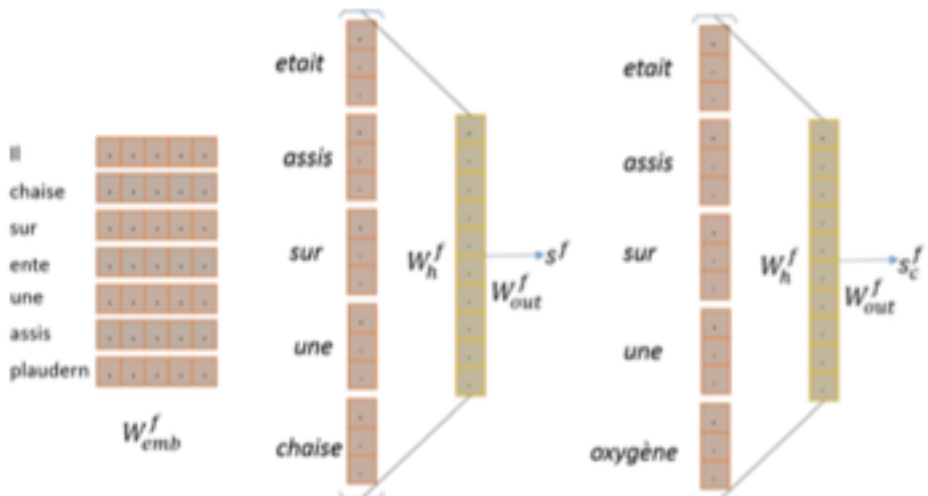




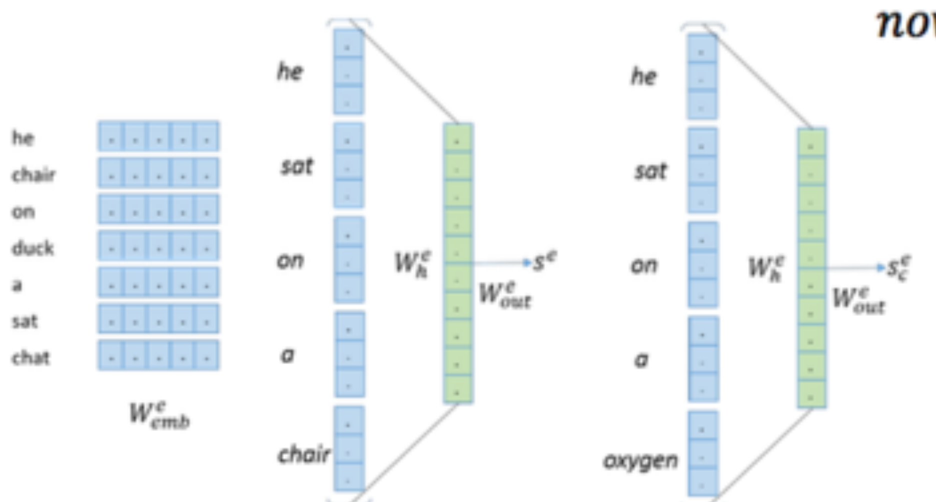
(Hermann & Blunson, 2014)

Cross-lingual alignment with parallel sentences

Fr positive: Il était assis sur une chaise
Fr negative: Il était assis sur une oxygène



En positive: he sat on a chair
En negative: he sat on a oxygen



Independently update θ^e and θ^f

$$\boxed{\begin{array}{l} maximize\ max(0, 1 - s^f + s^f_c) \\ w.r.t.\ \theta^e \end{array}}$$

+ Parallel data
En: he sat on a chair $[s_e = w^e_1, w^e_2, w^e_3, w^e_4, w^e_5]$
Fr : Il était assis sur une chaise $[s_f = w^f_1, w^f_2, w^f_3, w^f_4, w^f_5]$

$$now, also\ minimize \quad \Omega\left(W^e_{emb}, W^f_{emb}\right) = \left\| \frac{1}{m} \sum_{w_i \in s^e}^{w_m} W^e_{emb_i} - \frac{1}{n} \sum_{w_j \in s^e}^{w_n} W^f_{emb_i} \right\|^2$$

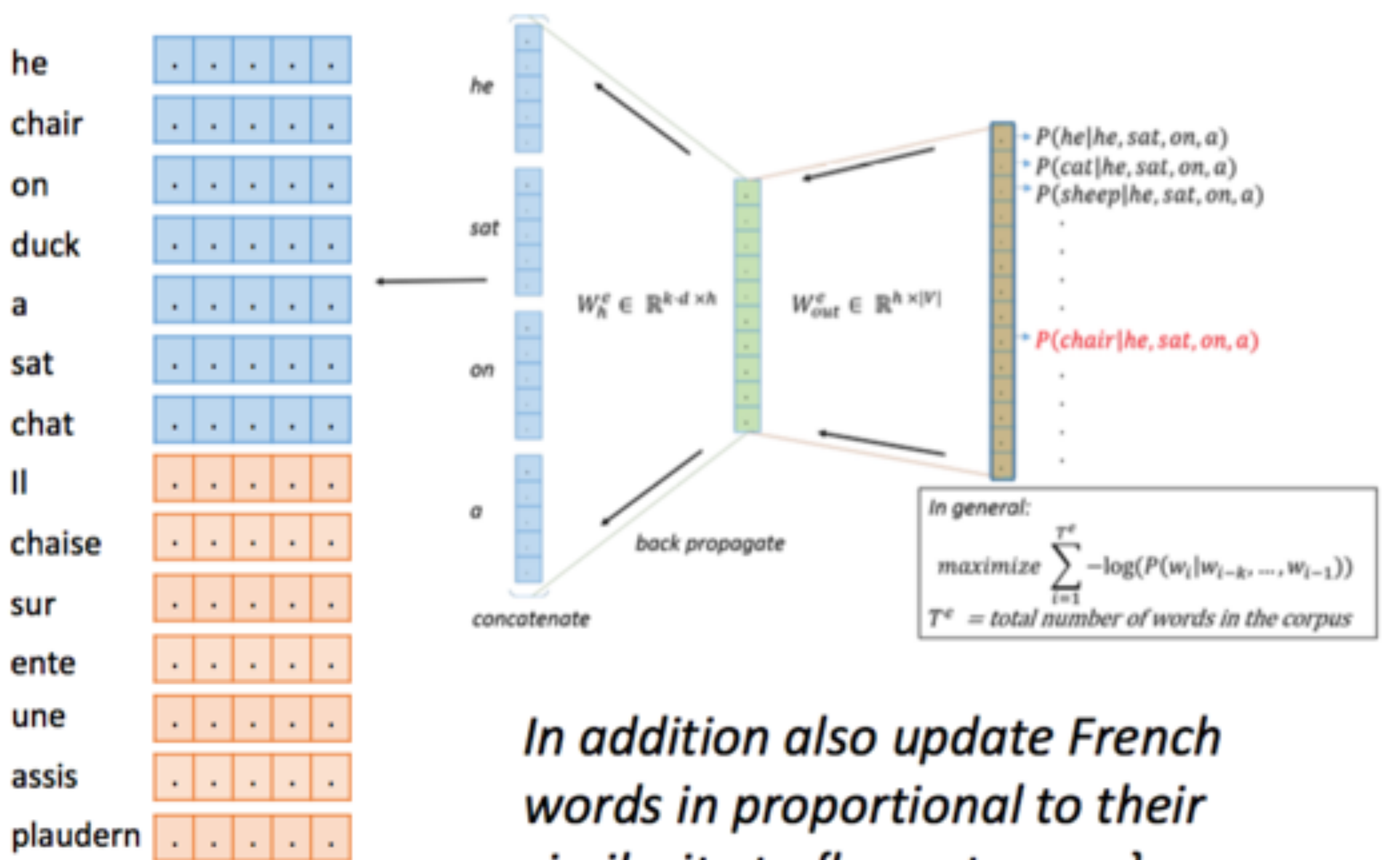
$$w.r.t\ W^e_{emb}, W^f_{emb}$$

$$\boxed{\begin{array}{l} maximize\ max(0, 1 - s^e + s^e_c) \\ w.r.t.\ \theta^f \end{array}}$$

(Gouws et. al., 2015)

Cross-lingual alignment with parallel sentences for MT

English Training instance: he sat on a chair



$W^e_{emb} \in \mathbb{R}^{|V^e| \times d}$

$W^f_{emb} \in \mathbb{R}^{|V^f| \times d}$

In addition also update French words in proportional to their similarity to {he, sat, on, a}

(Klementiev et. al., 2012)

	assis	il	une	sur	chaise
he	0.02	0.9	0.05	0.01	0.02
sat	0.85	0.01	0.02	0.03	0.09
chair	0.06	0.01	0.01	0.01	0.95
a	0.02	0.02	0.92	0.02	0.02
on	0.10	0.05	0.05	0.81	0.04

A

Each cell (i, j) of A stores $sim(w_i, w_j)$ using word alignment information from a parallel corpus

More formally,

$$W^f_{emb_i} = W^f_{emb_i} + \sum_{w_j \in V^e} A_{i,j} \frac{\partial \mathcal{L}(\theta^e)}{\partial W^e_{emb_j}}$$

$$\mathcal{L}(\theta^e) = \sum_{i=1}^{T_e} -\log(P(w_i | w_{i-k}, \ldots, w_{i-1}))$$

Similar words across the two languages undergo similar updates and hence remain close to each other

Unified framework for analysis of cross-lingual methods

- Minimize monolingual objective
- Constraint/Regularize with bilingual objective

$$\text{maximize} \quad \sum_{j \in \{e,f\}} \sum_{i=1}^{T_j} \underbrace{\mathcal{L}(\theta^j)}_{\textit{monolingual similarity}} + \underbrace{\lambda \cdot \Omega\,(W^e_{emb}, W^f_{emb})}_{\textit{bilingual similarity}}$$

$$w.r.t \;\; \theta_e, \theta_f$$

$$\theta_e = W_e, W^e_h, W^e_{out}$$

$$\theta_f = W_f, W^f_h, W^f_{out}$$

Evaluation: Cross-lingual document classification and translation

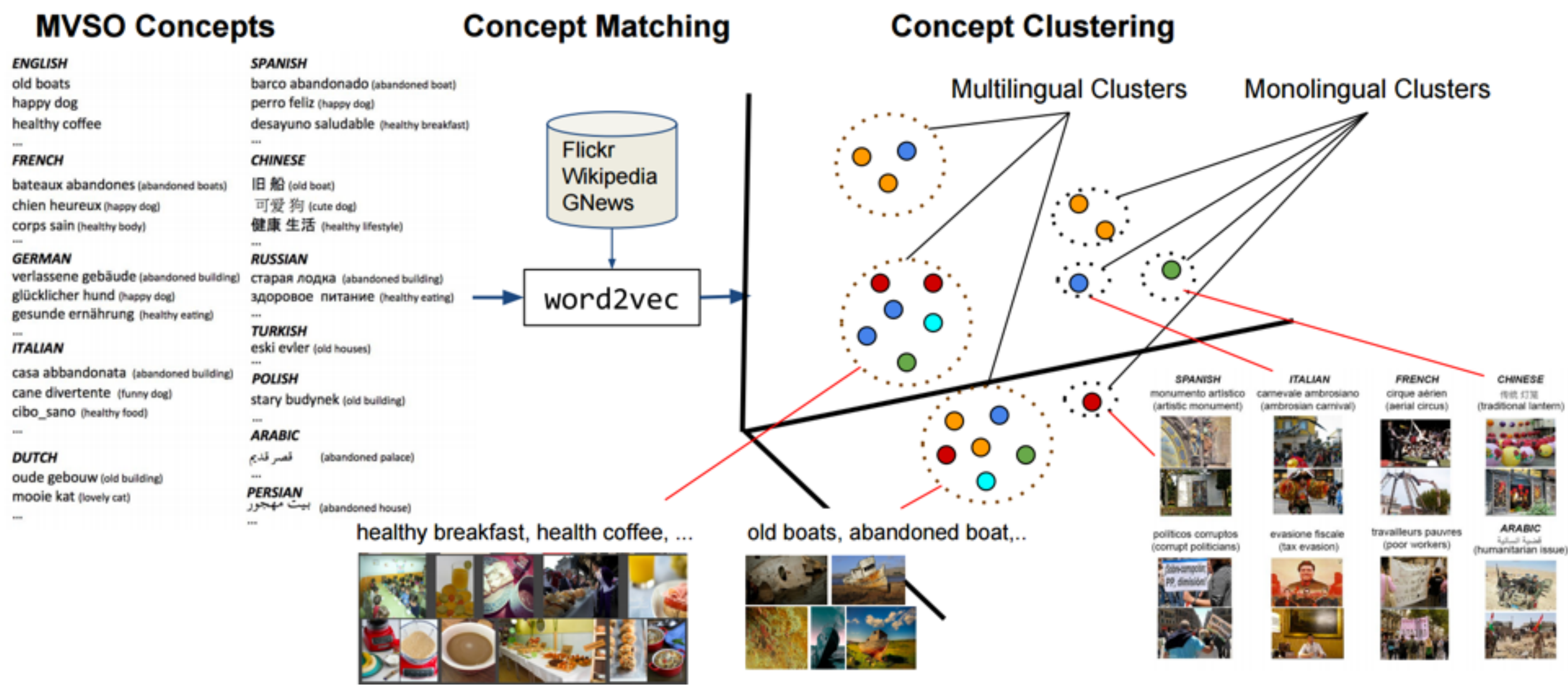
Method	*en → de*	*de → en*	Training Time (min)
Majority Baseline	46.8	46.8	-
Glossed Baseline	65.1	68.6	-
MT Baseline	68.1	67.4	-
Klementiev et al.	77.6	71.1	14,400
Bilingual Auto-encoders (BAEs)	**91.8**	72.8	4,800
BiCVM	83.7	71.4	15
BilBOWA (this work)	86.5	**75**	**6**

Method	En→Sp P@1	Sp→En P@1	En→Sp P@5	Sp→En P@5
Edit Distance	13	18	24	27
Word Co-occurrence	30	19	20	30
Mikolov et al., 2013	33	35	**51**	52
BilBOWA (this work)	**39** (+6)	**44** (+9)	**51**	**55** (+3)

(Gows et al., 2015)

Bonus: Multilingual visual sentiment concept matching

concept = adjective-noun-phrase (ANP)



(Pappas et al., 2016)

Multilingual visual sentiment concept ontology

- 7.36M+ Flickr images
- ~16K affective visual concepts: Adjective-Noun Pairs (ANPs)
- Co-occurrence (emotion, ANP)
- Sentiment value (text-based)
- 12 languages detected

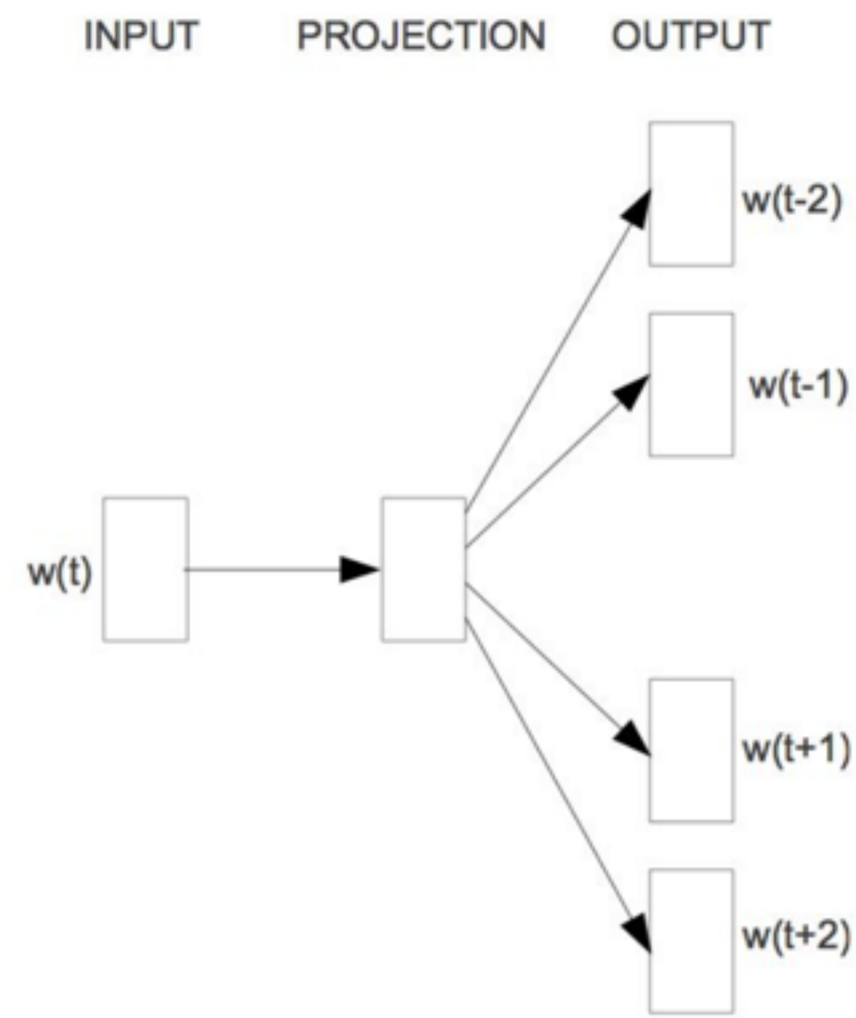
Language	Concepts	Images
English	4421	447997
Spanish	3381	37528
Italian	3349	25664
French	2349	16807
Chinese	504	5562
German	804	7335
Dutch	348	2226
Russian	129	800
Turkish	231	638
Polish	63	477
Persian	15	34
Arabic	29	23

(Jou et al., 2015)

Word embedding model

- Skip-gram model (`word2vec`)[1]
  - Google News 100B
  - Wikipedia 1.74B
  - Wikipedia + Reuters + WSJ 1.96B
  - Flickr 100 Million 0.75B
- Concept vectors
  - Sum of words composition
  - Directly learned (ANPs as tokens)

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado and Jeffrey Dean
Distributed Representations of Words and Phrases and their Compositionality
NIPS, Lake Tahoe, Nevada, USA, 2013



(Pappas et al., 2016)

Multilingual visual sentiment concept retrieval

- How often do two visual concepts appear together?
  - **Tag co-occurrence matrix ($n \times n$)**
- ANPs can be described as
  - **Co-occurrence vectors h_i, h_j in R^n**
    - n is the number of translated ANPs

- **Visual semantic distance between ANPs**

$$d(ANP_i, ANP_j) = 1 - cosine(h_i, h_j)$$

(Pappas et al., 2016)

Multilingual visual sentiment concept clustering

Visual **Semantic** Relatedness for different clustering methods

For each clustering method:

$$\text{sem}_C = \frac{1}{C} \sum_{c=1}^{C} \boxed{\frac{\sum_{j:j \neq i \ \& \ Uij \neq 0}^{|\{i,\ldots,N_c\}|} d(\text{ANP}_{c,i}, \text{ANP}_{c,j})}{N_c}}$$

Average over all clusters

Average visual semantic distance in a cluster for all ANP pairs whose semantic distance is greater than 0

C = number of non-unary clusters
Nc = number of ANPs for a cluster c

Inter-cluster distance was not significantly different

(Pappas et al., 2016)

Multilingual visual sentiment concept clustering

Visual **Sentiment** Consistency for different clustering methods

For each clustering method:

Average sentiment in a cluster

$$\mathrm{sen}_C = \frac{1}{C}\sum_{c=1}^{C} \boxed{\frac{\sum_{i=1}^{N_c}(\mathrm{sen}(\mathrm{ANP}_{c,i}) - \boxed{\mathrm{sen}_c})^2}{N_c}}$$

