Engaging Content
Engaging People

What has Deep Learning brought to Natural Language Processing?

Prof. Qun Liu, Dublin City University

Deep Learning Meetup Accenture The Dock

21st March 2018

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
SFI Science Foundation Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Background

- Deep Learning (DL) has brought great changes in the area of Natural Language Processing (NLP)
  - Almost all the states-of-the-art of NLP tasks have been refreshed
    - Parsing, Translation, …
  - Some previously difficult tasks become easy
    - Chitchat, Image caption generation
    - morphologically-rich language translation, adaptation, …
  - Some previously impossible tasks become possible
    - Interlingua, …

Questions

- What fundamental change has DL brought to NLP?
- What impacts have been made by this fundamental change?
- What is the weakness of DL in NLP?
- What is the future direction of NLP?

In this presentation, I will try to answer the above questions, followed by the introduction of some of our own work.

Outline

The Changes brought by DL to NLP

Our work on DL-based NLP

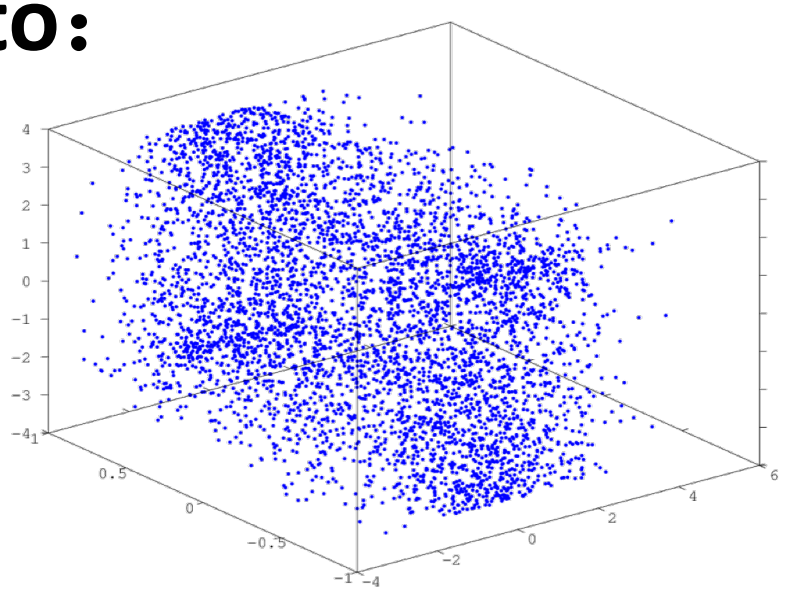
Weakness of DL-based NLP and Future Direction

The most Fundamental Change brought by DL to NLP

The spaces where NLP problems are defined have been moved

from:

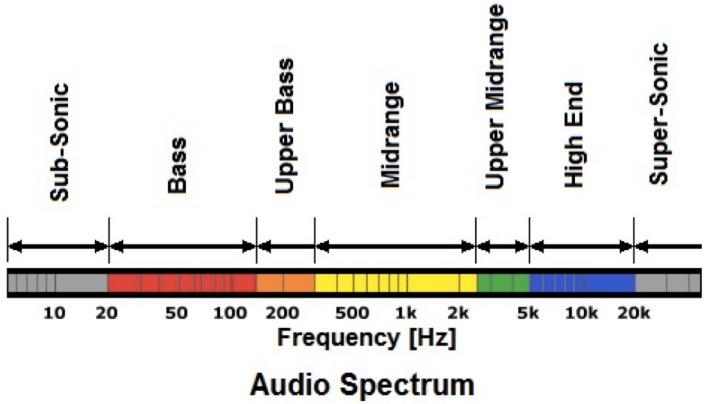
to:

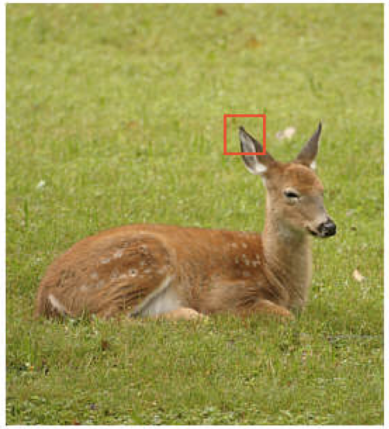


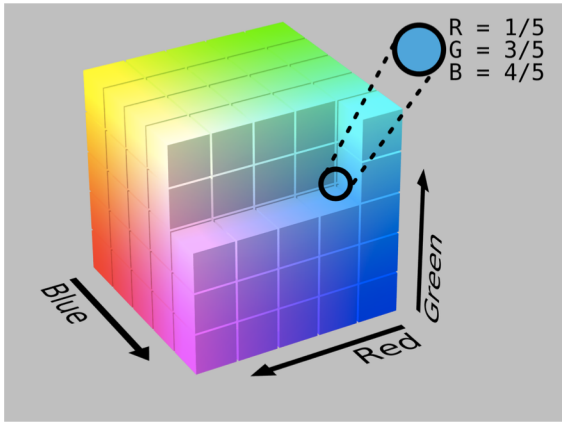
Symbolic Discrete Spaces

Numerical Continuous Spaces

Sounds and Images as Numbers



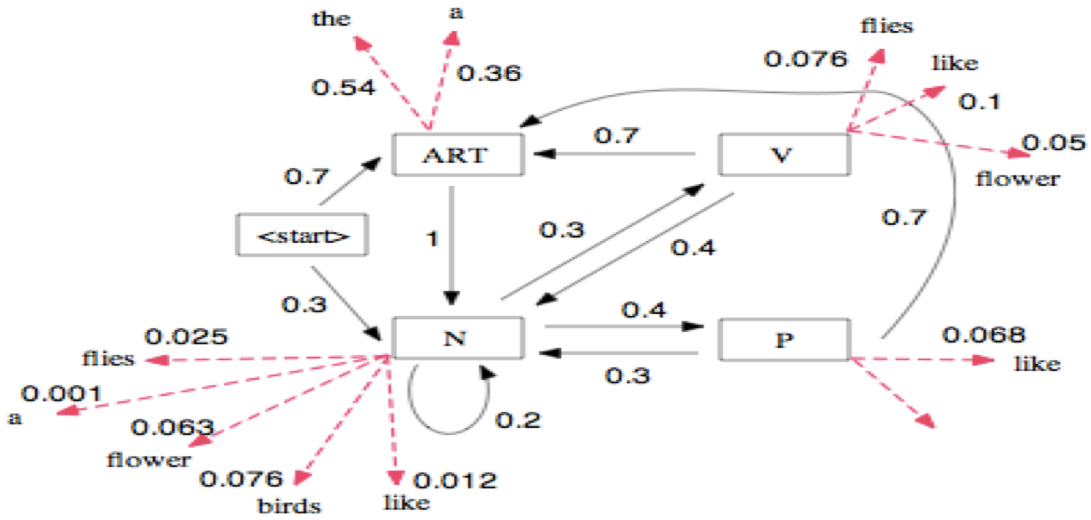




Language as Symbols

Statistical NLP – Symbols with Numbers

- HMM, MaxEnt, CRF, SCFG, SMT…



f	e	ϕ(f|e)	lex(f|e)	ϕ(e|f)	lex(e|f)	Alignment
失业 人数	unemployment figures	0.3	0.0037	0.0769	0.0018	0-0 1-1
失业 人数	number of unemployed	0.1333	0.0188	0.1025	0.0041	1-0 1-1 0-2
失业 人数	. unemployment was	0.3333	0.0015	0.0256	6.8e-06	0-1 1-1 1-2

Neural NLP – Pure Numbers

- Word Embeddings

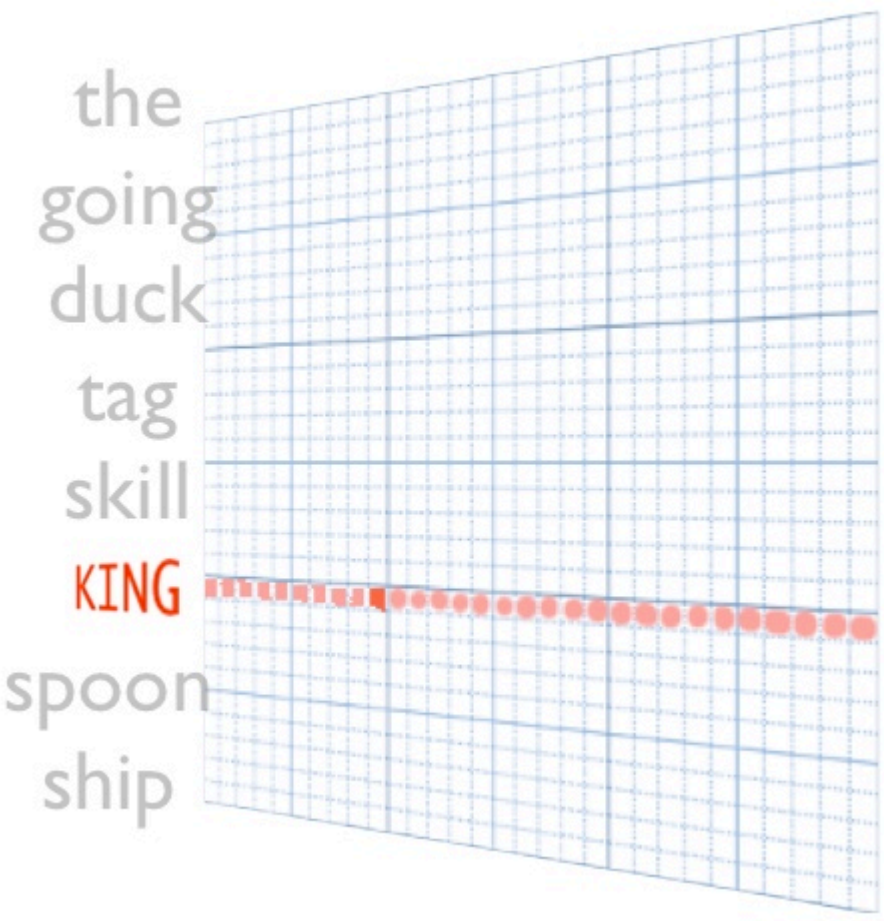