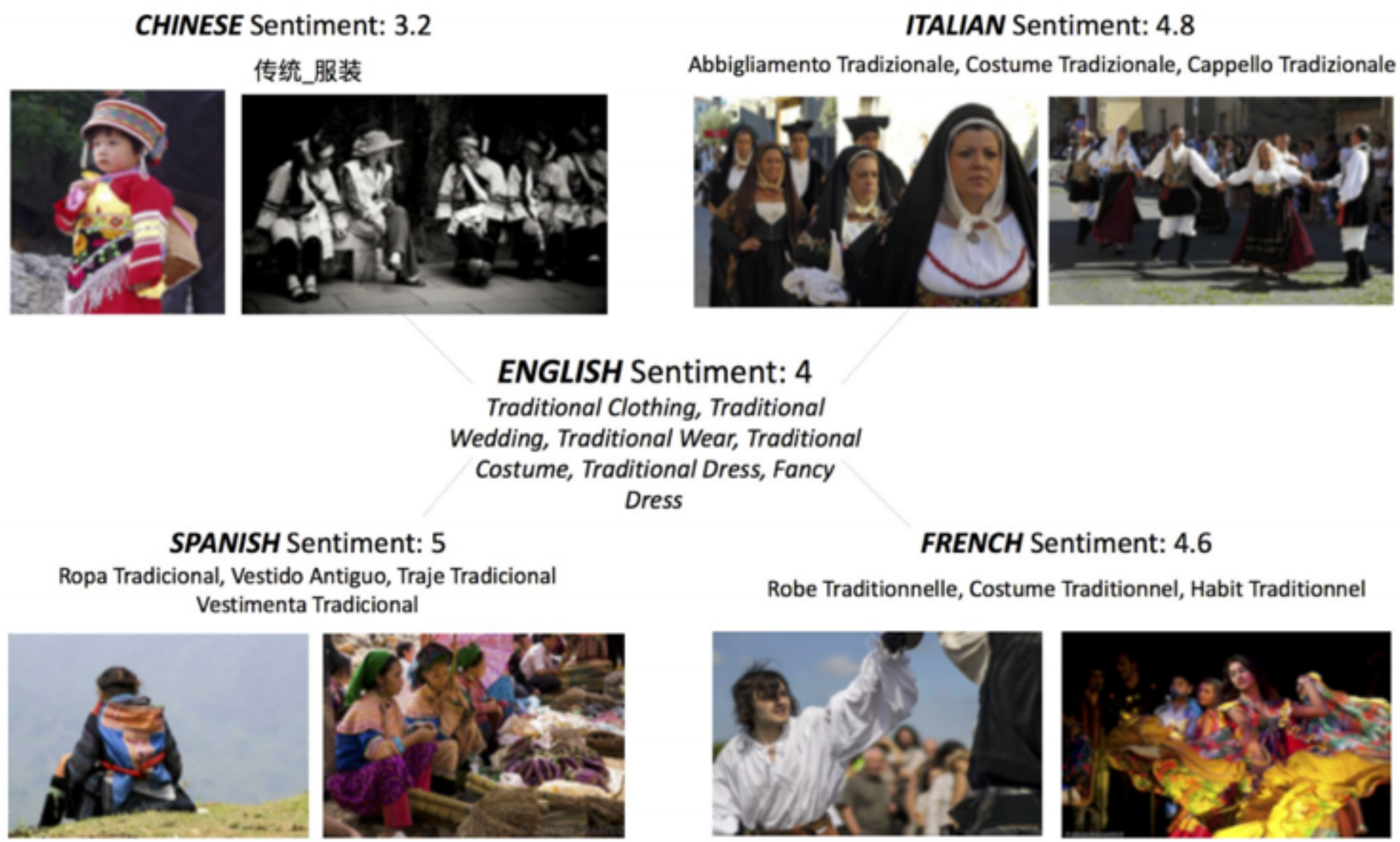
Average over all clusters

Average visual sentiment error in a cluster

C = number of non-unary clusters
Nc = number of ANPs for a cluster c

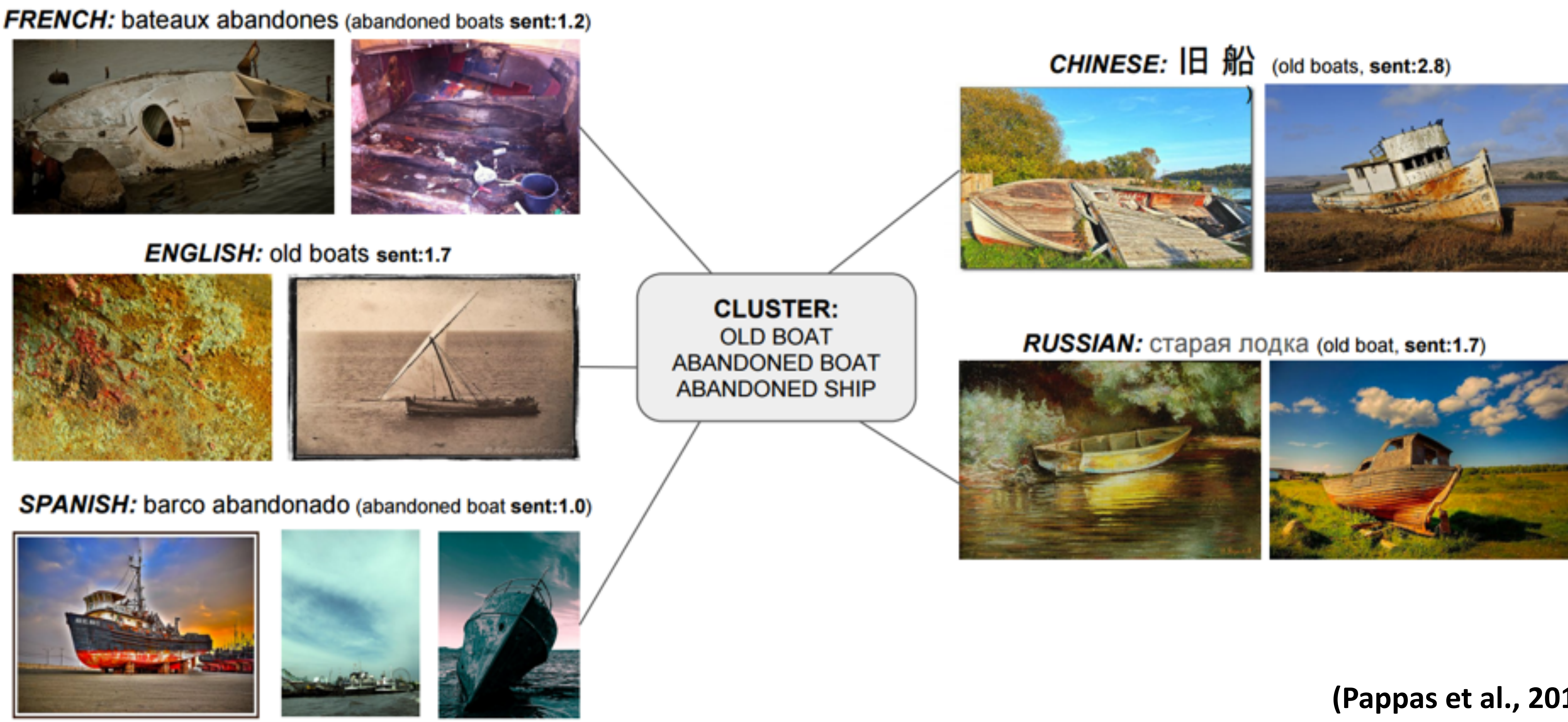
(Pappas et al., 2016)

Discovering interesting clusters: Multilingual



(Pappas et al., 2016)

Discovering interesting clusters: Western vs. Eastern



(Pappas et al., 2016)

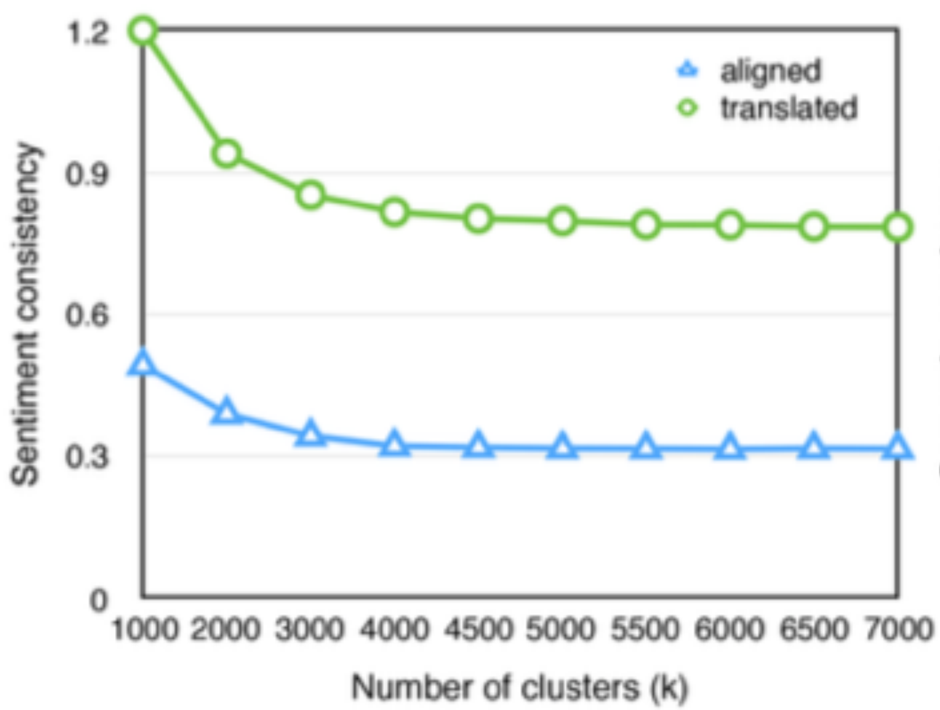
Discovering interesting clusters: Monolingual

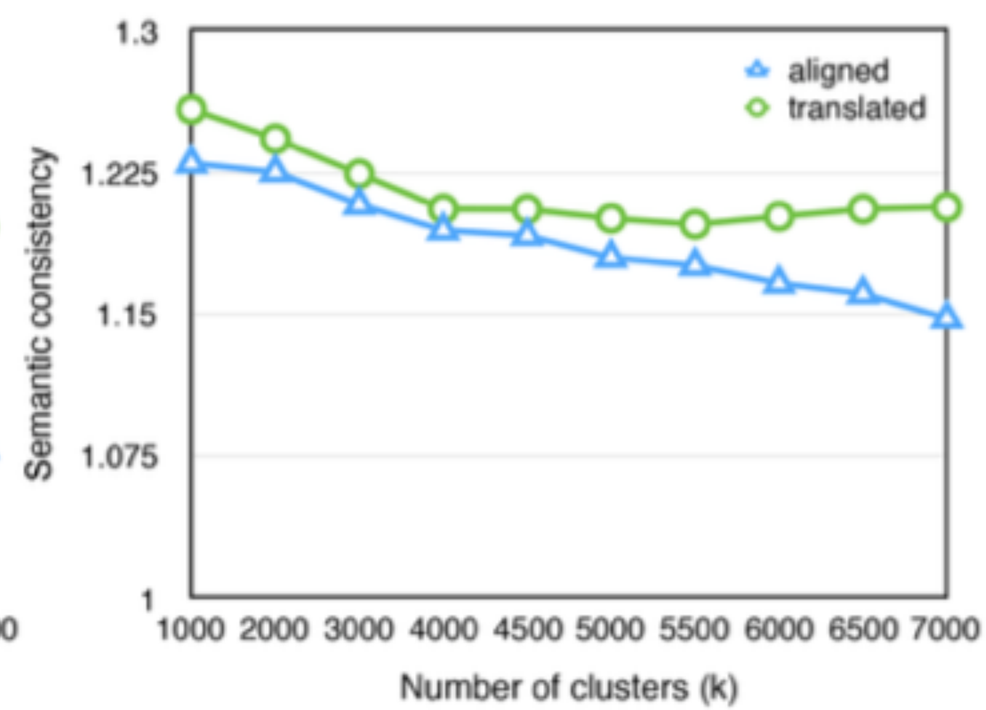
(Pappas et al., 2016)

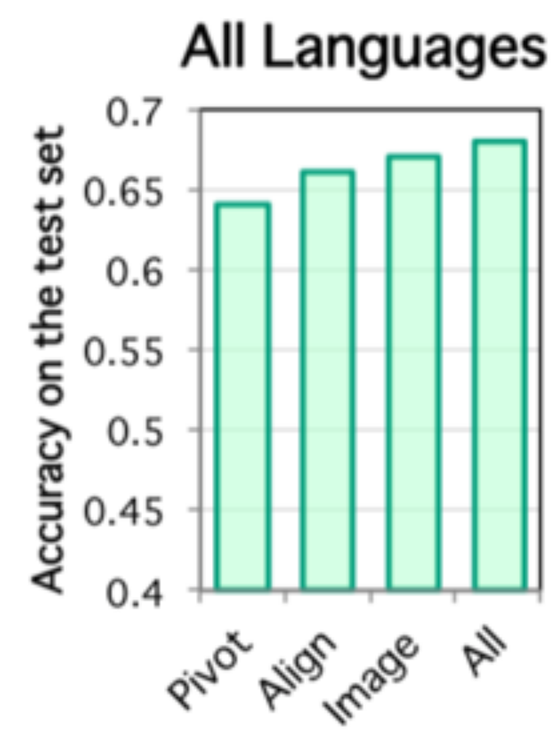
Evaluation: Multilingual visual sentiment concept analysis

- Aligned embeddings are better than translation in concept retrieval, clustering and sentiment prediction

Method \ Language	EN	ES	IT	FR	ZH	DE	NL	RU	TR
Translated concepts (w=5)	5.94	4.86	5.49	5.23	5.41	6.27	7.96	13.50	11.72
Aligned concepts (w=5)	5.94	3.05	3.77	4.20	2.22	4.08	6.60	17.83	15.85
Improvement (%)	+0.0	+59.3	+45.6	+24.5	+143.6	+53.6	+20.6	-32.0	-35.2







Conclusion

- Aligned embeddings are cheaper than translation and usually work better than it in several multilingual or crosslingual NLP tasks without parallel data
  - document classification Gows et al., 2015
  - named entity recognition Al-Rfou et al., 2014
  - dependency parsing Guo et al., 2015
  - concept retrieval and clustering Pappas et al., 2016

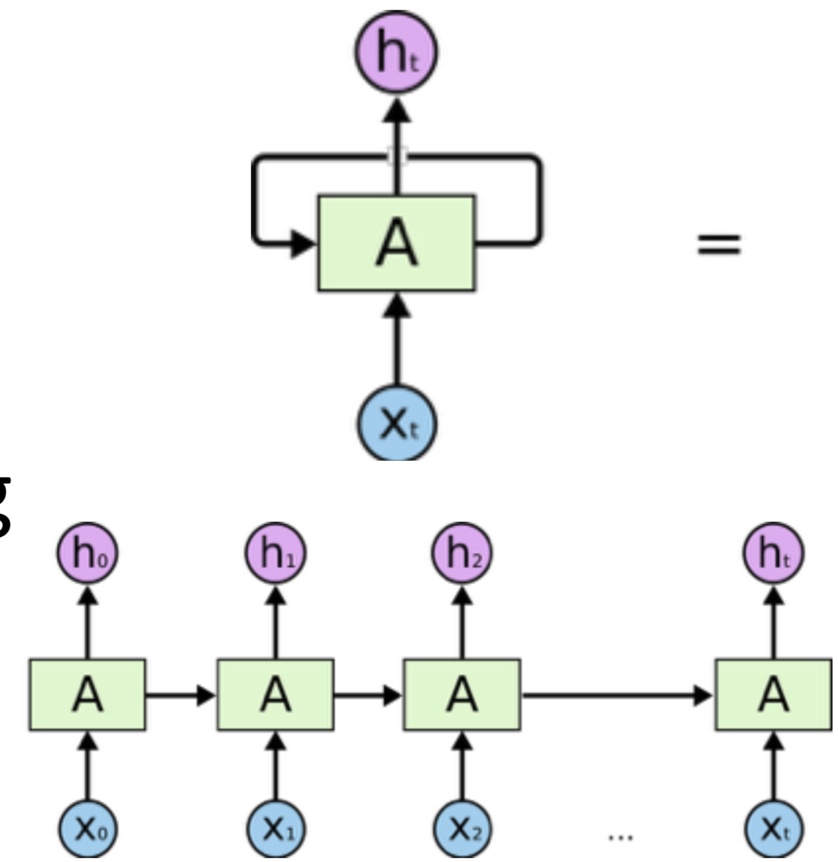
Outline of the talk

1. Recap: Word Representation Learning

2. Multilingual Word Representations
   - Alignment models
   - Evaluation tasks

3. Multilingual Word Sequence Modeling
   - Essentials: RNN, LSTM, GRU
   - Machine Translation
   - Document Classification

4. Summary



* Figure from Colah's blog, 2015.

Language Modeling

- Computes the probability of a sequence of words or simply "likelihood of a text": $P(w_1, w_2, \ldots, w_t)$

- N-gram models with Markov assumption:

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

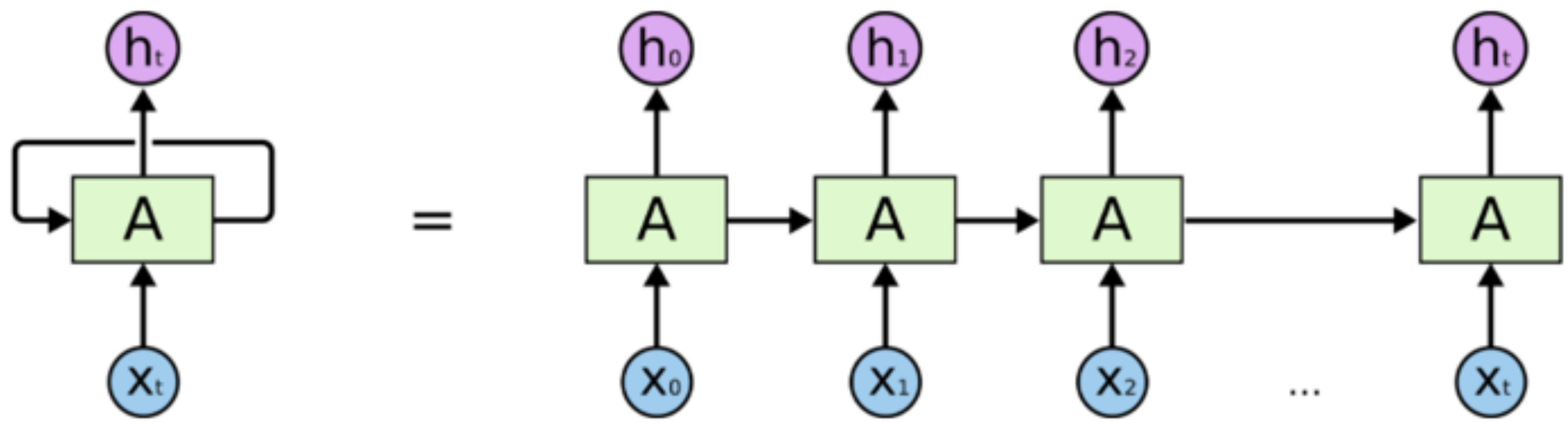
- **Where is it useful?**
  - speech recognition
  - machine translation
  - POS tagging and parsing

- **What are its limitations?**
  - unrealistic assumption
  - huge memory needs
    - back-off models

Recurrent Neural Network (RNN)

- Neural language model: $h_t = \sigma\left(W^{(hh)}h_{t-1} + W^{(hx)}x_{[t]}\right)$

$$\hat{P}(x_{t+1} = v_j \mid x_t, \ldots, x_1) = \hat{y}_t = \text{softmax}\left(W^{(S)}h_t\right)$$

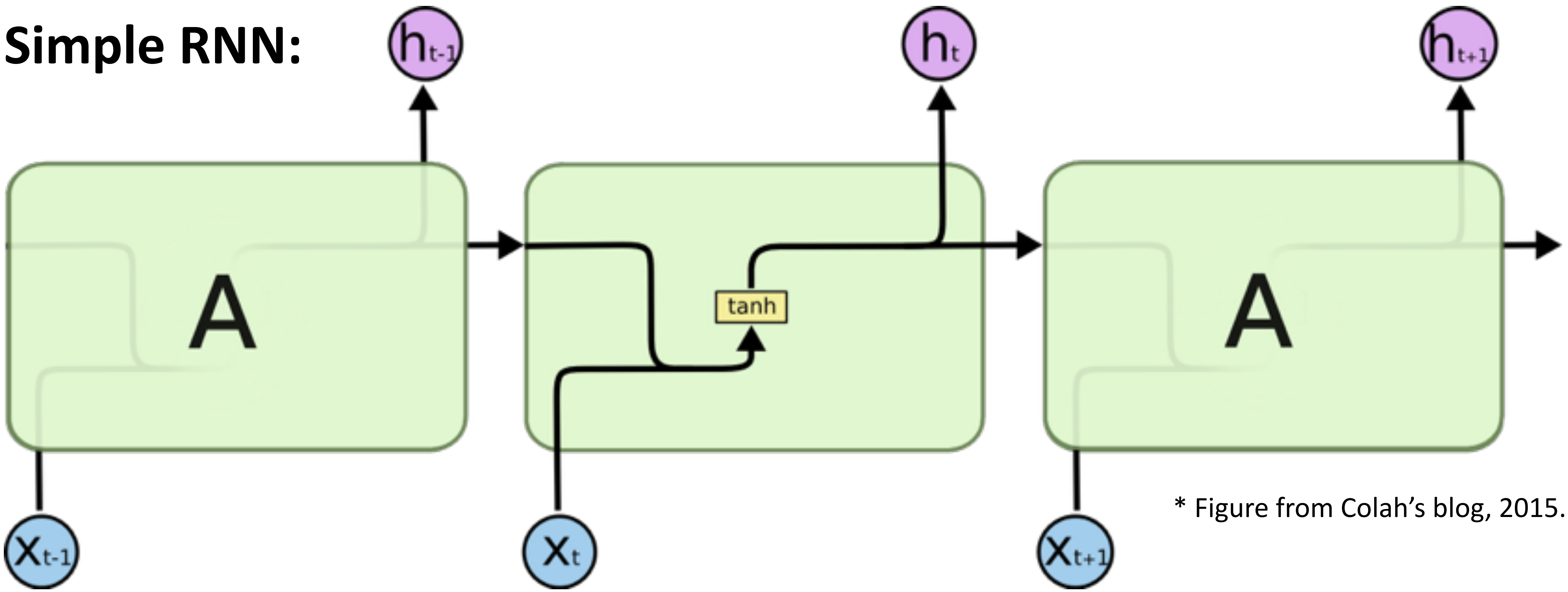


- **What are its main limitations?**
  - vanishing gradient problem (error doesn't propagate far)
  - fail to capture long-term dependencies
    - **tricks:** gradient clipping, identity initialization + ReLus

Long Short Term Memory (LSTM)

- Long-short term memory nets are able to learn long-term dependencies: Hochreiter and Schmidhuber 1997

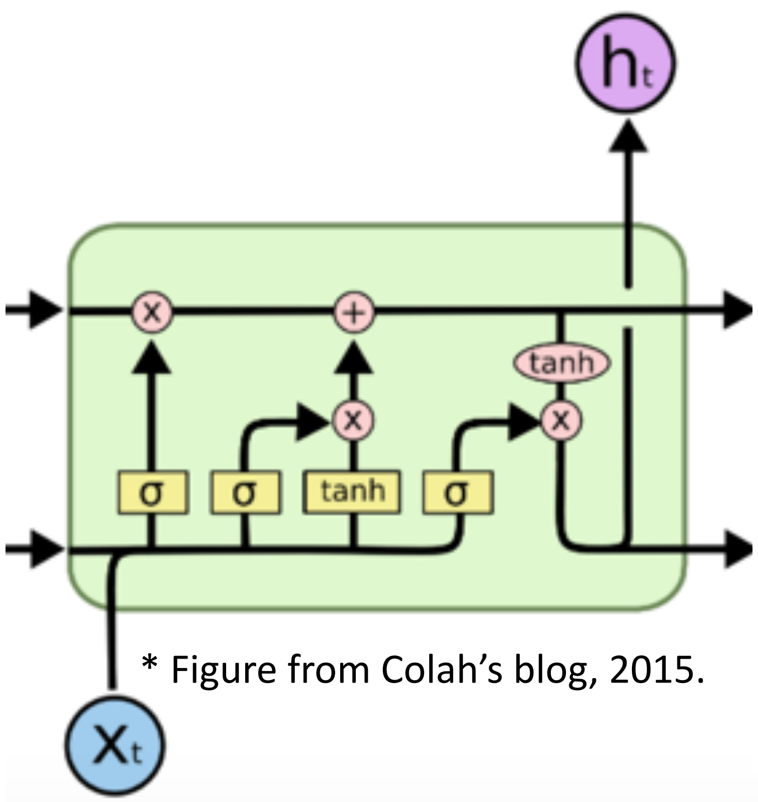
Simple RNN:



* Figure from Colah's blog, 2015.

Long Short Term Memory (LSTM)

- Long-short term memory nets are able to learn long-term dependencies: Hochreiter and Schmidhuber 1997
  - Ability to remove or add information to the cell state regulated by “gates”



* Figure from Colah’s blog, 2015.

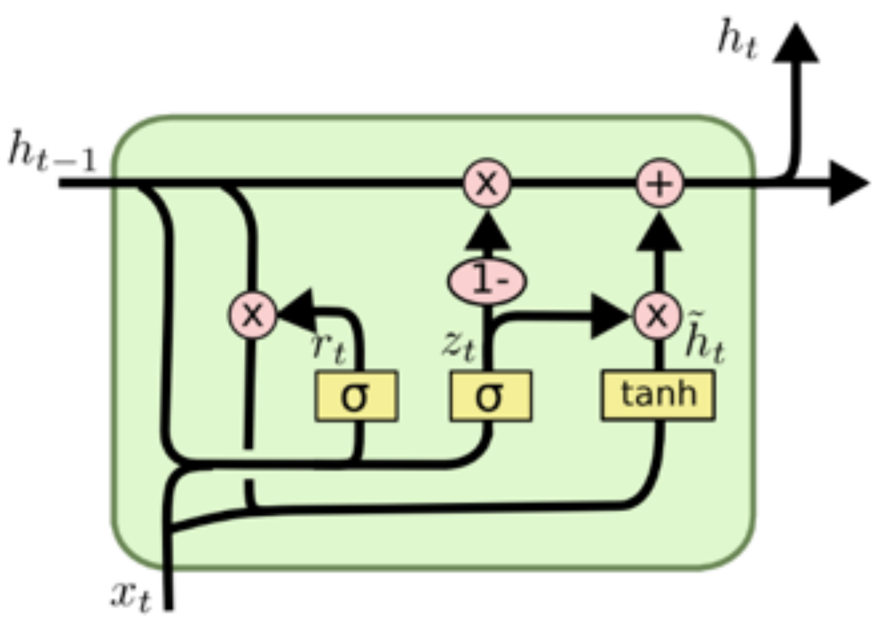
- Input gate (current cell matters) $i_t = \sigma\left(W^{(i)}x_t + U^{(i)}h_{t-1}\right)$
- Forget (gate 0, forget past) $f_t = \sigma\left(W^{(f)}x_t + U^{(f)}h_{t-1}\right)$
- Output (how much cell is exposed) $o_t = \sigma\left(W^{(o)}x_t + U^{(o)}h_{t-1}\right)$
- New memory cell $\tilde{c}_t = \tanh\left(W^{(c)}x_t + U^{(c)}h_{t-}\right)$

Final memory cell: $c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$

Final hidden state: $h_t = o_t \circ \tanh(c_t)$

Gated Recurrent Unit (GRU)

- Gated RNN by Chung et al, 2014 combines the forget and input gates into a single "update gate"
  - keep memories to capture long-term dependencies
  - allow error messages to flow at different strengths



* Figure from Colah's blog, 2015.

$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

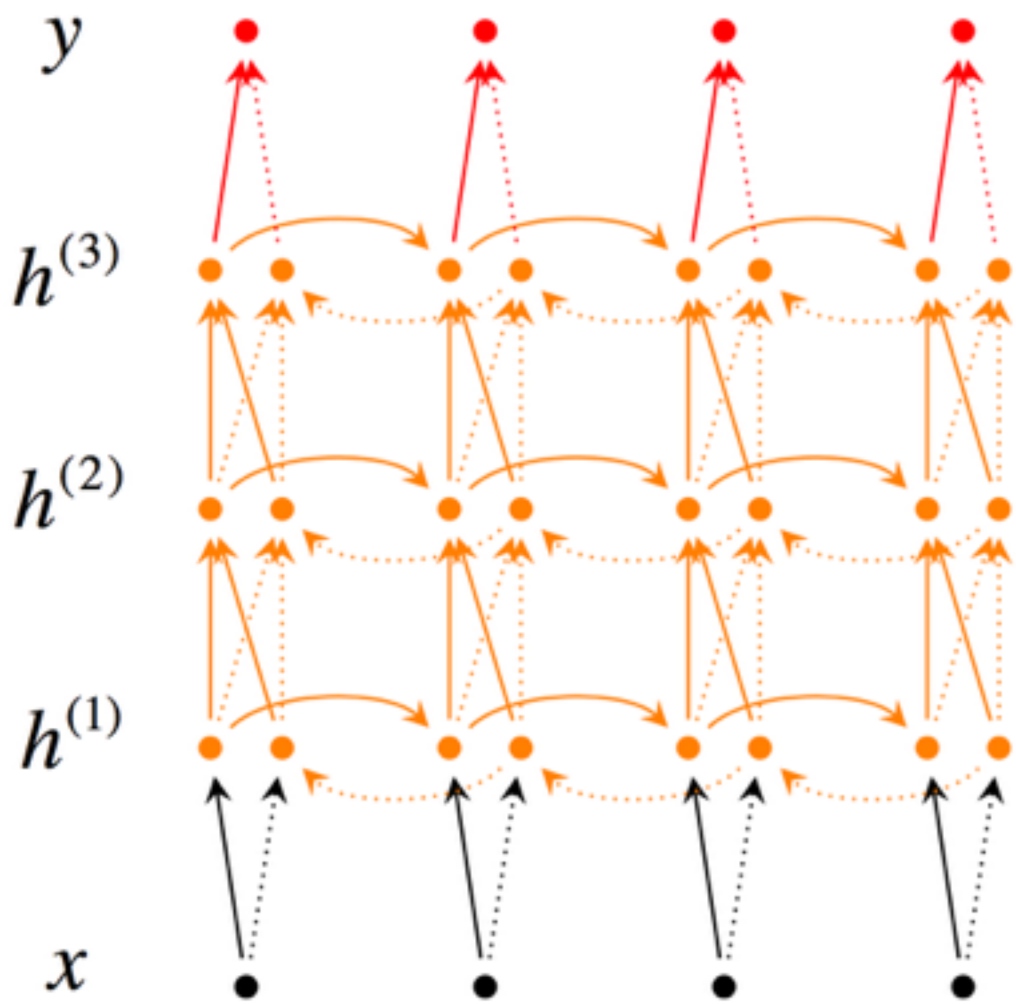
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

z_t: update gate — r_t: reset gate — h_t: regular RNN update

Deep Bidirectional Models

- Here RNN but it applies to LSTMs and GRUs too



$$\overrightarrow{h}_t^{(i)} = f(\overrightarrow{W}^{(i)} h_t^{(i-1)} + \overrightarrow{V}^{(i)} \overrightarrow{h}_{t-1}^{(i)} + \overrightarrow{b}^{(i)})$$

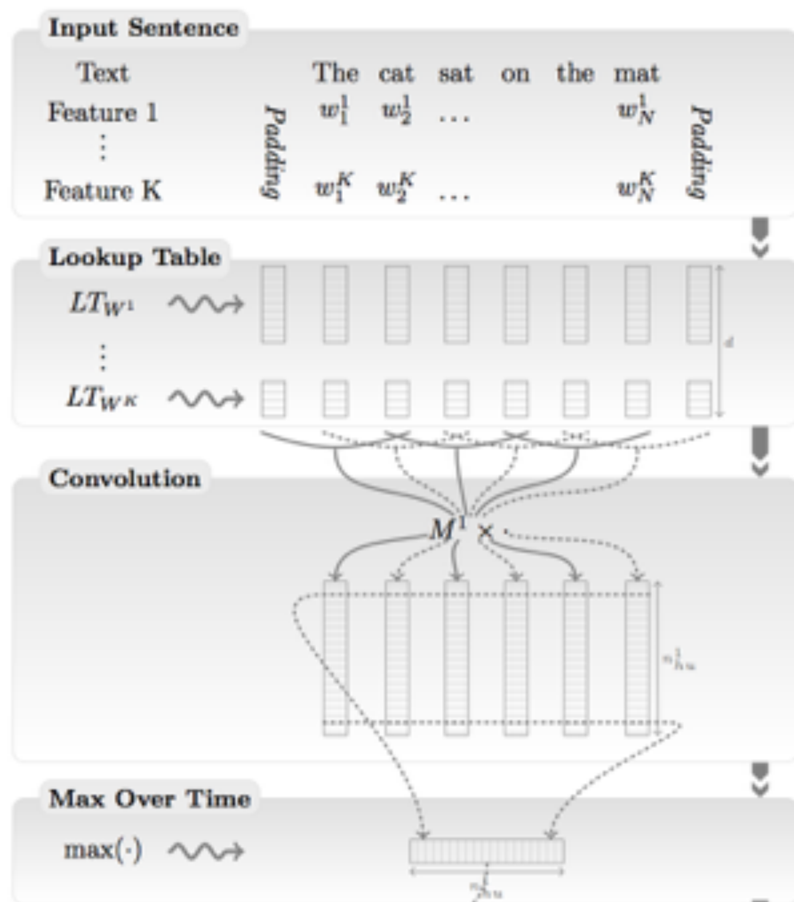
$$\overleftarrow{h}_t^{(i)} = f(\overleftarrow{W}^{(i)} h_t^{(i-1)} + \overleftarrow{V}^{(i)} \overleftarrow{h}_{t+1}^{(i)} + \overleftarrow{b}^{(i)})$$

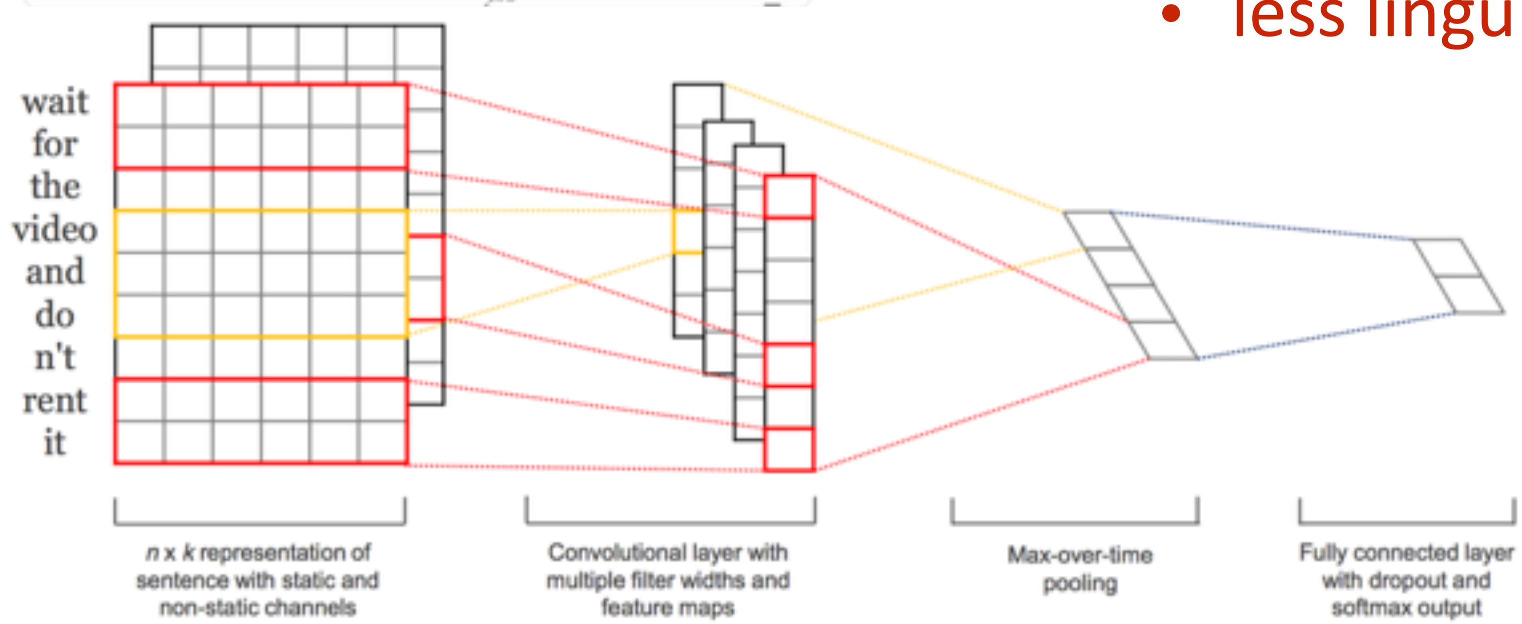
$$y_t = g(U[\overrightarrow{h}_t^{(L)}; \overleftarrow{h}_t^{(L)}] + c)$$

(Irsoy and Cardie, 2014)

Each memory layer passes an intermediate sequential representation to the next.

Convolutional Neural Network (CNN)





- Typically good for images
- Convolutional filter(s) is (are) applied every k words:

$$c_i = f(\mathbf{w}^T \mathbf{x}_{i:i+h-1} + b)$$

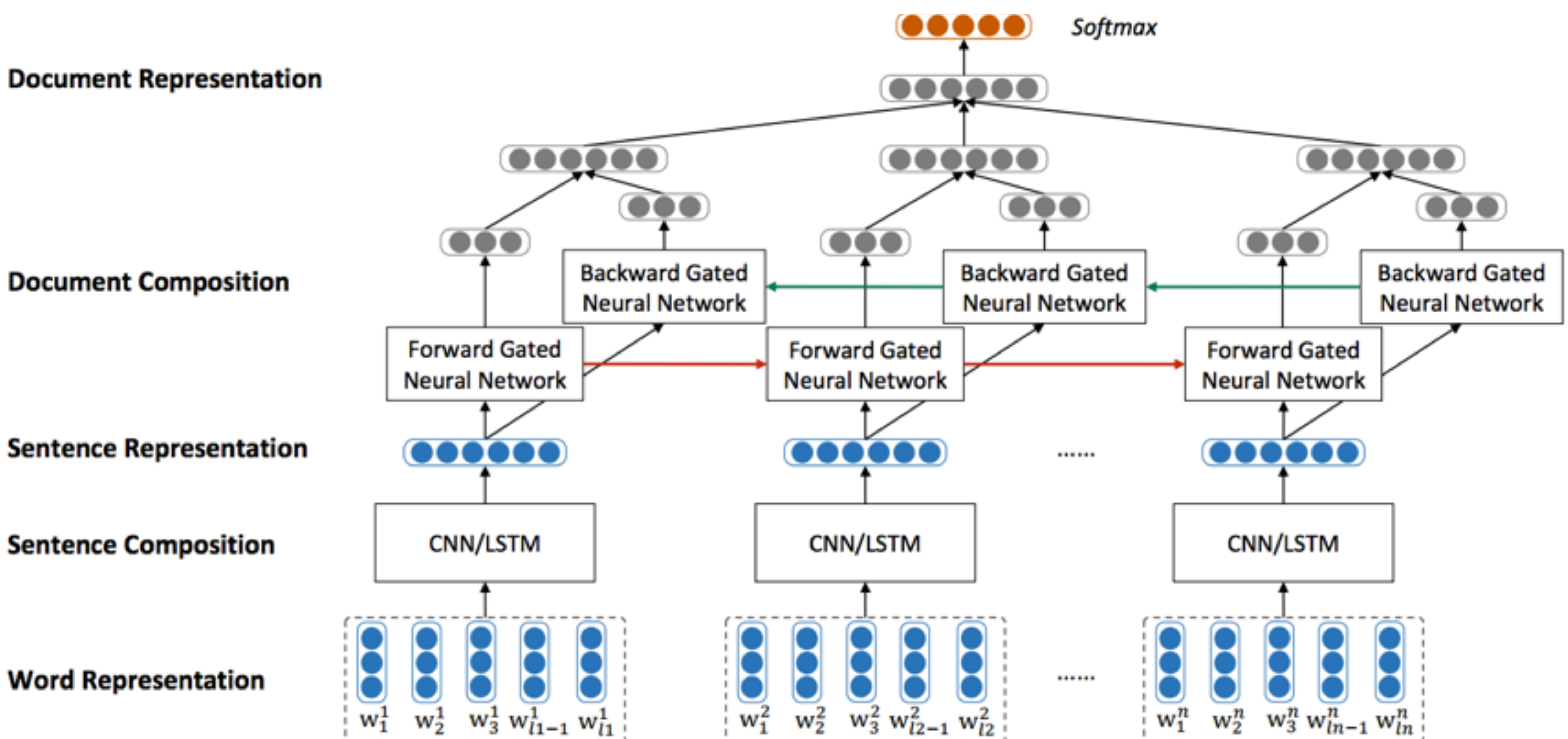
- Similar to Recursive NNs but without constraining to grammatical phrases only, as Socher et al., 2011
  - no need for a parser (!)
  - less linguistically motivated ?

(Collobert et al., 2011)

(Kim, 2014)

Hierarchical Models

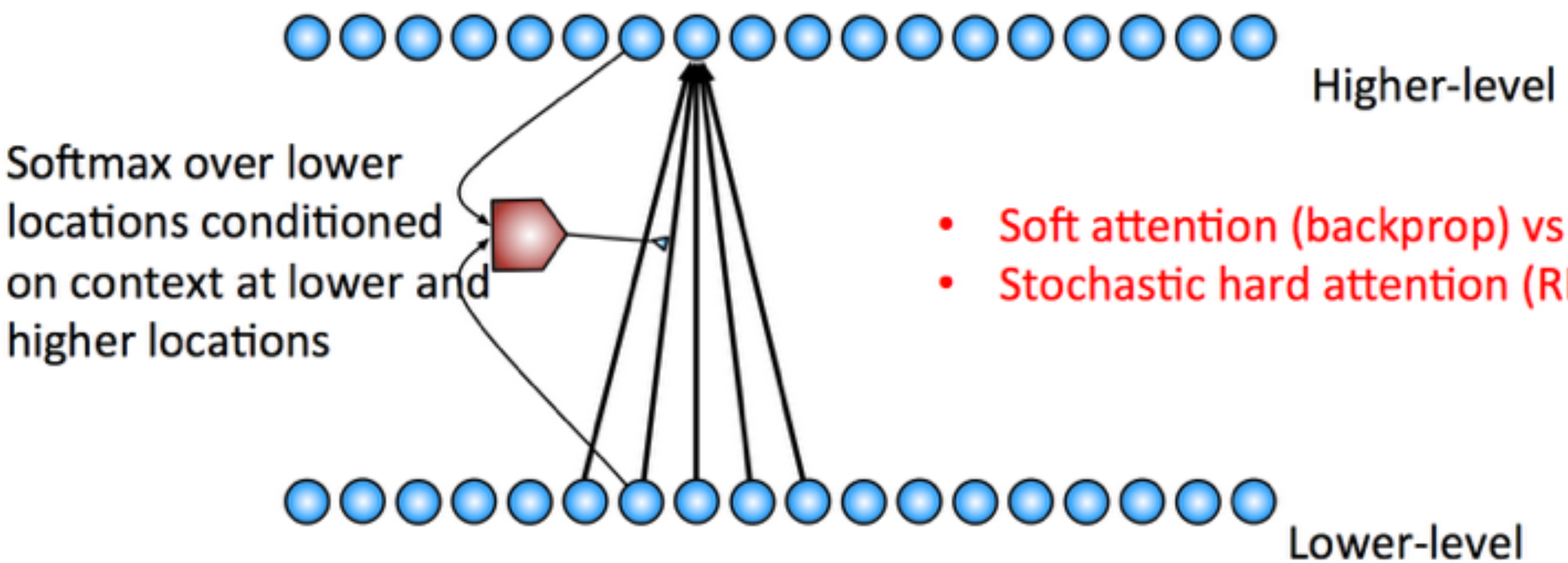
- Word-level and sentence-level modeling with any type of NN layers



(Tang et al., 2015)

Attention Mechanism for Machine Translation

- Chooses "where to look" or learns to assign a relevance to each input position given encoder hidden state for that position and the previous decoder state
  - learns a soft bilingual alignment model

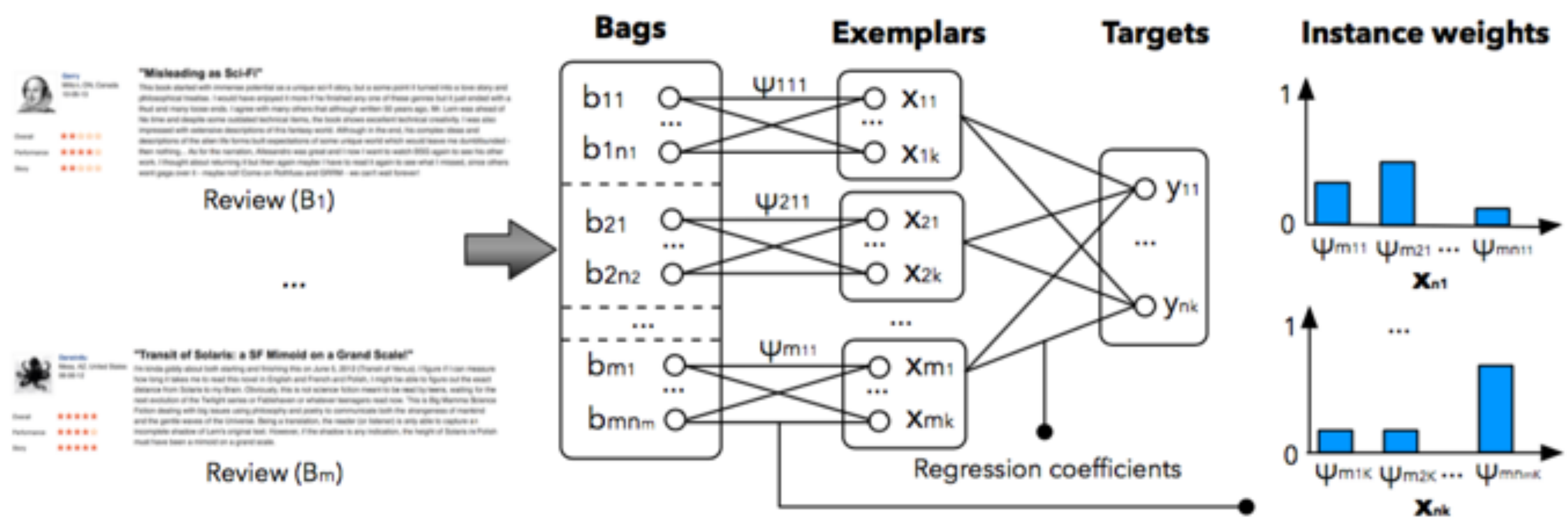


- Soft attention (backprop) vs
- Stochastic hard attention (RL)

(Bahdanau et al., 2015)

Attention Mechanism for Document Classification

- Operates on input word sequence (or intermediate hidden states: Pappas and Popescu-Belis 2016)
- Learns to focus on relevant parts of the input with respect to the target labels
  - learns a soft extractive summarization model



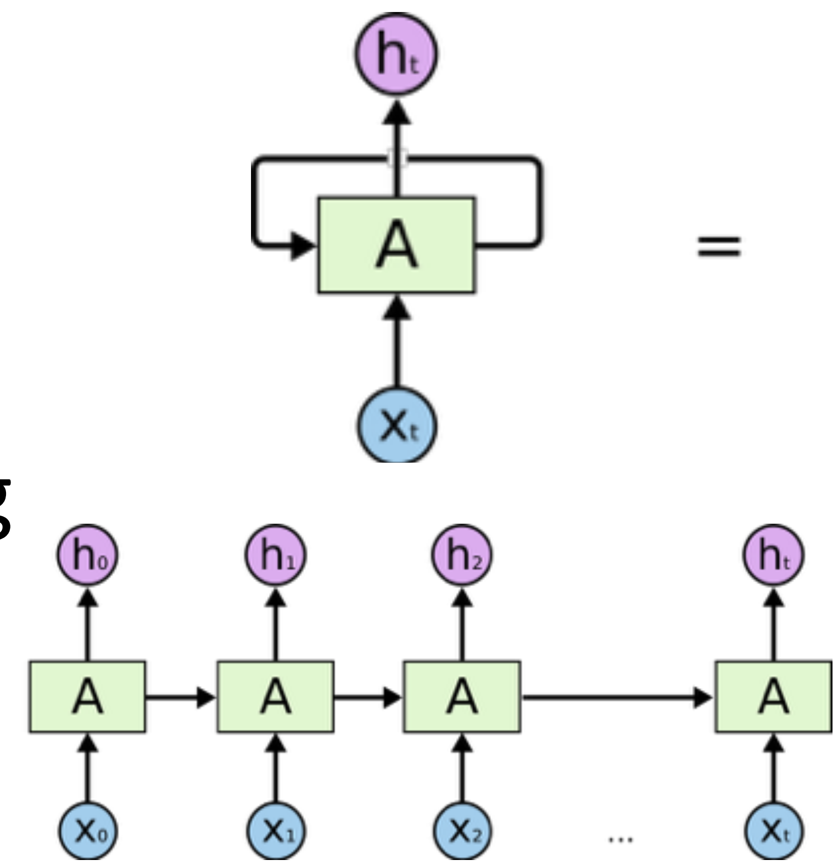
(Pappas and Popescu-Belis, 2014)

Outline of the talk

1. Recap: Word Representation Learning

2. Multilingual Word Representations
   - Alignment models
   - Evaluation tasks

3. Multilingual Word Sequence Modeling
   - Essentials: RNN, LSTM, GRU
   - Machine Translation
   - Document Classification



* Figure from Colah's blog, 2015.

4. Summary

RNN encoder-decoder for Machine Translation

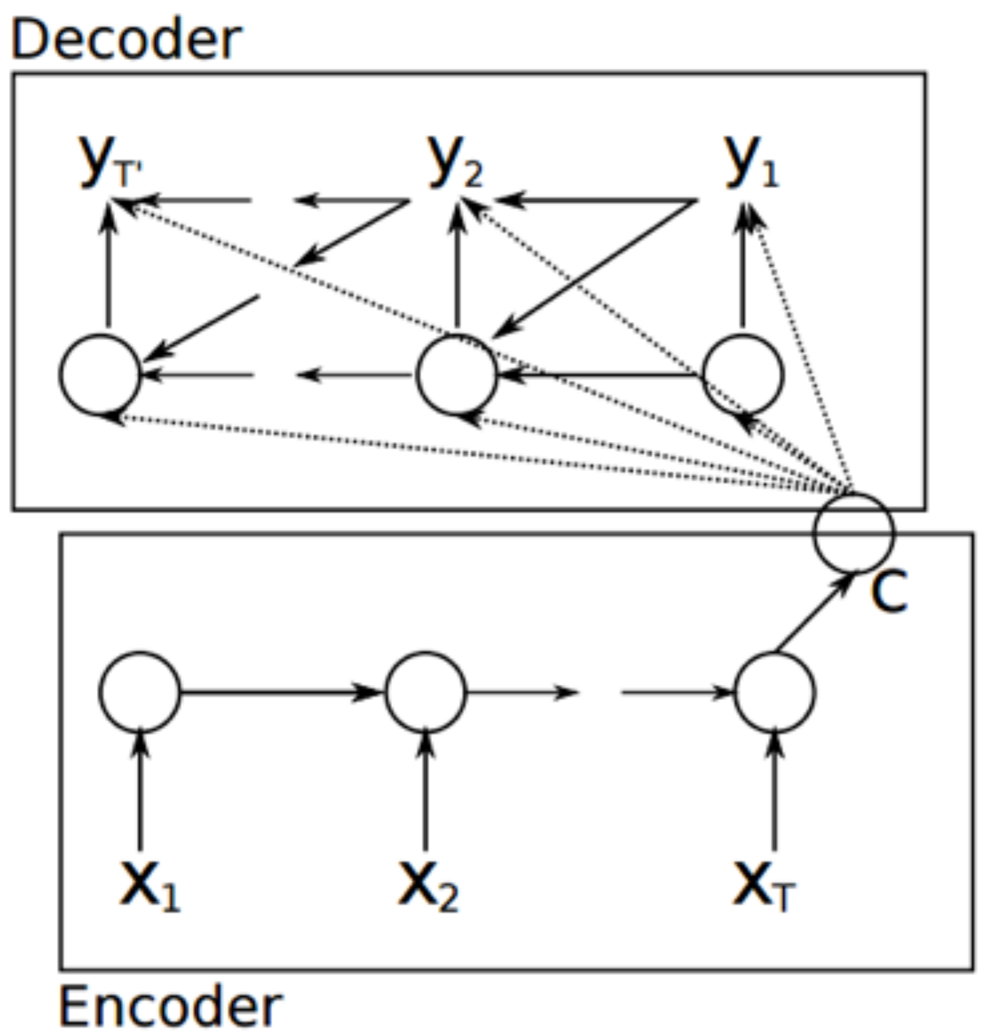


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

- GRU as hidden layer
- Maximize the log likelihood of the target sequence given the source sequence:

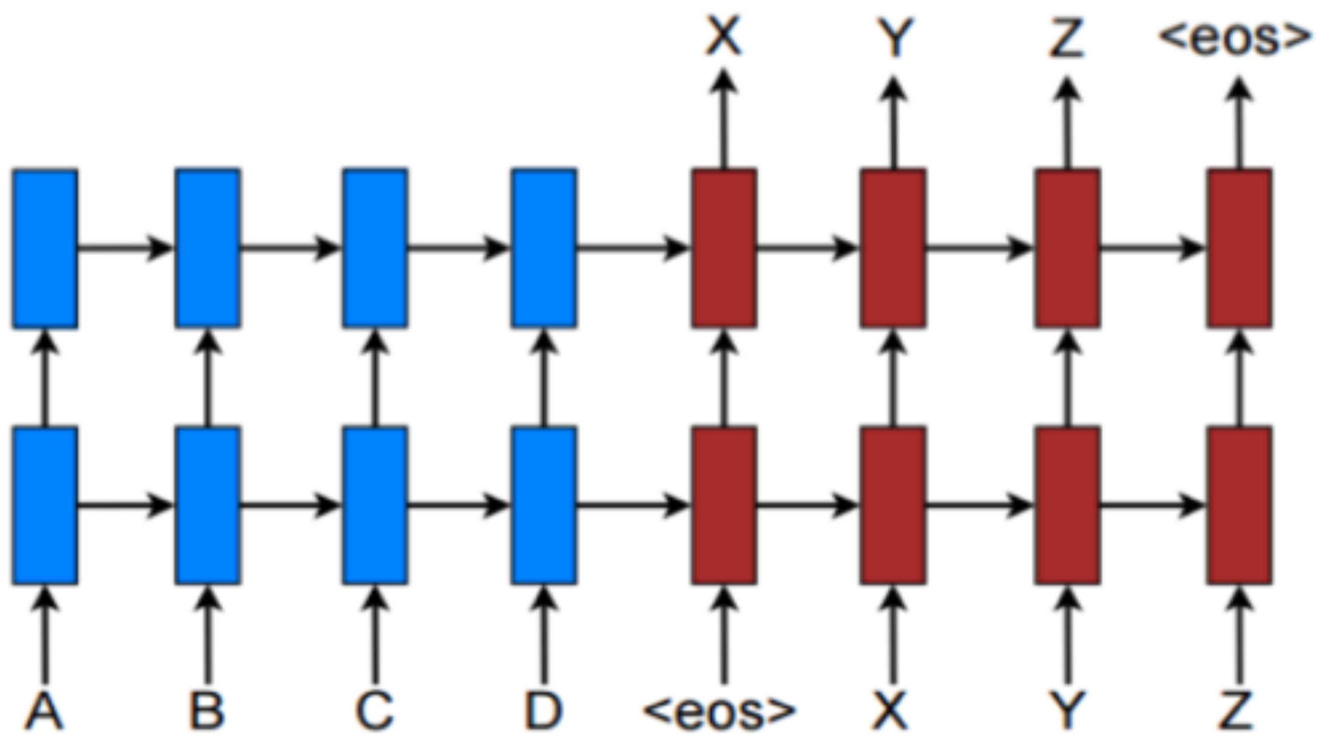
$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p_{\theta}(Y_n|X_n)$$

- WMT 2014 (EN→FR)

Models	BLEU dev	BLEU test
Baseline	30.64	33.30
RNN	31.20	33.87
CSLM + RNN	31.48	34.64
CSLM + RNN + WP	31.50	34.54

(Cho et al., 2014)

Sequence to sequence learning for Machine Translation



- LSTM hidden layers instead of GRU
- 4 layers deep instead of shallow encoder-decoder

(Sutskever et al., 2014)

Sequence to sequence learning for Machine Translation

- WMT 2014 (EN→FR)

Trick-1: Reverse the input sequence.

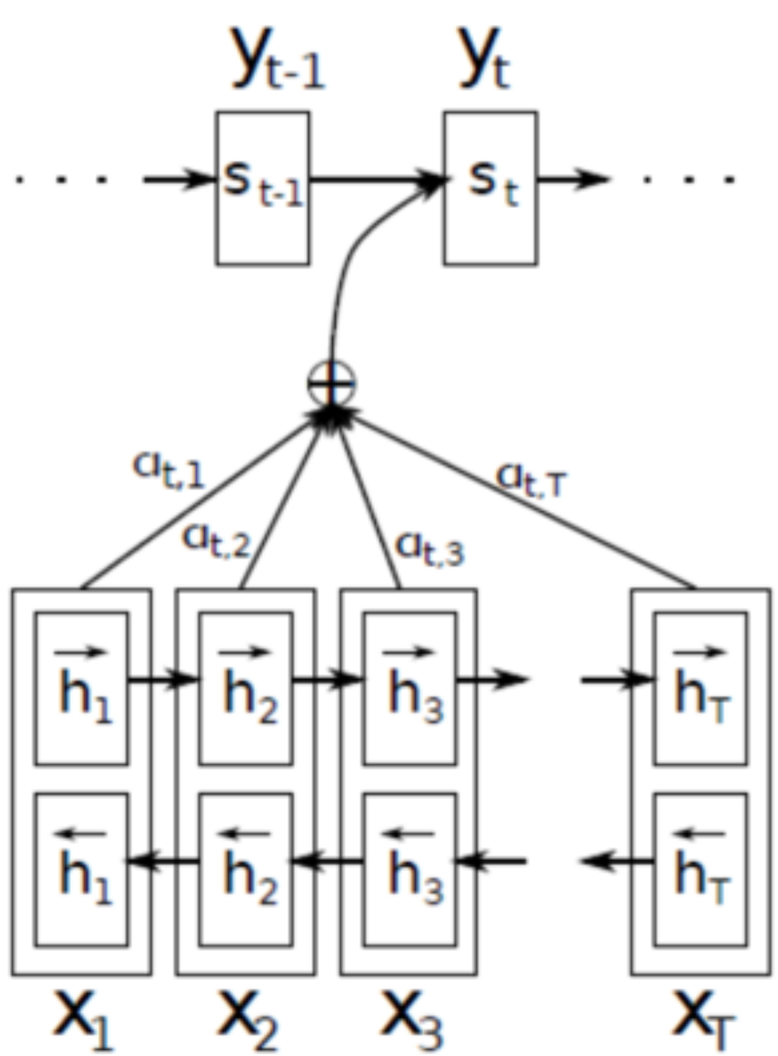
Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Trick-2: Ensemble Neural Nets.

- PCA projection of the hidden state of the last encoder layer

(Sutskever et al., 2014)

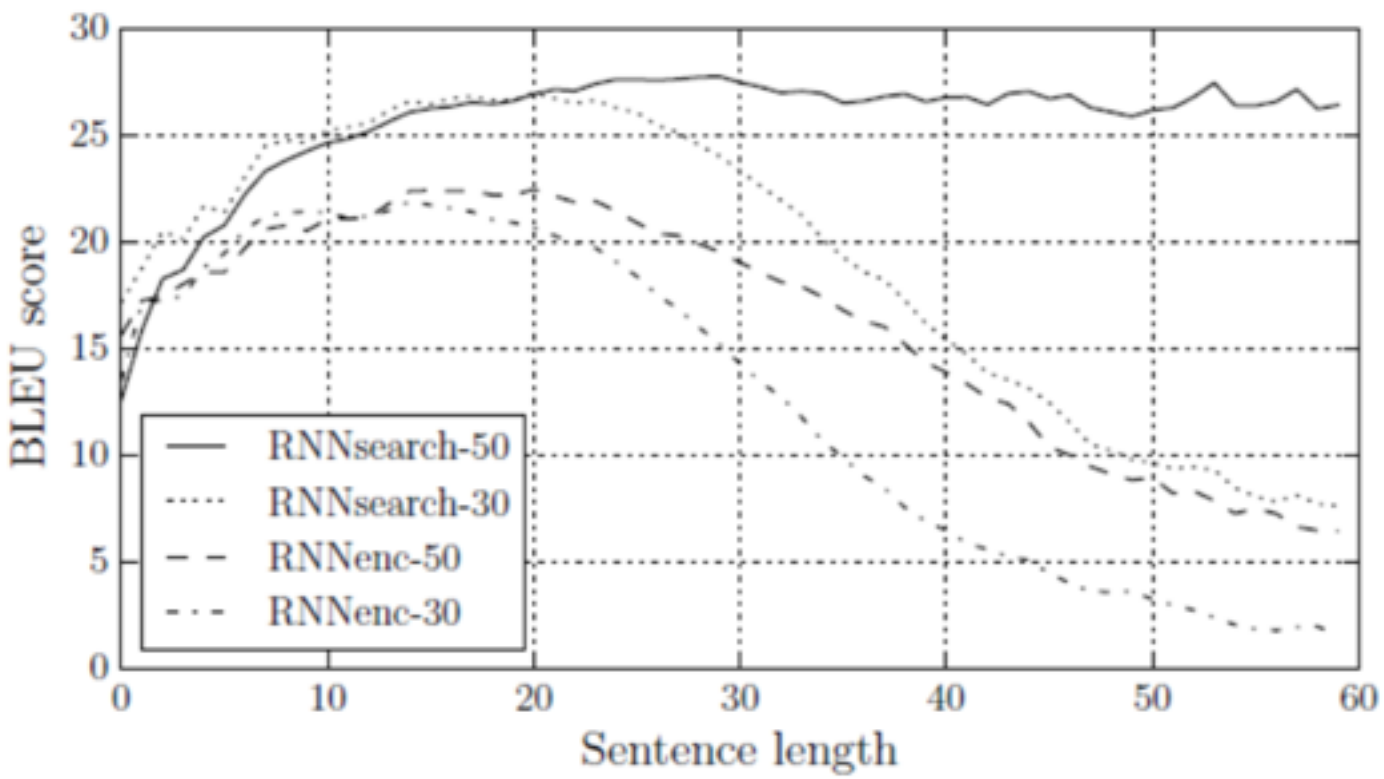
Jointly learning to align and translate for Machine Translation



- **Limitation:** can we compress all the needed information in the last encoder state?
- **Idea:** use all the hidden states of the encoder
  - length proportional to that of the sentence!
  - compute a weighted average of all the hidden states

(Bahdanau et al., 2015)

Jointly learning to align and translate for Machine Translation



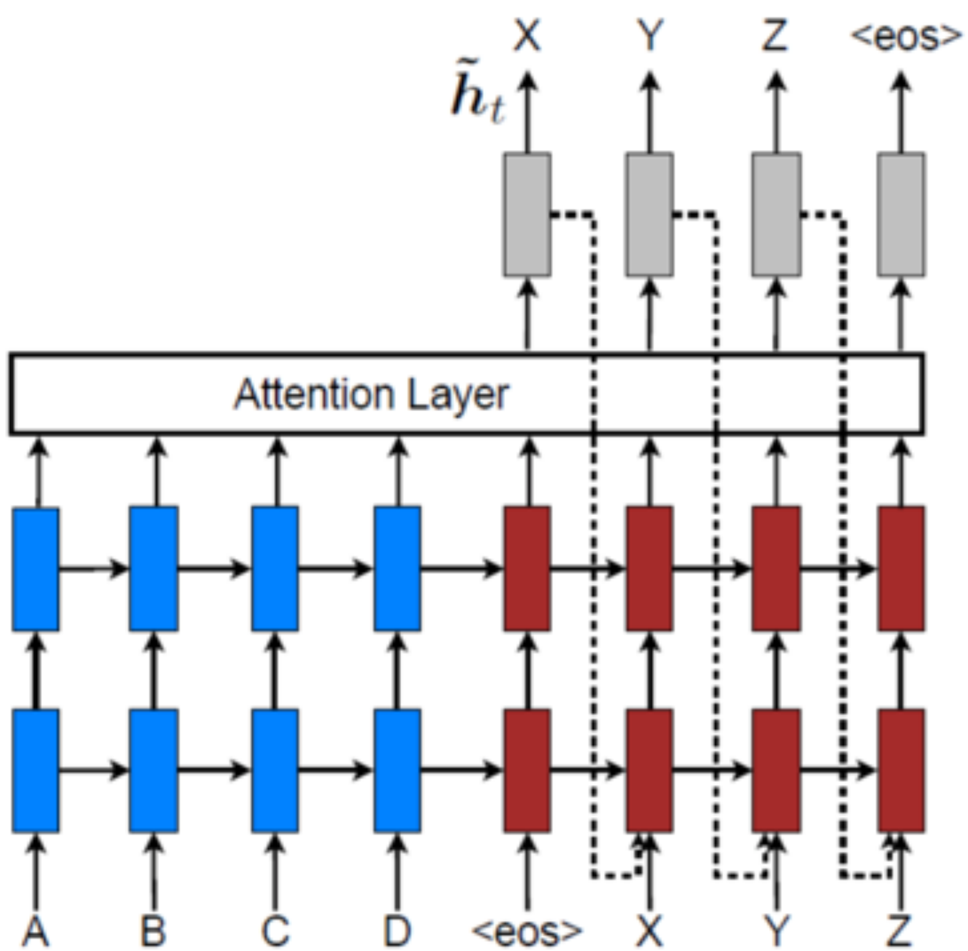
- WMT 2014 (EN→FR)

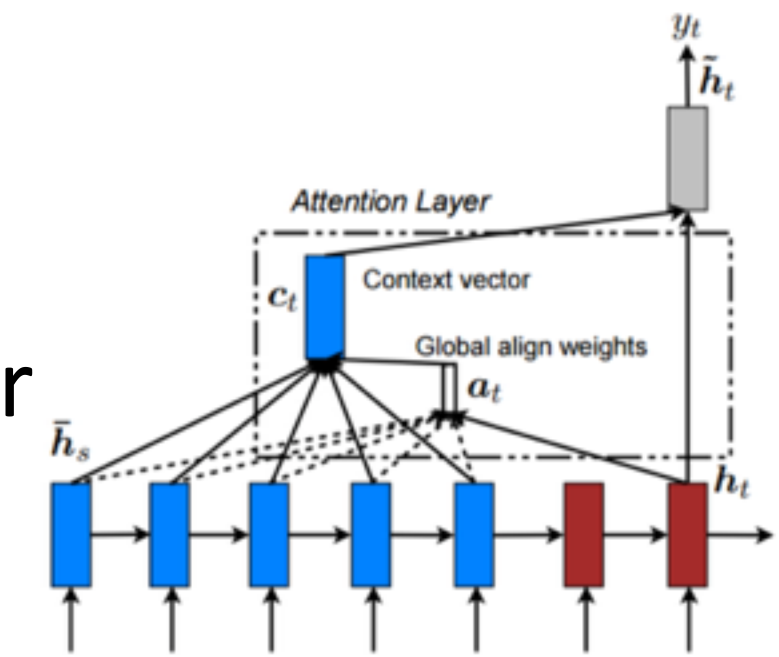
Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63

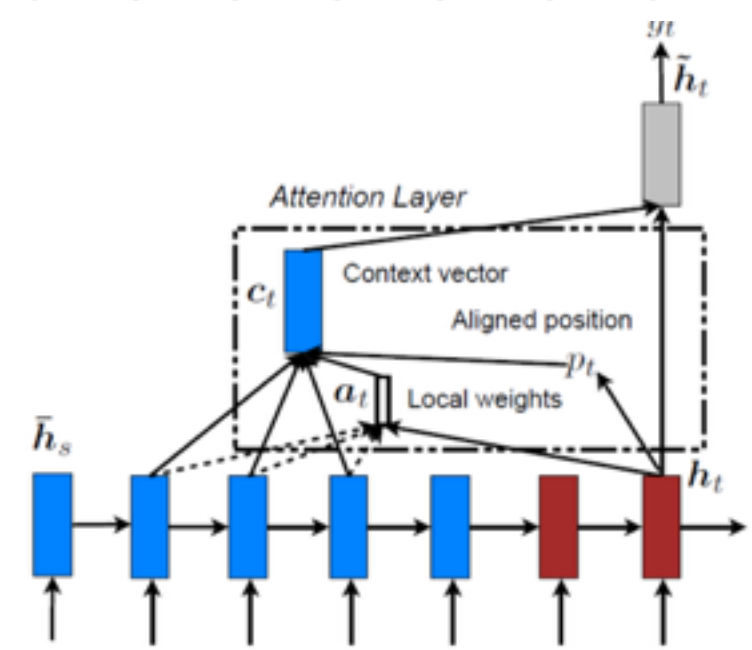
(Bahdanau et al., 2015)

Effective approaches to attention-based NMT

- Global and local attention
- Input-feeding approach
- Stacked LSTM instead of single-layer



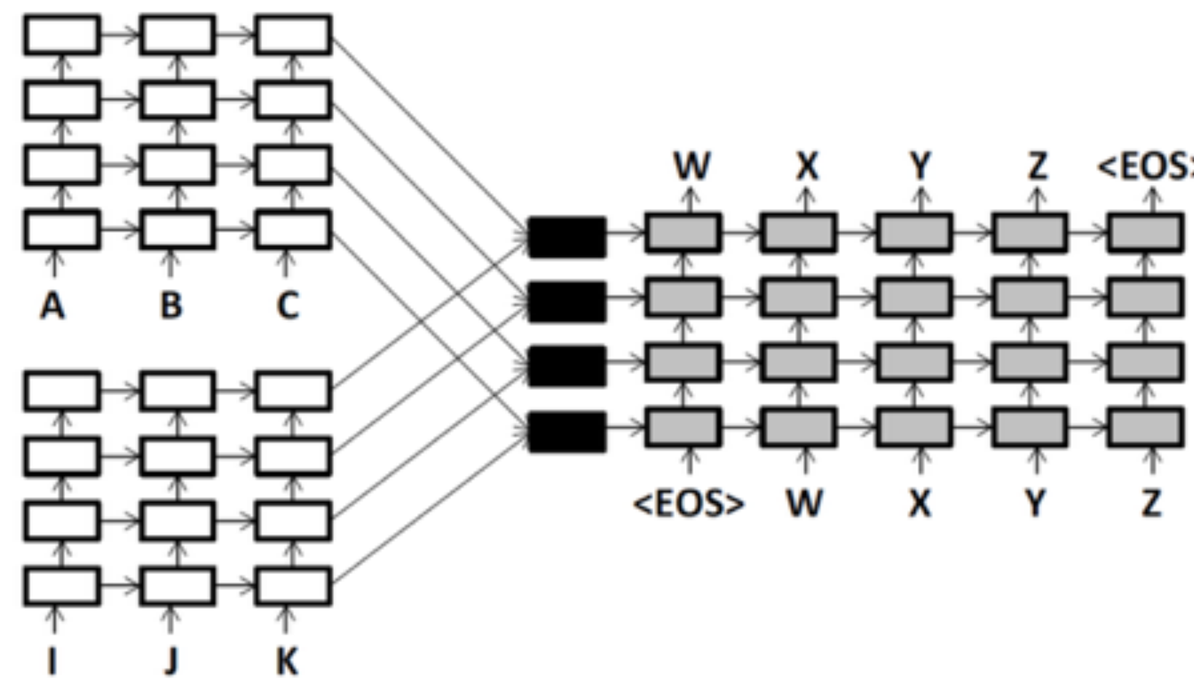


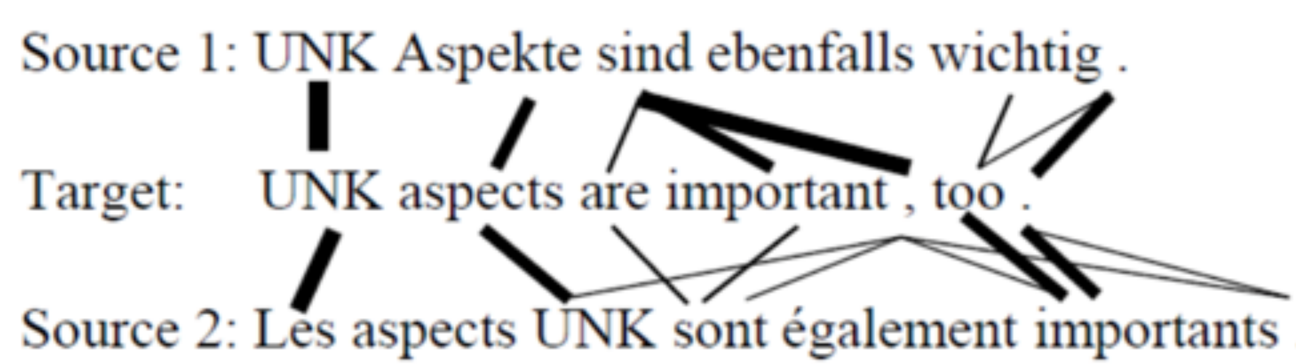


(Luong et al., 2015)

Multi-source NMT

- Train p(e|f, g) model directly on trilingual data
- Use it to decode e given any (f, g) pair
- Take local-attention NMT model and concatenate context from multiple sources





(Zoph and Knight, 2016)

Multi-source NMT

- Multi-source training improves over individual French English and German English pairs
  - **Best**: basic concatenation with attention

Target = English			
Source	**Method**	**Ppl**	**BLEU**
French	—	10.3	21.0
German	—	15.9	17.3
French+German	Basic	8.7	23.2
French+German	Child-Sum	9.0	22.5
French+French	Child-Sum	10.9	20.7
French	Attention	8.1	25.2
French+German	B-Attent.	5.7	30.0
French+German	CS-Attent.	6.0	29.6

Target = German			
Source	**Method**	**Ppl**	**BLEU**
French	—	12.3	10.6
English	—	9.6	13.4
French+English	Basic	9.1	14.5
French+English	Child-Sum	9.5	14.4
English	Attention	7.3	17.6
French+English	B-Attent.	6.9	18.6
French+English	CS-Attent.	7.1	18.2

(Zoph and Knight, 2016)

Multi-source NMT

- Multi-source training improves over individual French English and German English pairs
  - **Best**: basic concatenation with attention

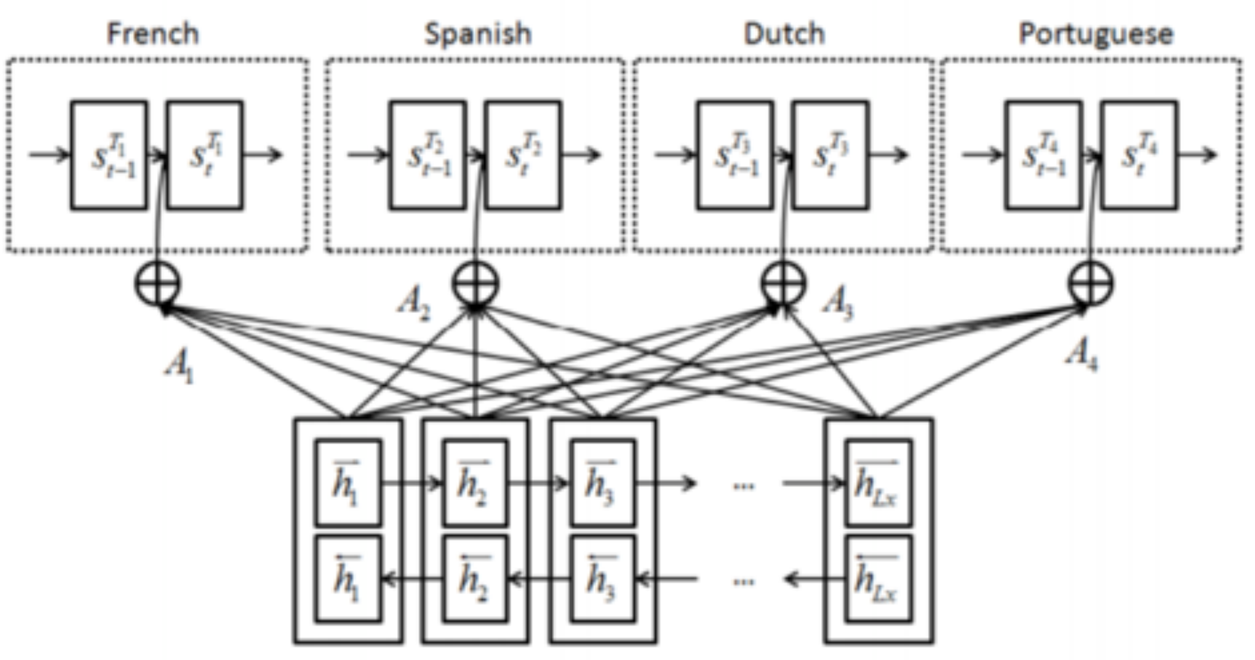
Target = English			
Source	**Method**	**Ppl**	**BLEU**
French	—	10.3	21.0
German	—	15.9	17.3
French+German	Basic	8.7	23.2
French+German	Child-Sum	9.0	22.5
French+French	Child-Sum	10.9	20.7
French	Attention	8.1	25.2
French+German	B-Attent.	5.7	30.0
French+German	CS-Attent.	6.0	29.6

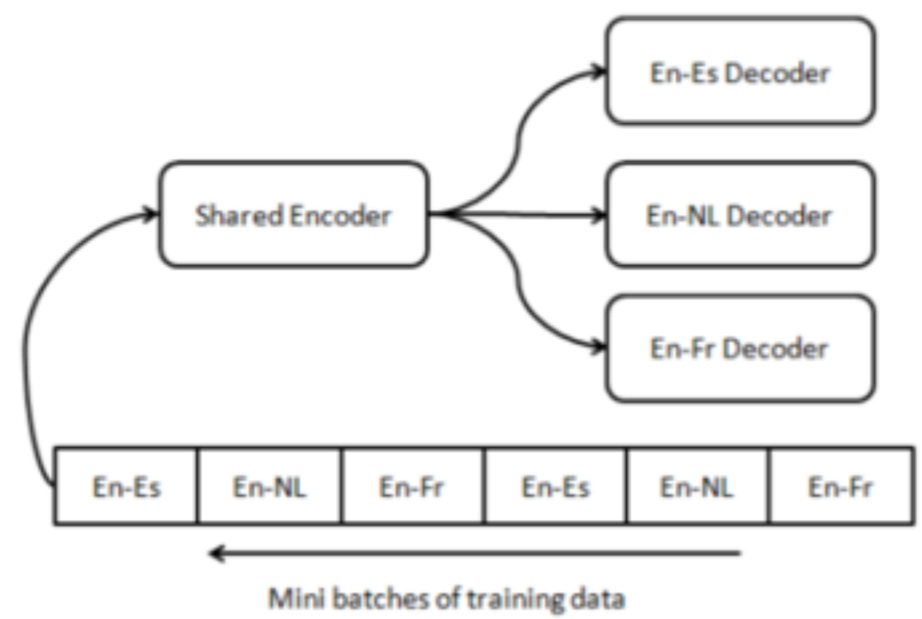
Target = German			
Source	**Method**	**Ppl**	**BLEU**
French	—	12.3	10.6
English	—	9.6	13.4
French+English	Basic	9.1	14.5
French+English	Child-Sum	9.5	14.4
English	Attention	7.3	17.6
French+English	B-Attent.	6.9	18.6
French+English	CS-Attent.	7.1	18.2

(Zoph and Knight, 2016)

Multi-target NMT

- Multi-task learning framework for multiple target language translation
  - Optimization for one to many model



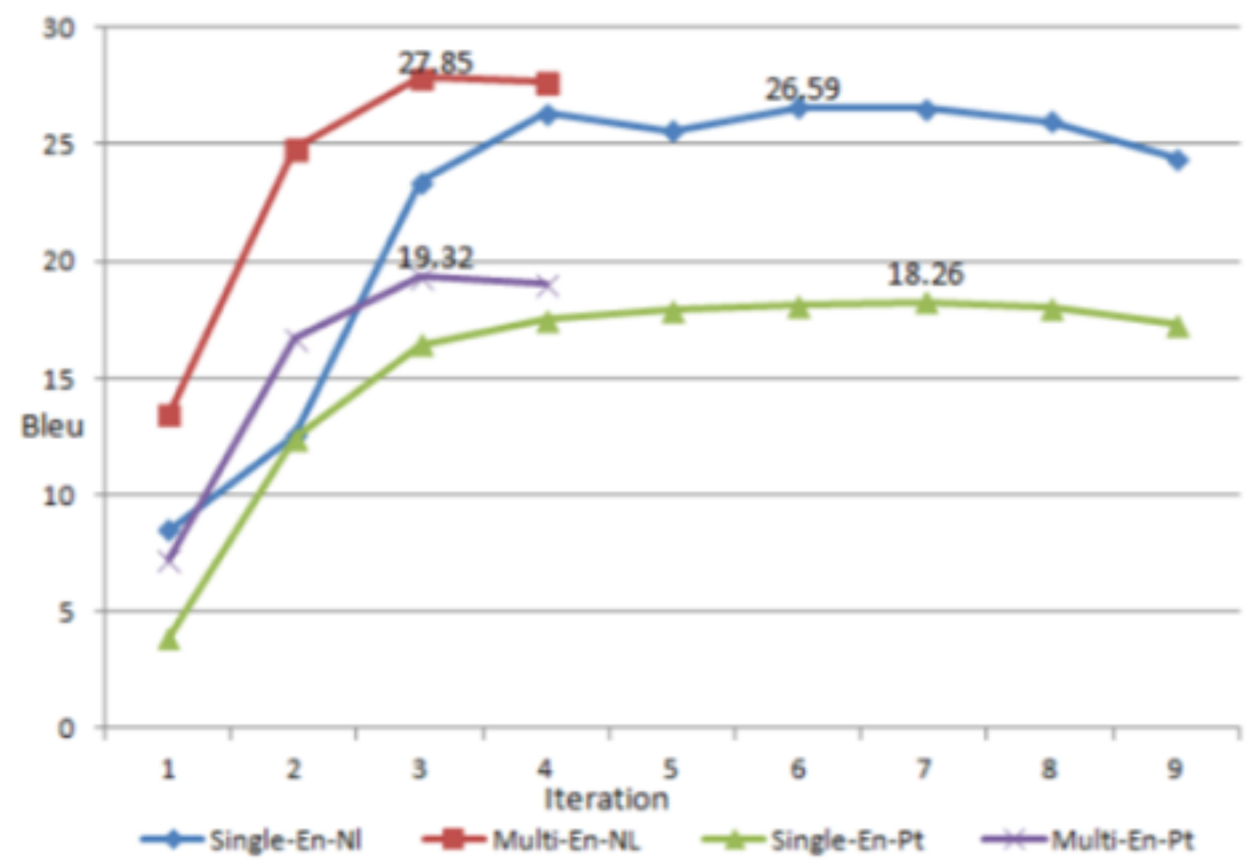


(Dong et al., 2015)

Multi-target NMT

- Improves over NMT and moses baselines over WMT 2013 test
  - but also on larger datasets
- Faster and better convergence in multiple language translation

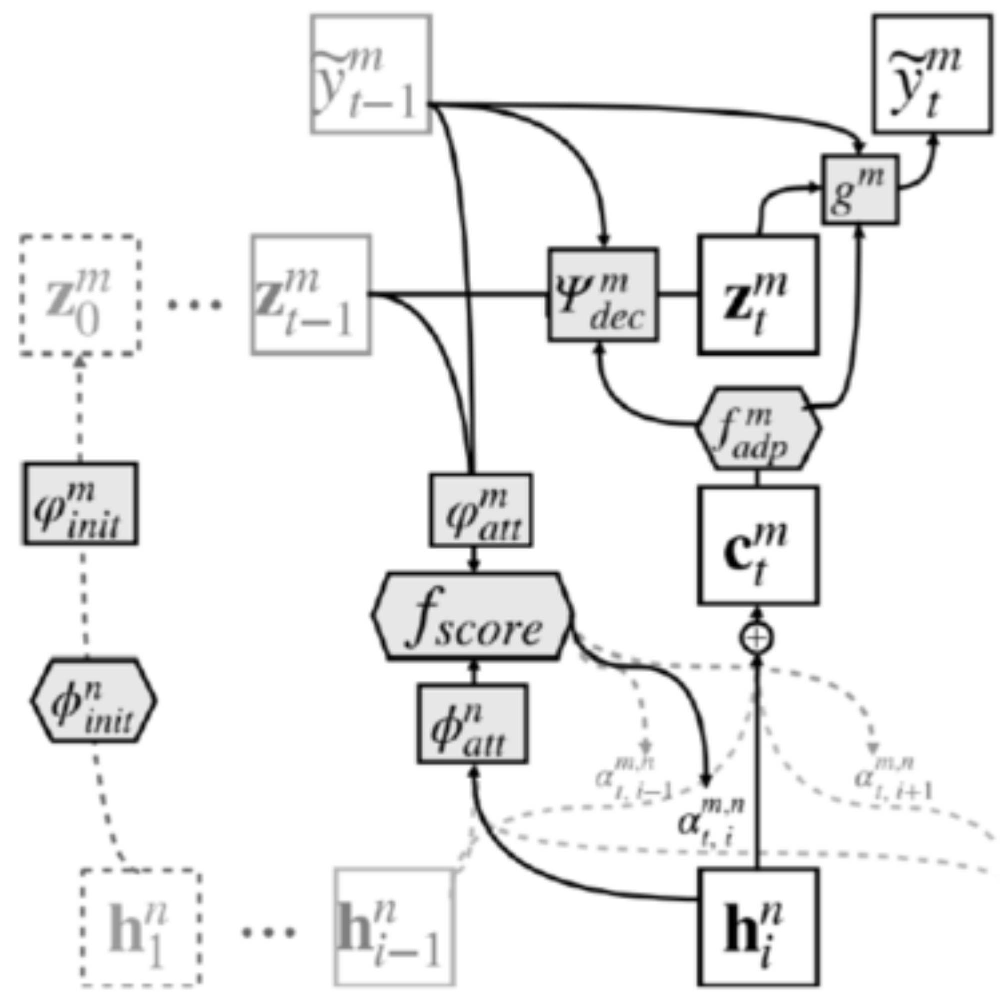
	Nmt Baseline	Nmt Multi-Full	Nmt Multi-Partial	Moses
En-Fr	23.89	26.02(**+2.13**)	25.01(**+1.12**)	23.83
En-Es	23.28	25.31(**+2.03**)	25.83(**+2.55**)	23.58



(Dong et al., 2015)

Multi-way, Multilingual NMT

- Encoder-decoder model with multiple encoders and decoders shared across pairs
  - share knowledge across langs
  - universal space for all langs
  - good for low-resource langs
- Attention is pair specific, **hence expensive** O(L^2)
  - instead share attention across all pairs!



<u>Figure</u>: n_th encoder and m_th decoder at timestep t / ϕ makes encoder & decoder states compatible with the attention mechanism / f_adp makes context vector compatible with the decoder
→ all these transformations to support different types of encoders/decoders for different languages!

(Firat et al., 2016)

Multi-way, Multilingual NMT

	Size	Single	Single+DF	Multi
En→Fi	100k	5.06/3.96	4.98/3.99	6.2/**5.17**
	200k	7.1/6.16	7.21/6.17	8.84/**7.53**
	400k	9.11/7.85	9.31/8.18	11.09/**9.98**
	800k	11.08/9.96	11.59/10.15	12.73/**11.28**
De→En	210k	14.27/13.2	14.65/13.88	16.96/**16.26**
	420k	18.32/17.32	18.51/17.62	19.81/**19.63**
	840k	21/19.93	21.69/20.75	22.17/**21.93**
	1.68m	23.38/23.01	23.33/22.86	23.86/**23.52**
En→De	210k	11.44/11.57	11.71/11.16	12.63/**12.68**
	420k	14.28/14.25	14.88/15.05	15.01/**15.67**
	840k	17.09/17.44	17.21/17.88	17.33/**18.14**
	1.68m	19.09/19.6	19.36/20.13	19.23/**20.59**

- **Consistent improvements for low-resource languages**
    - the lower the training data the bigger the improvement
- **In large-scale translation improves only translation to English**
    - hypothesis: EN appears always as source or target language for all pairs → better decoder ?

		Dir	Fr (39m) → En	Fr (39m) En →	Cs (12m) → En	Cs (12m) En →	De (4.2m) → En	De (4.2m) En →	Ru (2.3m) → En	Ru (2.3m) En →	Fi (2m) → En	Fi (2m) En →
(a) BLEU	Dev	Single	27.22	26.91	21.24	15.9	24.13	20.49	21.04	18.06	13.15	9.59
		Multi	26.09	25.04	21.23	14.42	23.66	19.17	21.48	17.89	12.97	8.92
	Test	Single	27.94	**29.7**	20.32	**13.84**	24	**21.75**	22.44	**19.54**	12.24	**9.23**
		Multi	**28.06**	27.88	**20.57**	13.29	**24.20**	20.59	**23.44**	19.39	**12.61**	8.98
(b) LL	Dev	Single	-50.53	-53.38	-60.69	-69.56	-54.76	-61.21	-60.19	-65.81	-88.44	-91.75
		Multi	-50.6	-56.55	-54.46	-70.76	-54.14	-62.34	-54.09	-63.75	-74.84	-88.02
	Test	Single	-43.34	**-45.07**	-60.03	**-64.34**	-57.81	**-59.55**	-60.65	-60.29	-88.66	-94.23
		Multi	**-42.22**	-46.29	**-54.66**	-64.80	**-53.85**	-60.23	**-54.49**	**-58.63**	**-71.26**	**-88.09**

(Firat et al., 2016)

Multi-way, Multilingual NMT

	Size	Single	Single+DF	Multi
En→Fi	100k	5.06/3.96	4.98/3.99	6.2/**5.17**
	200k	7.1/6.16	7.21/6.17	8.84/**7.53**
	400k	9.11/7.85	9.31/8.18	11.09/**9.98**
	800k	11.08/9.96	11.59/10.15	12.73/**11.28**
De→En	210k	14.27/13.2	14.65/13.88	16.96/**16.26**
	420k	18.32/17.32	18.51/17.62	19.81/**19.63**
	840k	21/19.93	21.69/20.75	22.17/**21.93**
	1.68m	23.38/23.01	23.33/22.86	23.86/**23.52**
En→De	210k	11.44/11.57	11.71/11.16	12.63/**12.68**
	420k	14.28/14.25	14.88/15.05	15.01/**15.67**
	840k	17.09/17.44	17.21/17.88	17.33/**18.14**
	1.68m	19.09/19.6	19.36/20.13	19.23/**20.59**

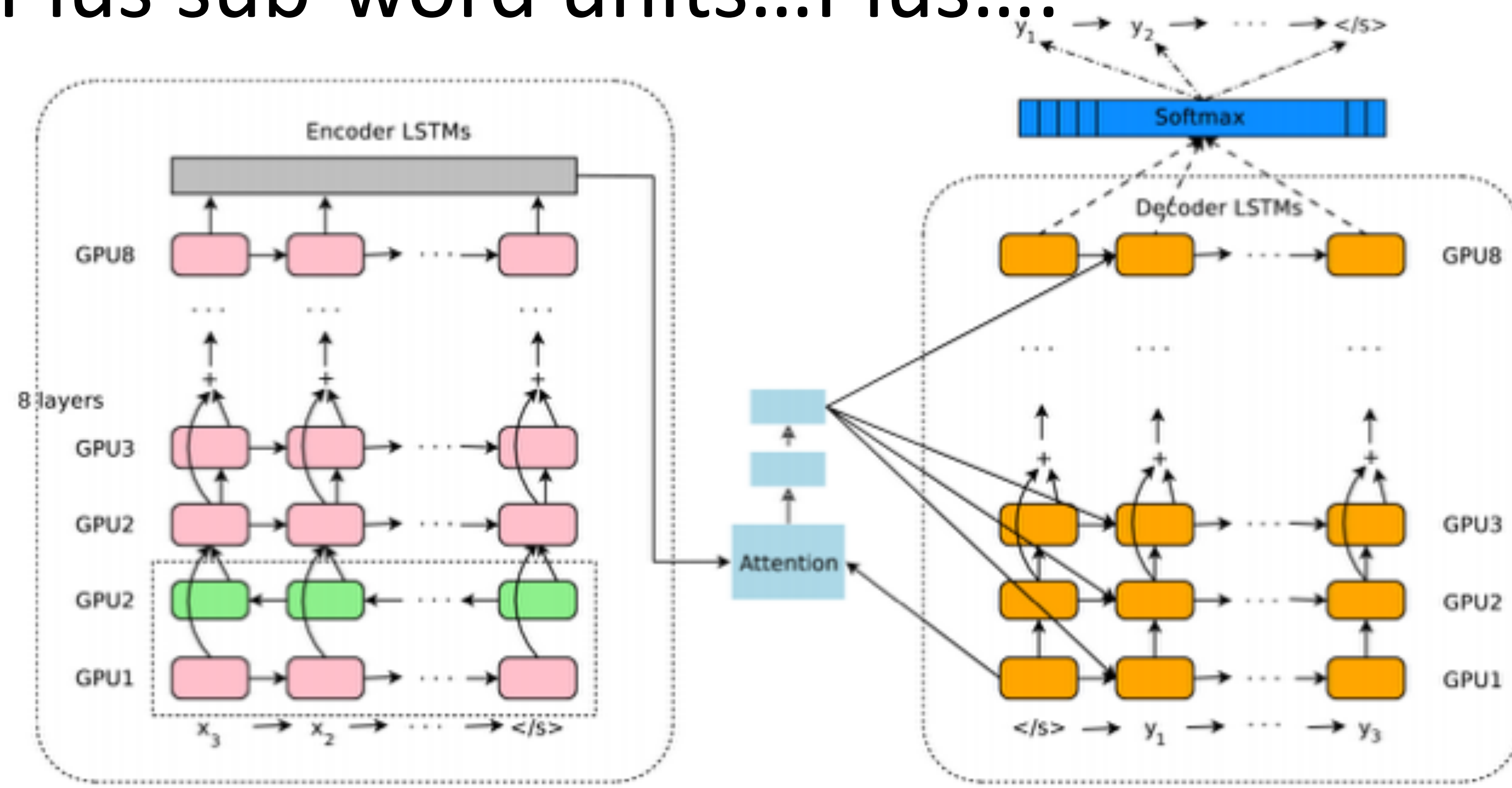
- **Consistent improvements for low-resource languages**
  - the lower the training data the bigger the improvement
- **In large-scale translation improves only translation to English**
  - hypothesis: EN appears always as source or target language for all pairs → better decoder ?

			Fr (39m)		Cs (12m)		De (4.2m)		Ru (2.3m)		Fi (2m)	
		Dir	→ En	En →	→ En	En →	→ En	En →	→ En	En →	→ En	En →
(a) BLEU	Dev	Single	27.22	26.91	21.24	15.9	24.13	20.49	21.04	18.06	13.15	9.59
		Multi	26.09	25.04	21.23	14.42	23.66	19.17	21.48	17.89	12.97	8.92
	Test	Single	27.94	**29.7**	20.32	**13.84**	24	**21.75**	22.44	**19.54**	12.24	**9.23**
		Multi	**28.06**	27.88	**20.57**	13.29	**24.20**	20.59	**23.44**	19.39	**12.61**	8.98
(b) LL	Dev	Single	-50.53	-53.38	-60.69	-69.56	-54.76	-61.21	-60.19	-65.81	-88.44	-91.75
		Multi	-50.6	-56.55	-54.46	-70.76	-54.14	-62.34	-54.09	-63.75	-74.84	-88.02
	Test	Single	-43.34	**-45.07**	-60.03	**-64.34**	-57.81	**-59.55**	-60.65	-60.29	-88.66	-94.23
		Multi	**-42.22**	-46.29	**-54.66**	-64.80	**-53.85**	-60.23	**-54.49**	**-58.63**	**-71.26**	**-88.09**

(Firat et al., 2016)

Google's Neural Machine Translation ~~System~~ "Monster"

- An encoder, a decoder and an attention network
  - Plus 8-layer deep with residual connections
  - Plus refinement with Reinforcement Learning
  - Plus sub-word units…Plus….



(Wu et al., 2016)

Google's Neural Machine Translation ~~System~~ "Monster"

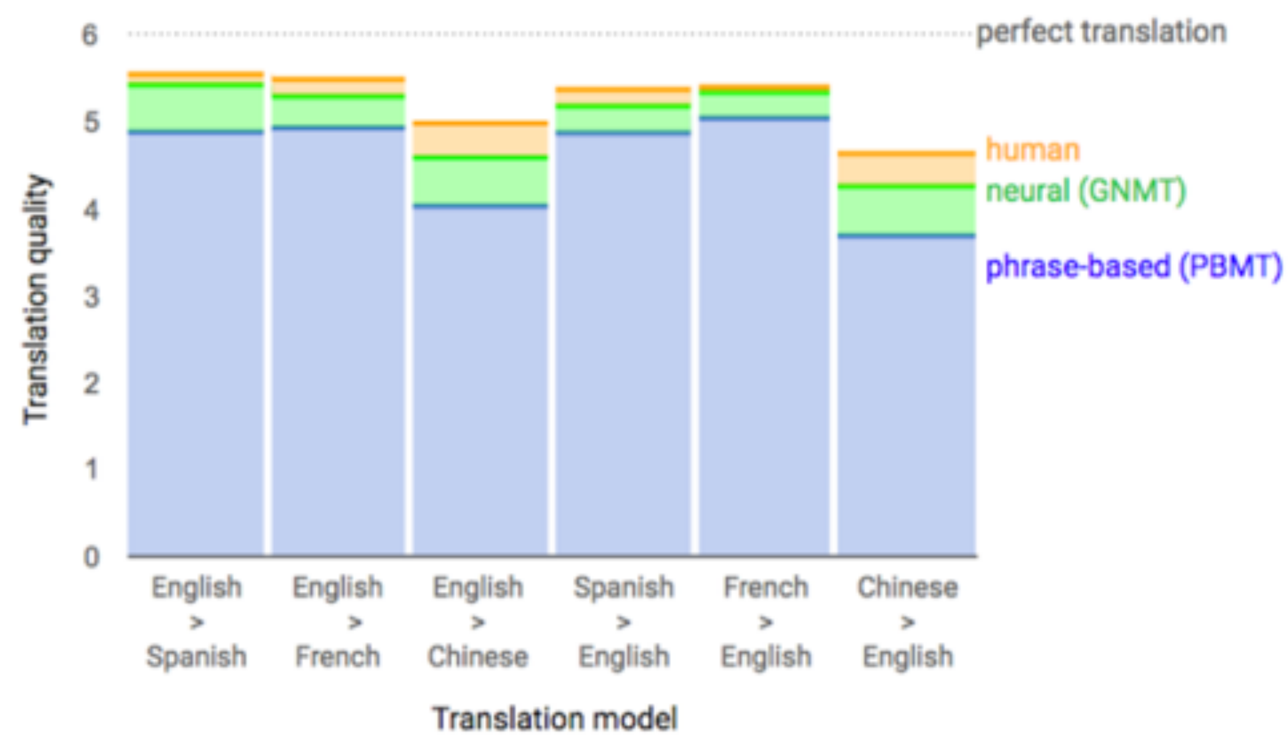
- **EN->FR training takes 6 days on 96GPUS !!!! and 3 more days for refinement...**

Table 7: Model ensemble results on WMT En→Fr (newstest2014)

Model	BLEU
WPM-32K (8 models)	40.35
RL-refined WPM-32K (8 models)	41.16
LSTM (6 layers) [31]	35.6
LSTM (6 layers + PosUnk) [31]	37.5
Deep-Att + PosUnk (8 models) [45]	40.4

Table 5: Single model results on WMT En→De (newstest2014)

Model	BLEU	CPU decoding time per sentence (s)
Word	23.12	0.2972
Character (512 nodes)	22.62	0.8011
WPM-8K	23.50	0.2079
WPM-16K	24.36	0.1931
WPM-32K	24.61	0.1882
Mixed Word/Character	24.17	0.3268
PBMT [6]	20.7	
RNNSearch [37]	16.5	
RNNSearch-LV [37]	16.9	
RNNSearch-LV [37]	16.9	
Deep-Att [45]	20.6	

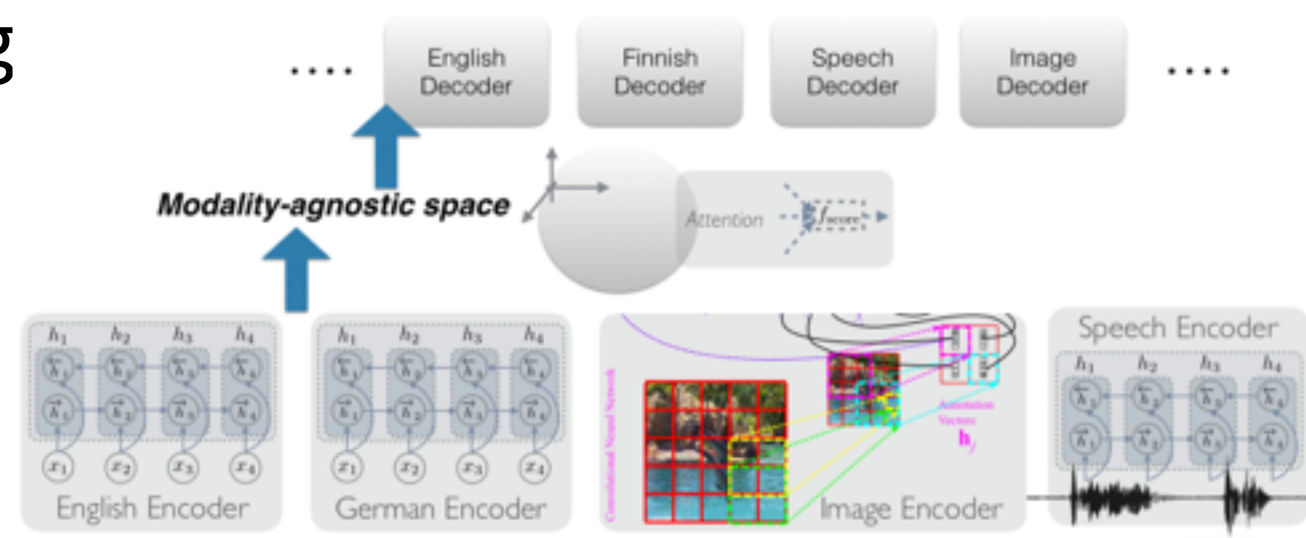


Data from side-by-side evaluations, where human raters compare the quality of translations for a given source sentence. Scores range from 0 to 6, with 0 meaning "completely nonsense translation", and 6 meaning "perfect translation."

(Wu et al., 2016)

Future of NMT and other possibilities

- **Multi-task learning**: Training multiple pairs of languages jointly and with other tasks
  → Image captioning, Speech recognition !



- **Larger context**: Modeling larger sequences than sentences as in document classification **will be key**
  - understanding long-term dependencies
  - leveraging structural information of the input
  - being able to reason over it to solve any task

  → Effective Attention / Memory?

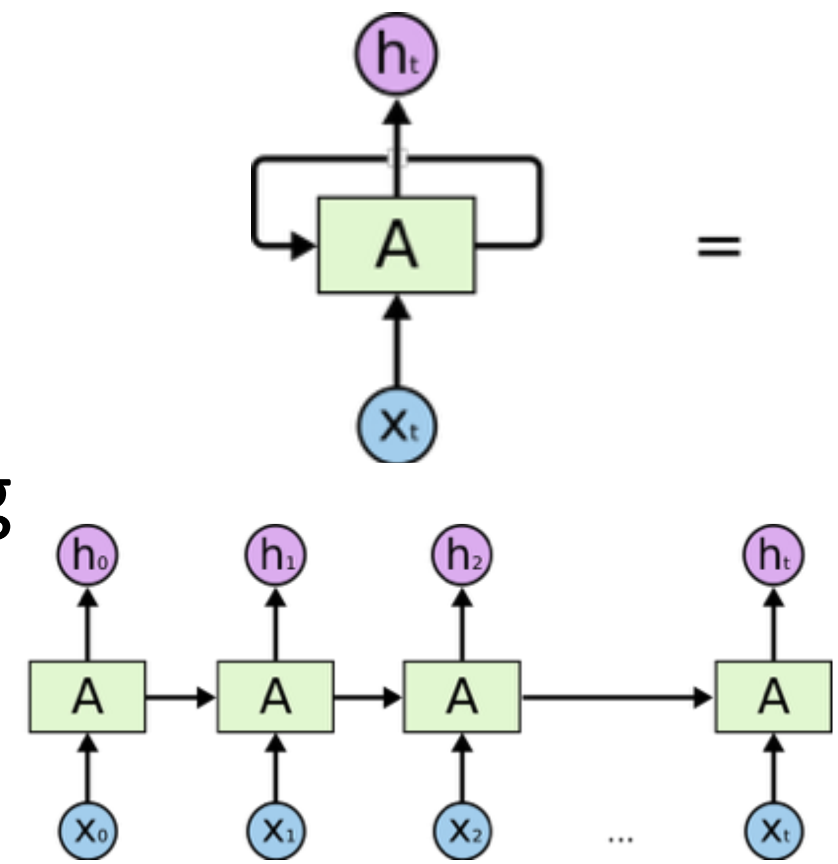
(Luong, Cho, Manning tutorial, 2016)

Outline of the talk

1. Recap: Word Representation Learning

2. Multilingual Word Representations
   - Alignment models
   - Evaluation tasks

3. Multilingual Word Sequence Modeling
   - Essentials: RNN, LSTM, GRU
   - Machine Translation
   - Document Classification

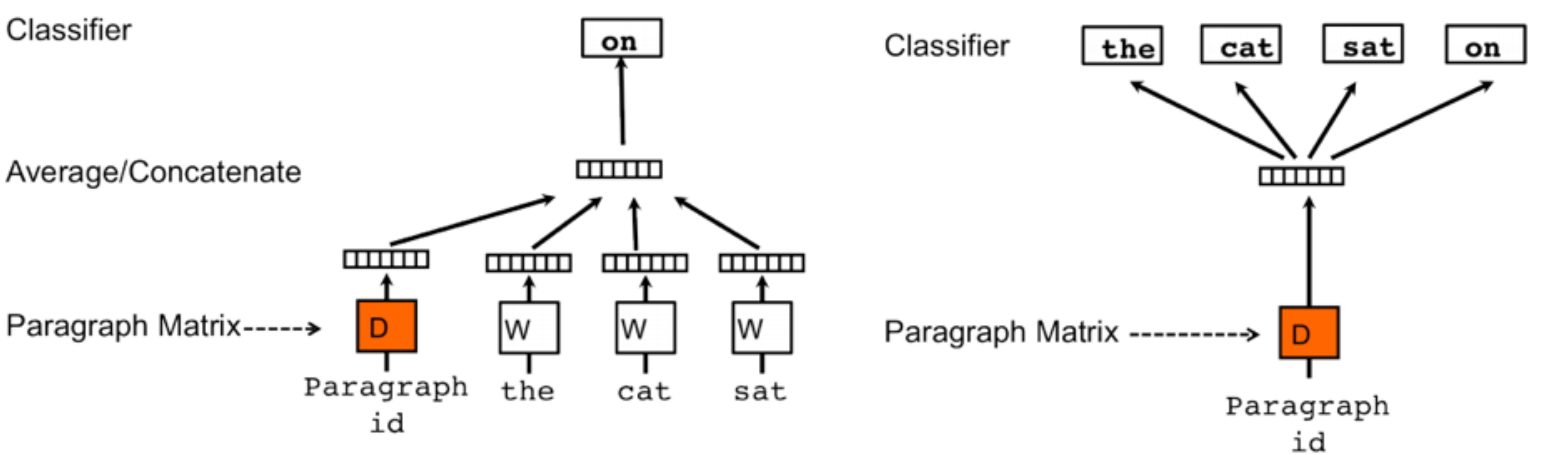


* Figure from Colah's blog, 2015.

4. Summary

Paragraph vectors for Document Classification

- Learning vectors of paragraphs inspired by word2vec
  - trained without supervision on a large corpus
  - preferably similar domain as the target
- **Two methods**: with or without word ordering



(Le et al., 2014)

Paragraph vectors for Document Classification

- Learned paragraph vectors + logistic regression
- Outperformed previous method on sentence-level and document-level sentiment classification

Table 1. The performance of our method compared to other approaches on the Stanford Sentiment Treebank dataset. The error rates of other methods are reported in (Socher et al., 2013b).

Model	Error rate (Positive/ Negative)	Error rate (Fine-grained)
Naïve Bayes (Socher et al., 2013b)	18.2 %	59.0%
SVMs (Socher et al., 2013b)	20.6%	59.3%
Bigram Naïve Bayes (Socher et al., 2013b)	16.9%	58.1%
Word Vector Averaging (Socher et al., 2013b)	19.9%	67.3%
Recursive Neural Network (Socher et al., 2013b)	17.6%	56.8%
Matrix Vector-RNN (Socher et al., 2013b)	17.1%	55.6%
Recursive Neural Tensor Network (Socher et al., 2013b)	14.6%	54.3%
Paragraph Vector	**12.2%**	**51.3%**

Table 2. The performance of Paragraph Vector compared to other approaches on the IMDB dataset. The error rates of other methods are reported in (Wang & Manning, 2012).

Model	Error rate
BoW (bnc) (Maas et al., 2011)	12.20 %
BoW (bΔt'c) (Maas et al., 2011)	11.77%
LDA (Maas et al., 2011)	32.58%
Full+BoW (Maas et al., 2011)	11.67%
Full+Unlabeled+BoW (Maas et al., 2011)	11.11%
WRRBM (Dahl et al., 2012)	12.58%
WRRBM + BoW (bnc) (Dahl et al., 2012)	10.77%
MNB-uni (Wang & Manning, 2012)	16.45%
MNB-bi (Wang & Manning, 2012)	13.41%
SVM-uni (Wang & Manning, 2012)	13.05%
SVM-bi (Wang & Manning, 2012)	10.84%
NBSVM-uni (Wang & Manning, 2012)	11.71%
NBSVM-bi (Wang & Manning, 2012)	8.78%
Paragraph Vector	**7.42%**

(Le et al., 2014)

Convolutional neural network for Document Classification

- Used multiple filter widths
- Dropout regularization (randomly dropping portion of hidden units during back-propagation)

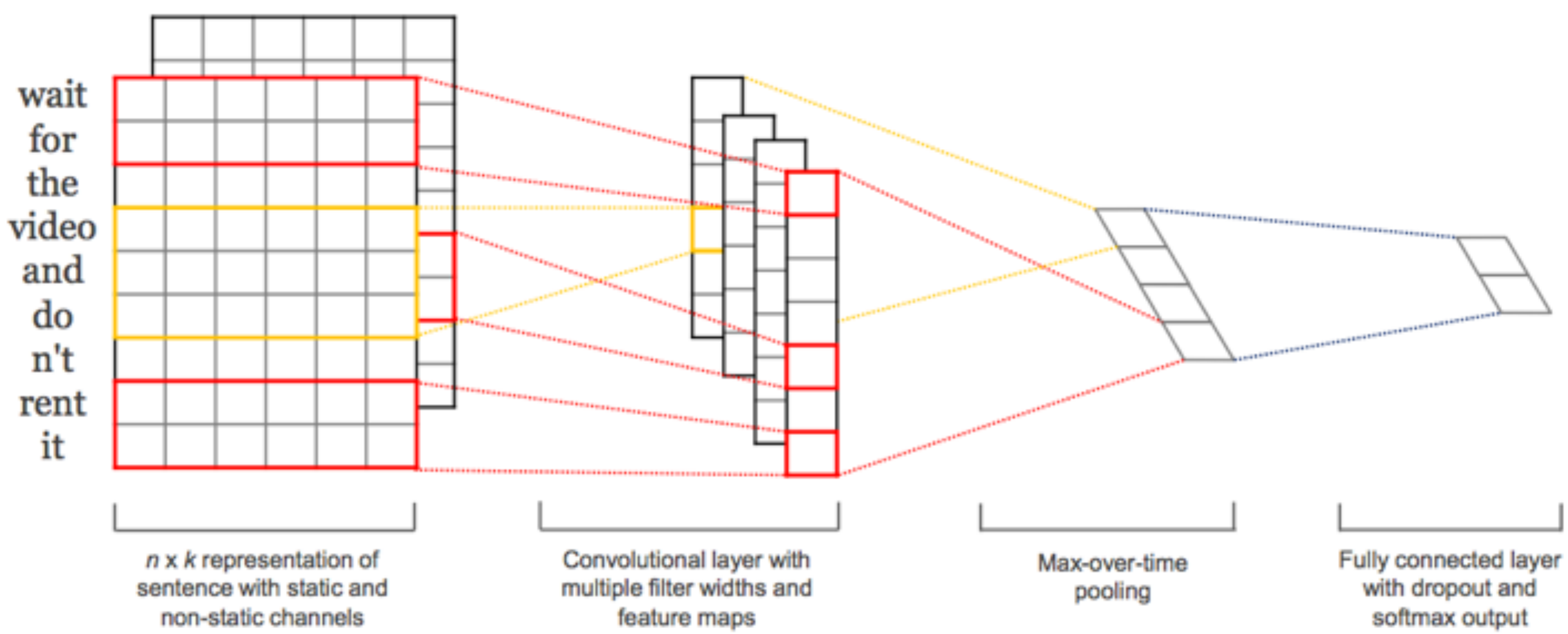


Figure 1: Model architecture with two channels for an example sentence.

(Kim et al., 2014)

Convolutional neural network for Document Classification

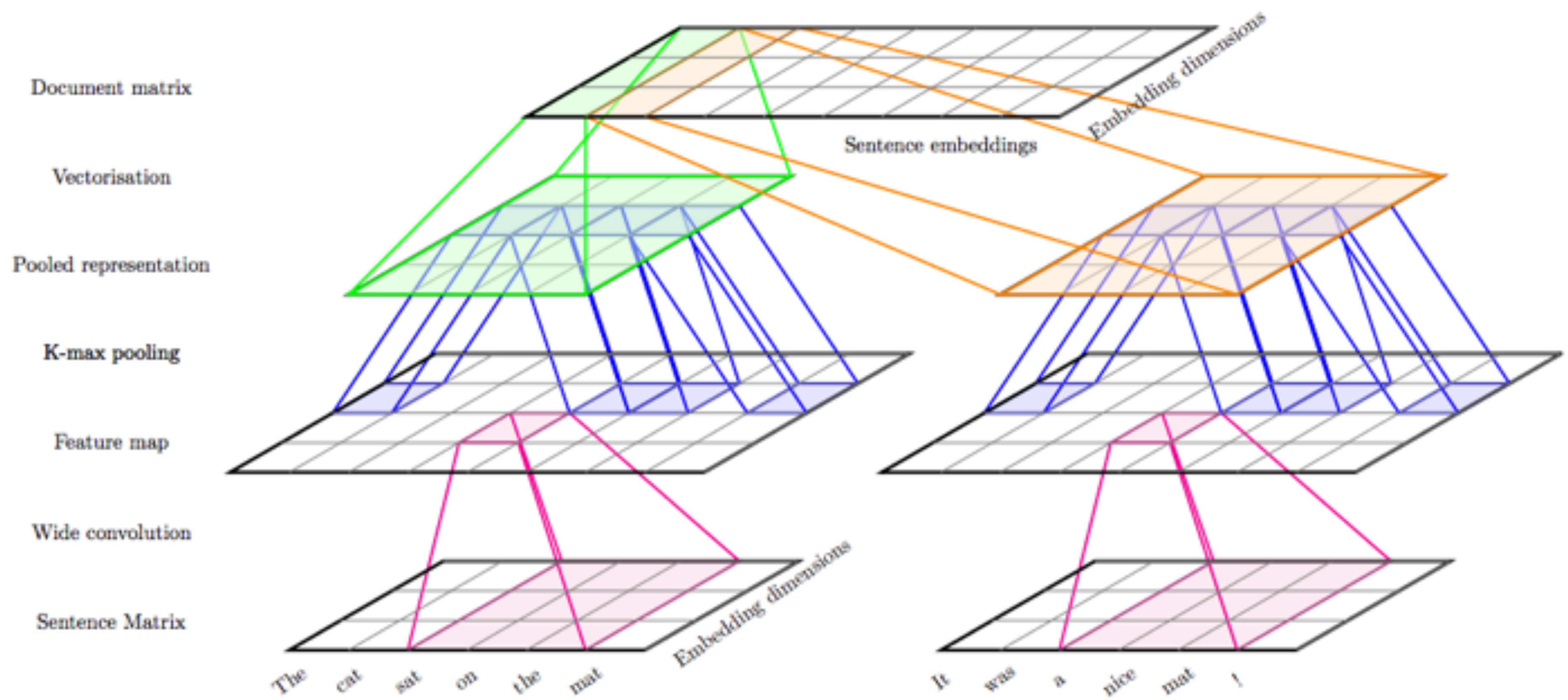
Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	**89.6**
CNN-non-static	**81.5**	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	**85.0**	89.4
RAE (Socher et al., 2011)	77.7	43.2	82.4	–	–	–	86.4
MV-RNN (Socher et al., 2012)	79.0	44.4	82.9	–	–	–	–
RNTN (Socher et al., 2013)	–	45.7	85.4	–	–	–	–
DCNN (Kalchbrenner et al., 2014)	–	48.5	86.8	–	93.0	–	–
Paragraph-Vec (Le and Mikolov, 2014)	–	**48.7**	87.8	–	–	–	–
CCAE (Hermann and Blunsom, 2013)	77.8	–	–	–	–	–	87.2
Sent-Parser (Dong et al., 2014)	79.5	–	–	–	–	–	86.3
NBSVM (Wang and Manning, 2012)	79.4	–	–	93.2	–	81.8	86.3
MNB (Wang and Manning, 2012)	79.0	–	–	**93.6**	–	80.0	86.3
G-Dropout (Wang and Manning, 2013)	79.0	–	–	93.4	–	82.1	86.1
F-Dropout (Wang and Manning, 2013)	79.1	–	–	**93.6**	–	81.9	86.3
Tree-CRF (Nakagawa et al., 2010)	77.3	–	–	–	–	81.4	86.1
CRF-PR (Yang and Cardie, 2014)	–	–	–	–	–	82.7	–
SVM_S (Silva et al., 2011)	–	–	–	–	**95.0**	–	–

- Not all baseline methods used drop-out though

(Kim et al., 2014)

Modeling and Summarizing Documents with a Convolutional Network

- Similar to Kim et al, 2014 however different
  - K-max pooling instead of max pooling
  - Two layers of convolutions



(Denil et al., 2014)

Modeling and Summarizing Documents with a Convolutional Network

Model	Errors
SVM	66
BiNB	62
MaxEnt	61
Max-TDNN	76
NBoW	68
DCNN	45
Our Model	46

Model	Accuracy
BoW (bΔt'c)	88.23%
Full+BoW	88.33%
Full+Unlabelled+BoW	88.89%
WRRBM	87.42%
WRRBM+BoW (bnc)	89.23%
SVM-bi	86.95%
NBSVM-uni	88.29%
NBSVM-bi	91.22%
Paragraph Vector	92.58%
Our model	89.38%

Table 1: **Left:** Number of test set errors on the twitter sentiment dataset. The first block of three entries is from Go *et al.* [5], the second block is from Kalchbrenner *et al.* [13]. **Right:** Error rates on the IMDB movie review data set. The first block is from Maas *et al.* [16], the second from Dahl *et al.* [3], the third from Wang and Manning [24] and the fourth from Le and Mikolov [15].

(Denil et al., 2014)

Modeling and Summarizing Documents with a Convolutional Network

Proportion	Summary	Random	Margin	Fixed	Summary	Random	Margin
100%	83.03	83.03	—				
50%	83.53	79.79	+3.74	Pick 5	83.07	80.02	+3.05
33%	83.10	76.72	+6.38	Pick 4	83.09	79.05	+4.04
25%	82.91	74.87	+8.04	Pick 3	82.88	77.15	+5.73
20%	82.67	73.20	+9.47	Pick 2	82.04	74.48	+7.56
First and last	68.62						

Table 2: Results of classifying summaries with Naïve Bayes. Results labelled proportion indicate selecting up to the indicated percentage of sentences in the review, and results labelled fixed show the result of selecting a fixed number of sentences from each. The summary column shows the accuracy of Naïve Bayes on summaries produced by our model. The random column shows the same model classifying summaries created by selecting sentences at random. The margin column shows the difference in accuracy between our model and the random summaries.

(Denil et al., 2014)

Modeling and Summarizing Documents with a Convolutional Network

Graphics is far from the best part of the game. **This is the number one best TH game in the series.** Next to Underground. **It deserves strong love. It is an insane game.** There are massive levels, massive unlockable characters... it's just a massive game. **Waste your money on this game. This is the kind of money that is wasted properly.** And even though graphics suck, thats doesn't make a game good. Actually, the graphics were good at the time. Today the graphics are crap. WHO CARES? As they say in Canada, This is the fun game, aye. (You get to go to Canada in THPS3) Well, I don't know if they say that, but they might. who knows. Well, Canadian people do. Wait a minute, I'm getting off topic. This game rocks. Buy it, play it, enjoy it, love it. It's PURE BRILLIANCE.

The first was good and original. I was a not bad horror/comedy movie. So I heard a second one was made and I had to watch it . What really makes this movie work is Judd Nelson's character and the sometimes clever script. **A pretty good script for a person who wrote the Final Destination films and the direction was okay.** Sometimes there's scenes where it looks like it was filmed using a home video camera with a grainy - look. Great made - for - TV movie. **It was worth the rental and probably worth buying just to get that nice eerie feeling and watch Judd Nelson's Stanley doing what he does best.** I suggest newcomers to watch the first one before watching the sequel, just so you'll have an idea what Stanley is like and get a little history background.

When the movie was released it was the biggest hit and it soon became the Blockbuster. But honestly the movie is a ridiculous watch with a plot which glorifies a loser. The movie has a Tag - line - "Preeti Madhura, Tyaga Amara" which means Love's Sweet but Sacrifice is Immortal. **In the movie the hero of the movie (Ganesh) sacrifices his love for the leading lady (Pooja Gandhi) even though the two loved each other!** His justification is the meaning of the tag - line. This movie influenced so many young broken hearts that they found this "Loser - like Sacrificial" attitude very thoughtful and hence became the cult movie it is, when they could have moved on with their lives. **Ganesh's acting in the movie is Amateurish, Crass and Childishly stupid.** He actually looks funny in a song, (Onde Ondu Sari ...) when he's supposed to look all stylish and cool. His looks don't help the leading role either. **His hair style is badly done in most part of the movie.** POOJA GANDHI CANT ACT. Her costumes are horrendous in the movie and very inconsistent. **The good part about the movie is the excellent cinematography and brilliant music by Mano Murthy which are actually the true saving graces of the movie.** Also the lyrics by Jayant Kaikini are very well penned. The Director Yograj Bhat has to be lauded picturization the songs in a tasteful manner. Anyway all - in - all except for the songs, the movie is a very ordinary one !!!!!!

A friend and I went through a phase some (alot of) years ago of selecting the crappest horror films in the video shop for an evening's entertainment. For some reason, I ended up buying this one (probably v. v. cheap). **The cheap synth soundtrack is a classic of its time and genre.** There's also a few very amusing scenes. Among them is a scene where a man's being attacked and defends himself with a number of unlikely objects, it made me laugh at the time (doesn't seem quite so funny in retrospect but there you go). **Apart from that it's total crap, mind you.** But probably worth a watch if you like films like "Chopping Mall". Yes, I've seen that too.

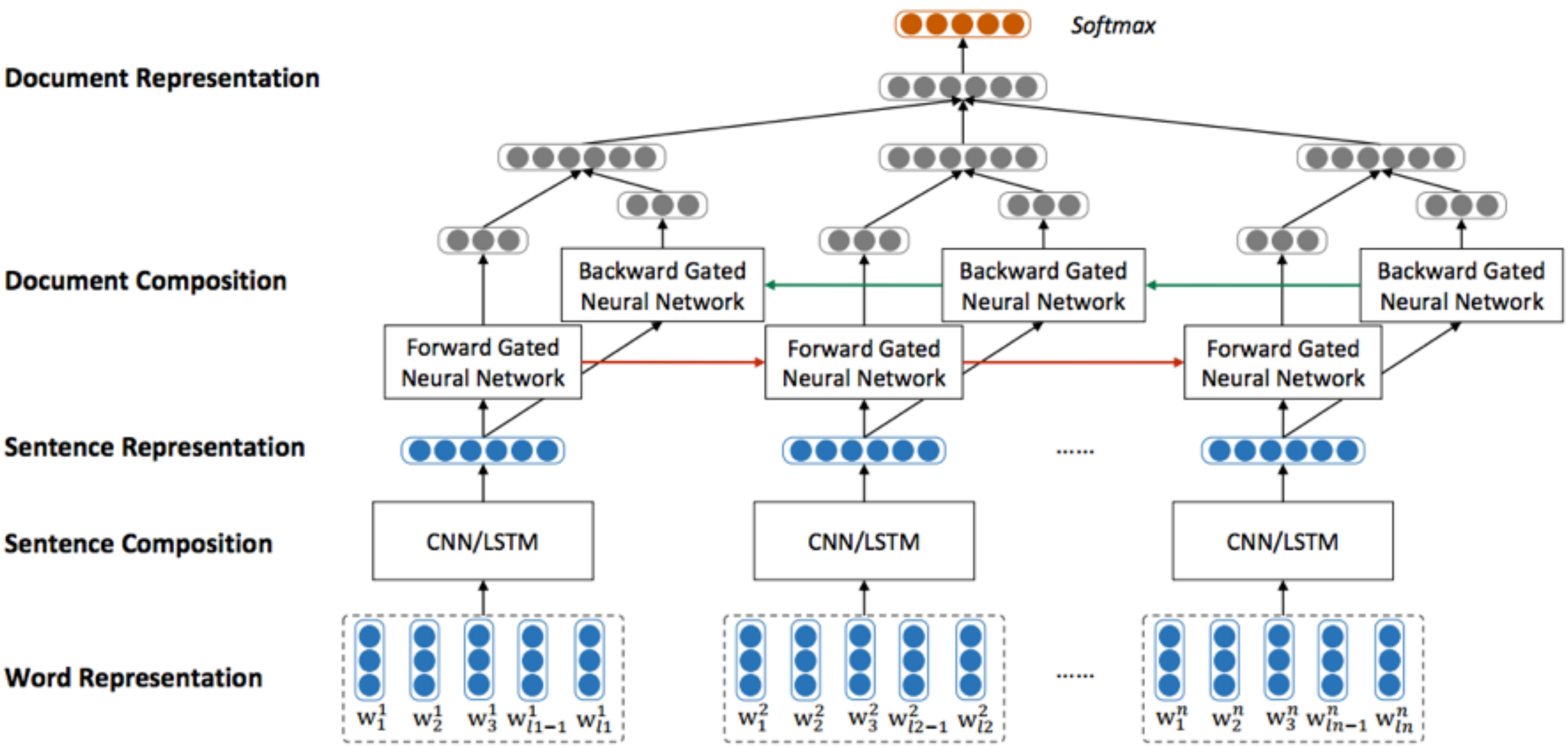
I tried restarting the movie twice. I put it in three machines to see what was wrong . Did Steven Seagal's voice change? **Did he die during filming and the studio have to dub the sound with someone who doesn't even resemble him?** Or was the sound on the DVD destroyed? After about 10 minutes, you finally hear the actor's real voice. Though throughout most of the film, it sounds like the audio was recorded in a bathroom. **I would be ashamed to donate a copy of this movie to Goodwill, if I owned a copy.** I rented it, but I will never do that again. I will check this database before renting any more of his movies, all of which were (more or less) good movies. **You usually knew what you were getting when you watched a Steven Seagal movie.** I guess that is no more.

Vertigo co - stars Stewart (in his last turn as a romantic lead) and Novak elevate this, Stewart's other "Christmas movie," movie to above mid - level entertainment. **The chemistry between the two stars makes for a fairly moving experience and further revelation can be gleaned from the movie if witchcraft is seen as a metaphor for the private pain that hampers many people's relationships.** All in all, a nice diversion with legendary stars, 7/10

Figure 3: Several example summaries created by our ConvNet. The full text of the review is shown in black and the sentences selected by the ConvNet appear in colour. While summarising a review with the first sentence is a popular pragmatic approach, it is clear in these examples that this heuristic is not as effective as the ConvNet summarisation scheme. Each summary is created by selecting up to 20% of the sentences in the review.

il et al., 2014)

Gated recurrent neural network for Document Classification



(Tang et al., 2015)

Gated recurrent neural network for Document Classification

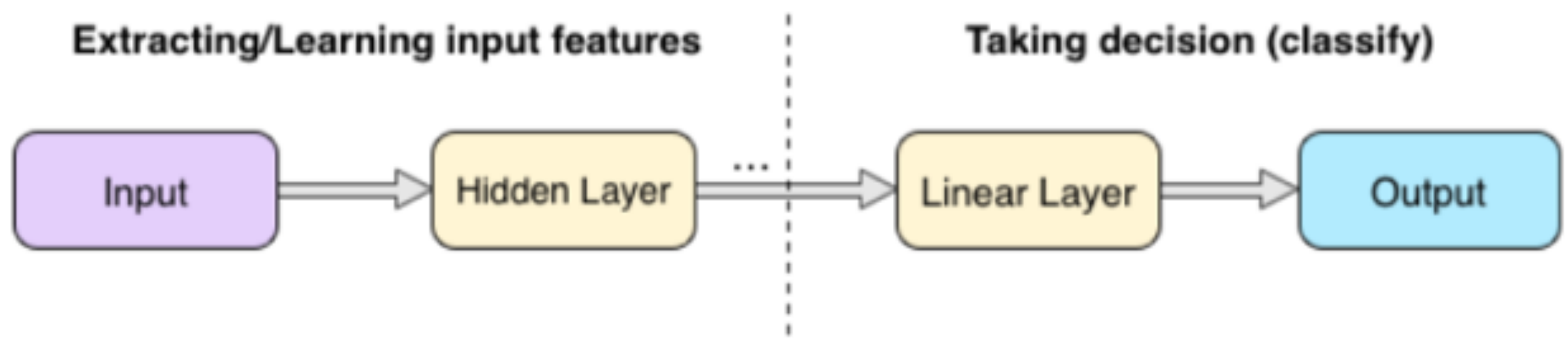
	Yelp 2013		Yelp 2014		Yelp 2015		IMDB	
	Accuracy	MSE	Accuracy	MSE	Accuracy	MSE	Accuracy	MSE
Majority	0.356	3.06	0.361	3.28	0.369	3.30	0.179	17.46
SVM + Unigrams	0.589	0.79	0.600	0.78	0.611	0.75	0.399	4.23
SVM + Bigrams	0.576	0.75	0.616	0.65	0.624	0.63	0.409	3.74
SVM + TextFeatures	0.598	0.68	0.618	0.63	0.624	0.60	0.405	3.56
SVM + AverageSG	0.543	1.11	0.557	1.08	0.568	1.04	0.319	5.57
SVM + SSWE	0.535	1.12	0.543	1.13	0.554	1.11	0.262	9.16
JMARS	N/A	–	N/A	–	N/A	–	N/A	4.97
Paragraph Vector	0.577	0.86	0.592	0.70	0.605	0.61	0.341	4.69
Convolutional NN	0.597	0.76	0.610	0.68	0.615	0.68	0.376	3.30
Conv-GRNN	0.637	0.56	0.655	0.51	0.660	0.50	0.425	**2.71**
LSTM-GRNN	**0.651**	**0.50**	**0.671**	**0.48**	**0.676**	**0.49**	**0.453**	3.00

Table 2: Sentiment classification on Yelp 2013/2014/2015 and IMDB datasets. Evaluation metrics are accuracy (higher is better) and MSE (lower is better). The best method in each setting is in **bold**.

(Tang et al., 2015)

Standard Pipeline for Document Classification

- **Feature engineering**: BOW, n-grams, topic models, etc.
- **Feature learning**: auto-encoders, convolutional, recurrent, recursive NNs



Limitations

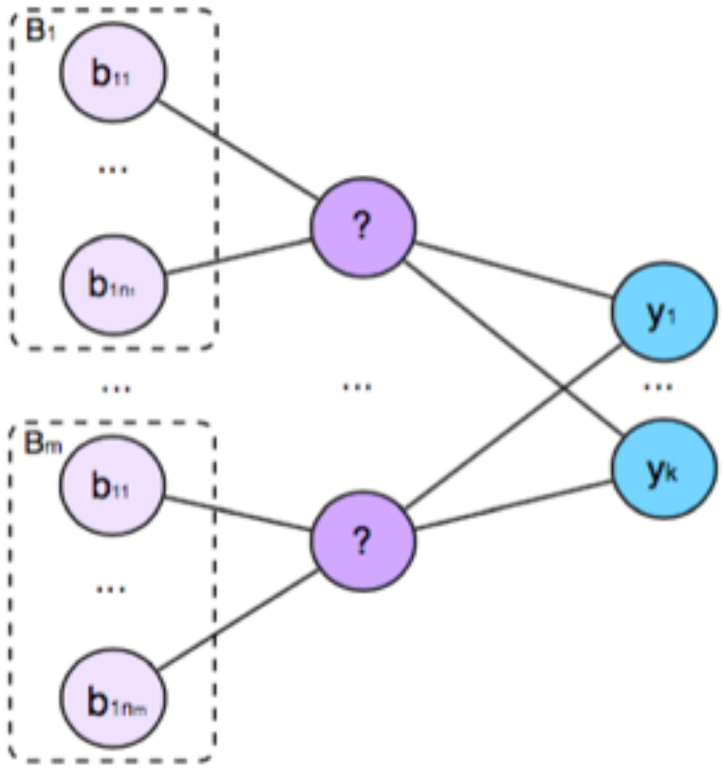
→ Treat the text globally and ignore the weak nature of labels

→ Make simplistic assumptions when aggreagating or pooling features

→ Offer few means for model interpretation

(Pappas and Popescu-Belis, 2014)

Multiple-instance Learning for Document Classification



Given $\mathcal{D} = \{(b_{ij}, y_i) \mid j = 1 \dots n_i\}^m$,
find $\Phi_k : \mathcal{B} \overset{?}{\to} \mathcal{X} \to \mathcal{Y}_k$

- The bag B_i is a review represented by n_i instances b_{ij}, its sentences
- The labels $y_i \in \mathbb{R}^k$ are the aspect ratings of the review
- The exemplar (representation) $x_i \in \mathbb{R}^d$ of B_i is initially unknown

Advantages

$\rightarrow$ Several input assumptions (Aggregated, Instance, Prime, Clustering)
$\rightarrow$ Subsumes traditional supervised regression (Aggregated)
$\rightarrow$ Better suited for weak labels, interpretable and flexible

(Pappas and Popescu-Belis, 2014)

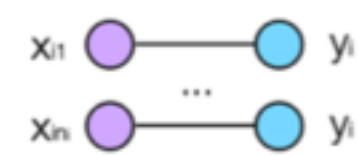
How to combine vectors? Structural assumptions

1. **Aggregated instances**: sum or average instances

$$f \leftarrow D_{agg} = \{(x_i, y_i) \mid i = 1, \ldots, m\}$$
$$\hat{y}(B_i) = f(x_i) = f(mean(\{b_{ij} \mid wj = 1, \ldots, n_i\}))$$

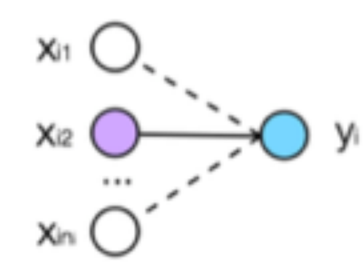
2. **Instance-as-example**: instances inherit bag labels

$$f \leftarrow D_{ins} = \{(b_{ij}, y_i) \mid j = 1, \ldots, n_i; i = 1, \ldots, m\}$$
$$\hat{y}(B_i) = mean(\{f(b_{ij}) \mid j = 1, \ldots, n_i\})$$



3. **Prime instance**: a single instance is selected

$$f \leftarrow D_{pri} = \{(b_i^p, y_i) \mid i = 1, \ldots, m\}$$
$$\hat{y}(B_i) = mean(\{f(b_{ij}) \mid j = 1, \ldots, n_i\})$$

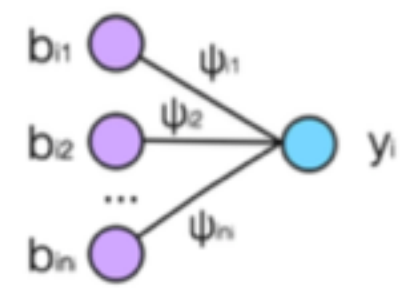


(Pappas and Popescu-Belis, 2014)

Joint learning of an instance relevance mechanism and a classifier

Inspired from method proposed by Wagstaff and Lane (2007):

$$x_i = \sum_{j=1}^{n_i} \psi_{ij} b_{ij}, \ \psi_{ij} \geq 0 \ \text{ and } \sum_{j=1}^{n_i} \psi_{ij} = 1$$



1. Models both instance weights and target labels
   - Target labels model: $\hat{y}_i = f(\Phi, B_i) = \Phi^T(B_i\psi_i)$
   - Instance weights model: $\hat{\psi}_i = f(O, B_i)O^T B_i$
2. Defines loss based on regularized least squares
   - Supports large datasets and high dimensionality $\mathcal{O}(md^2)$
   - Adapts to domain data through regularization

(Pappas and Popescu-Belis, 2014)

Joint differentiable objective for solving with SGD

Based on stochastic gradient descent

$$\sigma(B_i, O) = P(\psi = y_i | B_i) = \frac{\exp^{(O^T B_i)}}{\sum_{k=1}^{n_i} \exp^{(O^T B_{ik})}}$$

$$O, \Phi = \underset{O,\Phi}{arg\ min} \sum_{i=1}^{\cdots} (y_i - \Phi^T(B_i \cdot \sigma(B_i, O)))^2 + \Omega(\Phi, O)$$

- Preserves constraints of instance relevance assumption
- Achieves similar performance to alternating projections
- Makes the learning procedure more scalable

Shared material

→ Code: wmil, wmil-sgd

https://github.com/nik0spapp/

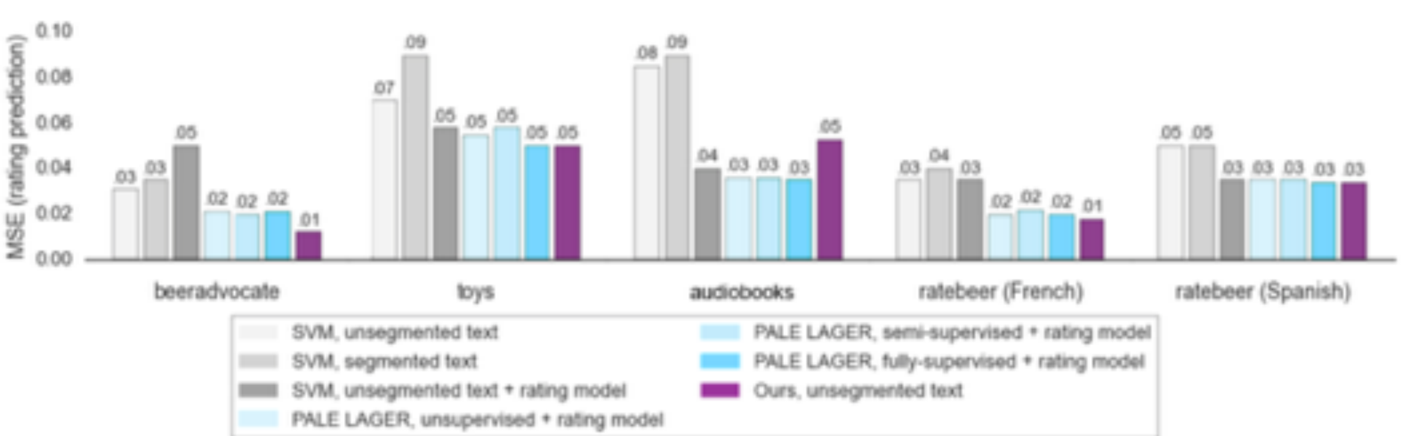
(Pappas and Popescu-Belis, 2014)

Observations on aspect rating prediction

Model \ Error	BOW		TF-IDF		word2vec	
	MAE	MSE	MAE	MSE	MAE	MSE
Aggregated (ℓ_1)	17.08	4.17	16.59	3.97	16.03	3.84
Aggregated (ℓ_2)	16.88	4.47	16.25	4.16	14.62	3.30
Instance (ℓ_1)	17.69	4.37	18.11	4.50	16.37	3.86
Instance (ℓ_2)	16.93	4.24	16.88	4.23	15.60	3.67
Prime (ℓ_1)	17.39	4.37	17.72	4.43	16.13	3.89
Prime (ℓ_2)	18.03	4.91	17.10	4.29	15.71	3.72
Ours (ℓ_2)	**15.97**	**3.97**	**15.36**	**3.63**	**14.25**	**3.29**

Mean Squared Error x 100 (%)

Methods	beeradvocate	toys	audible	ratebeer-fr	ratebeer-sp
Aggregated MIR	3.68	5.93	2.70	5.99	3.41
Instance MIR	3.28	6.59	2.40	6.04	3.39
Prime MIR	3.64	6.92	2.98	6.59	3.68
Clustering MIR	3.26	6.52	2.60	6.48	3.64
Weighted MIR	**2.66**	**5.57**	**2.27**	**5.71**	**3.28**



- The proposed mechanism is superior than alternatives
  - all text regions are useful but to a different extent
- Benefit regardless of the input features used
- Reaches state-of-the-art without using:
  - structured output learning
  - segmented text

(Pappas and Popescu-Belis, 2014)

Comparison with neural network models

- This mechanism can be used as a parametric pooling function of NNs
  - operating on intermediate hidden states
- Works better than Dense, GRU neural methods + average pooling
- Outperforms RCNN and uses far less parameters

Methods	**Vocabulary**	d_{hidden}	**Depth**	$\lvert\theta\rvert$	**MSE**
SVM (Lei et al., 2016)	bigram (>147k)	-	-	2.5M	0.0154
MIR (this work)	unigram (19k)	-	-	38k	**0.0115**
Dense (Rumelhart et al., 1986)	unigram (19k)	200	1	41.2k	0.0101
LSTM (Hochreiter et al, 1997)	unigram (147k)	200	2	644k	0.0094
GRU (Chung et al., 2014)	unigram (19k)	200	1	241.6k	**0.0079**
RCNN (Lei et al., 2016)	unigram (147k)	200	2	323k	0.0087
Dense+MIR (this work)	unigram (19k)	200	1	41.4k	0.0091
GRU+MIR (this work)	unigram (19k)	200	1	241.8k	**0.0078**

Table 2: Comparison of our instance relevance mechanism (MIR) integrated within neural networks, with state-of-the-art neural networks, on the aspect rating prediction task in terms of mean squared error (MSE). $|\theta|$ indicates the number of parameters.

(Pappas and Popescu-Belis, 2016)

Hierarchical attention networks for Document Classification

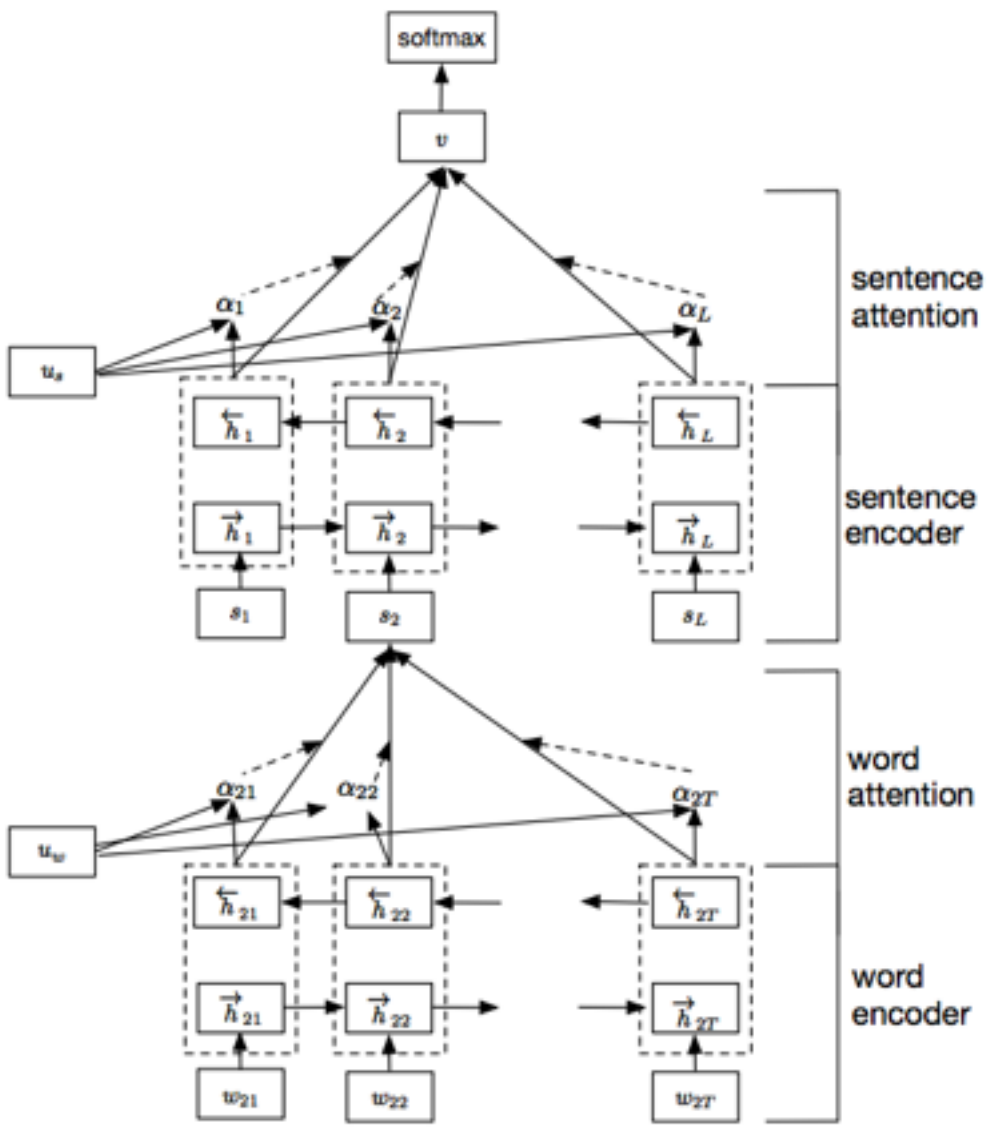


Figure 2: Hierarchical Attention Network.

- Very similar hierarchical structure as Tang et al., 2015 except average pooling
  - attention mechanism at the word and document levels

$$u_{it} = \tanh(W_w h_{it} + b_w)$$
$$\alpha_{it} = \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)}$$
$$s_i = \sum_t \alpha_{it} h_{it}.$$

(Yang et al., 2016)

Hierarchical attention networks for Document Classification

	Methods	Yelp'13	Yelp'14	Yelp'15	IMDB	Yahoo Answer	Amazon
Zhang et al., 2015	BoW	-	-	58.0	-	68.9	54.4
	BoW TFIDF	-	-	59.9	-	71.0	55.3
	ngrams	-	-	56.3	-	68.5	54.3
	ngrams TFIDF	-	-	54.8	-	68.5	52.4
	Bag-of-means	-	-	52.5	-	60.5	44.1
Tang et al., 2015	Majority	35.6	36.1	36.9	17.9	-	-
	SVM + Unigrams	58.9	60.0	61.1	39.9	-	-
	SVM + Bigrams	57.6	61.6	62.4	40.9	-	-
	SVM + TextFeatures	59.8	61.8	62.4	40.5	-	-
	SVM + AverageSG	54.3	55.7	56.8	31.9	-	-
	SVM + SSWE	53.5	54.3	55.4	26.2	-	-
Zhang et al., 2015	LSTM	-	-	58.2	-	70.8	59.4
	CNN-char	-	-	62.0	-	71.2	59.6
	CNN-word	-	-	60.5	-	71.2	57.6
Tang et al., 2015	Paragraph Vector	57.7	59.2	60.5	34.1	-	-
	CNN-word	59.7	61.0	61.5	37.6	-	-
	Conv-GRNN	63.7	65.5	66.0	42.5	-	-
	LSTM-GRNN	65.1	67.1	67.6	45.3	-	-
This paper	HN-AVE	67.0	69.3	69.9	47.8	75.2	62.9
	HN-MAX	66.9	69.3	70.1	48.2	75.2	62.9
	HN-ATT	**68.2**	**70.5**	**71.0**	**49.4**	**75.8**	**63.6**

Table 2: Document Classification, in percentage

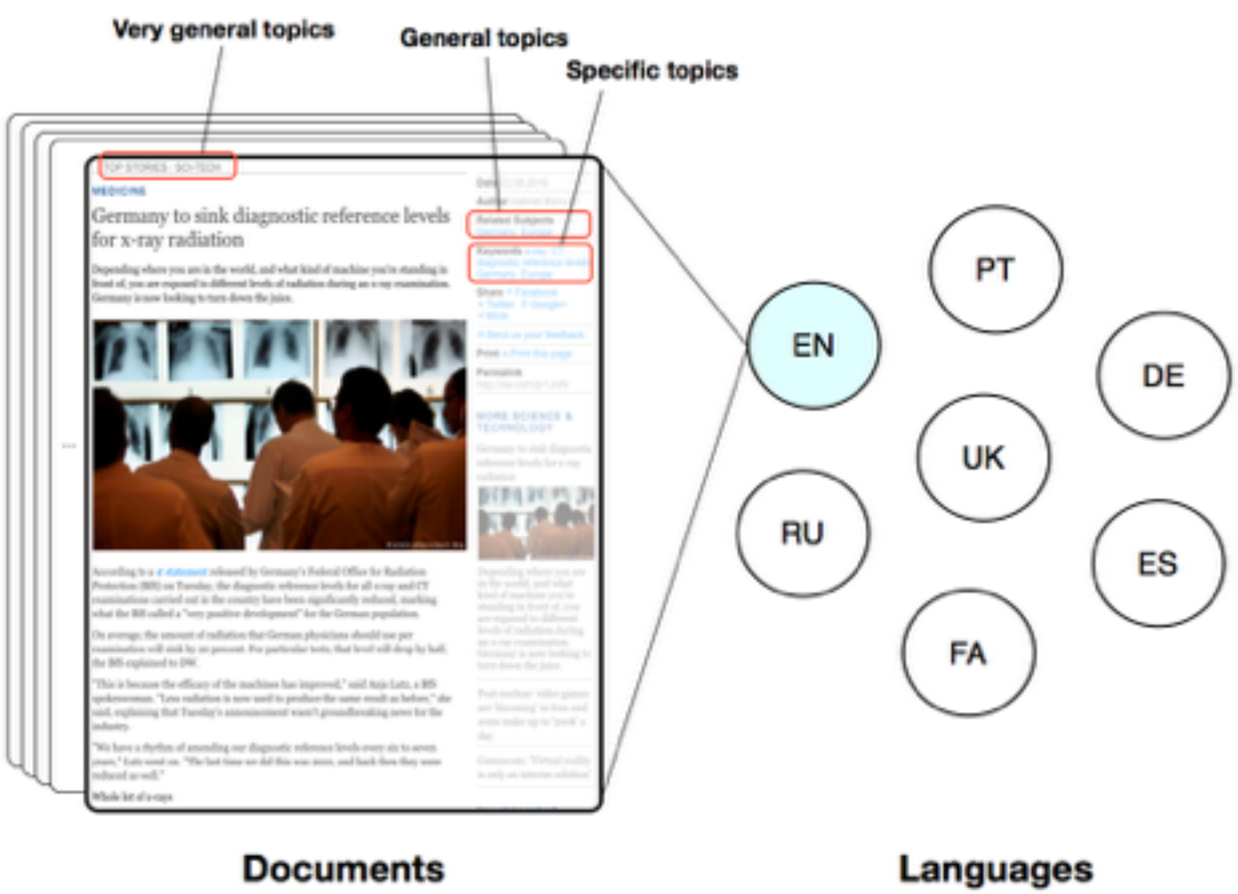
(Yang et al., 2016)

Reflections on Multilingual Document Classification

- **What are the present limitations?**
  - Current evaluation datasets contain small number of target classes and examples
    - RCV1/RCV2 → 6,000 documents, 2 langs, 4 labels
    - TED corpus → 12,078 documents, 12 langs, 15 labels
  - Requires the labels to be common across languages
  - Data are not enough to train SOA neural architectures

- **Observation**: currently there are several domains which support multiple languages but only monolingual classification is possible

New dataset: Deutsche Welle corpus (600k docs, 8 langs)



Language	Documents	Classes (topics)	
L	X	Y_g	Y_s
English	112,816	327	1,058
German	132,709	367	809
Spanish	75,827	159	684
Portuguese	39,474	95	301
Ukrainian	35,423	28	260
Russian	108,076	102	814
Arabic	57,697	91	344
Persian	36,282	71	127

Table 1: Deutche Welle corpus statistics.

Conclusion

- Multilingual word embeddings are useful for tasks where there is lack of parallel data
- Word sequence modeling is advancing quickly with the establishment of neural methods
    - Machine Translation
    - Document Classification
- **Multilingual Neural Machine Translation**
    - is useful for low-resourced languages
    - transfers knowledge in large-scale setting
- **Multilingual Document Classification**
    - several large resources available but with disjoint labels
    - could possibly benefit from NMT lessons

References (1/3)

- Le, Quoc V., and Tomas Mikolov. "Distributed Representations of Sentences and Documents." In ICML, vol. 14, pp. 1188-1196. 2014.
- Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882, 2014.
- Denil, Misha, Alban Demiraj, Nal Kalchbrenner, Phil Blunsom, and Nando de Freitas. "Modelling, visualising and summarising documents with a single convolutional neural network." arXiv preprint arXiv:1406.3830, 2014.
- Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. "Hierarchical attention networks for document classification." In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.
- Tang, Duyu, Bing Qin, and Ting Liu. "Document modeling with gated recurrent neural network for sentiment classification." In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1422-1432, 2015.
- Firat, Orhan, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural, and Yoshua Bengio. "Multi-way, multilingual neural machine translation." Computer Speech & Language, 2016.
- Pappas, Nikolaos, Miriam Redi, Mercan Topkara, Brendan Jou, Hongyi Liu, Tao Chen, and Shih-Fu Chang. "Multilingual visual sentiment concept matching." In International Conference of Multimedia Retrieval, 2016.
- Pappas, Nikolaos, and Andrei Popescu-Belis. "Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis." In Conference on Empirical Methods in Natural Language Processing, 2014.
- Pappas Nikolaos, Andrei Popescu-Belis. "Explicit Document Modeling through Weighted Multiple-Instance Learning", Under review.
- Yoav Goldberg. "A primer on neural network models for natural language processing" arXiv preprint:1510.00726, 2015.
- Ian Goodfellow, Aaron Courville, and Joshua Bengio. "Deep learning". Book in preparation for MIT Press., 2015.

References (2/3)

- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." arXiv preprint arXiv:1609.08144, 2016.
- Zoph, Barret, and Kevin Knight. "Multi-Source Neural Translation." arXiv preprint arXiv:1601.00710, 2016.
- Dong, Daxiang, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. "Multi-task learning for multiple language translation." In Proceedings of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing, pp. 1723-1732. 2015.
- Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025, 2015.
- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473, 2014.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." In Advances in neural information processing systems, pp. 3104-3112. 2014.
- Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078, 2014.
- Irsoy, Ozan, and Claire Cardie. "Deep recursive neural networks for compositionality in language." In Advances in Neural Information Processing Systems, pp. 2096-2104. 2014.
- Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).
- Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9, no. 8 (1997): 1735-1780.
- Levy, Omer, and Yoav Goldberg. "Dependency-Based Word Embeddings." In ACL (2), pp. 302-308. 2014.

References (3/3)

- Pappas, Nikolaos, Miriam Redi, Mercan Topkara, Brendan Jou, Hongyi Liu, Tao Chen, and Shih-Fu Chang. "Multicultural Visual Concept Retrieval and Clustering.”, under review, 2016.
- Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai. "Inducing crosslingual distributed representations of words." 2012.
- Bengio, Yoshua, and Greg Corrado. "Bilbowa: Fast bilingual distributed representations without word alignments." 2014.
- Hermann, Karl Moritz, and Phil Blunsom. "Multilingual distributed representations without word alignment." arXiv preprint arXiv:1312.6173, 2013.
- Faruqui, Manaal, and Chris Dyer. "Improving vector space word representations using multilingual correlation." Association for Computational Linguistics, 2014.
- Søgaard, Anders, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. "Inverted indexing for cross-lingual nlp." In The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015), 2015.
- Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. "Massively multilingual word embeddings." arXiv preprint arXiv:1602.01925, 2016.
- Guo, Jiang, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. "Cross-lingual dependency parsing based on distributed representations." In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 1, pp. 1234-1244, 2015.
- Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. "Combining language and vision with a multimodal skip-gram model." arXiv preprint arXiv:1501.02598, 2015.
- Li, Jiwei, and Dan Jurafsky. "Do multi-sense embeddings improve natural language understanding?." arXiv preprint arXiv:1506.01070 (2015).
- Faruqui, Manaal, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. "Retrofitting word vectors to

Resources (1/2)

➡ **Online courses**

- Coursera course on “Neural networks for machine learning” by Geoffrey Hinton
  https://www.coursera.org/learn/neural-networks
- Coursera course on “Machine learning” by Andrew Ng
  https://www.coursera.org/learn/machine-learning
- Stanford CS224d “Deep learning for NLP” by Richard Socher
  http://cs224d.stanford.edu/

➡ **Conference tutorials**

- Richard Socher and Christopher Manning, “Deep learning for NLP”, EMNLP 2013 tutorial.
  http://nlp.stanford.edu/courses/NAACL2013/
- David Jurgens and Mohammad Taher Pilehvar, “Semantic Similarity Frontiers: From Concepts to Documents”, EMNLP 2015 tutorial.
  http://www.emnlp2015.org/tutorials.html#t1
- Mitesh M Kharpa, Sarath Chandar, “Multilingual and Multimodal Language Processing”, NAACL 2016 tutorial.
  http://naacl.org/naacl-hlt-2016/t2.html

Resources (2/2)

➡ **Deep learning toolkits**

- Theano http://deeplearning.net/software/theano
- Torch http://www.torch.ch/
- Tensorflow http://www.tensorflow.org/
- Keras http://keras.io/

➡ **Pre-trained word vectors and codes**

- Word2vec toolkit and vectors
  https://code.google.com/p/word2vec/
- GloVe code and vectors
  http://nlp.stanford.edu/projects/glove/
- Hellinger PCA
  https://github.com/rlebret/hpca
- Online word vector evaluation
  http://wordvectors.org/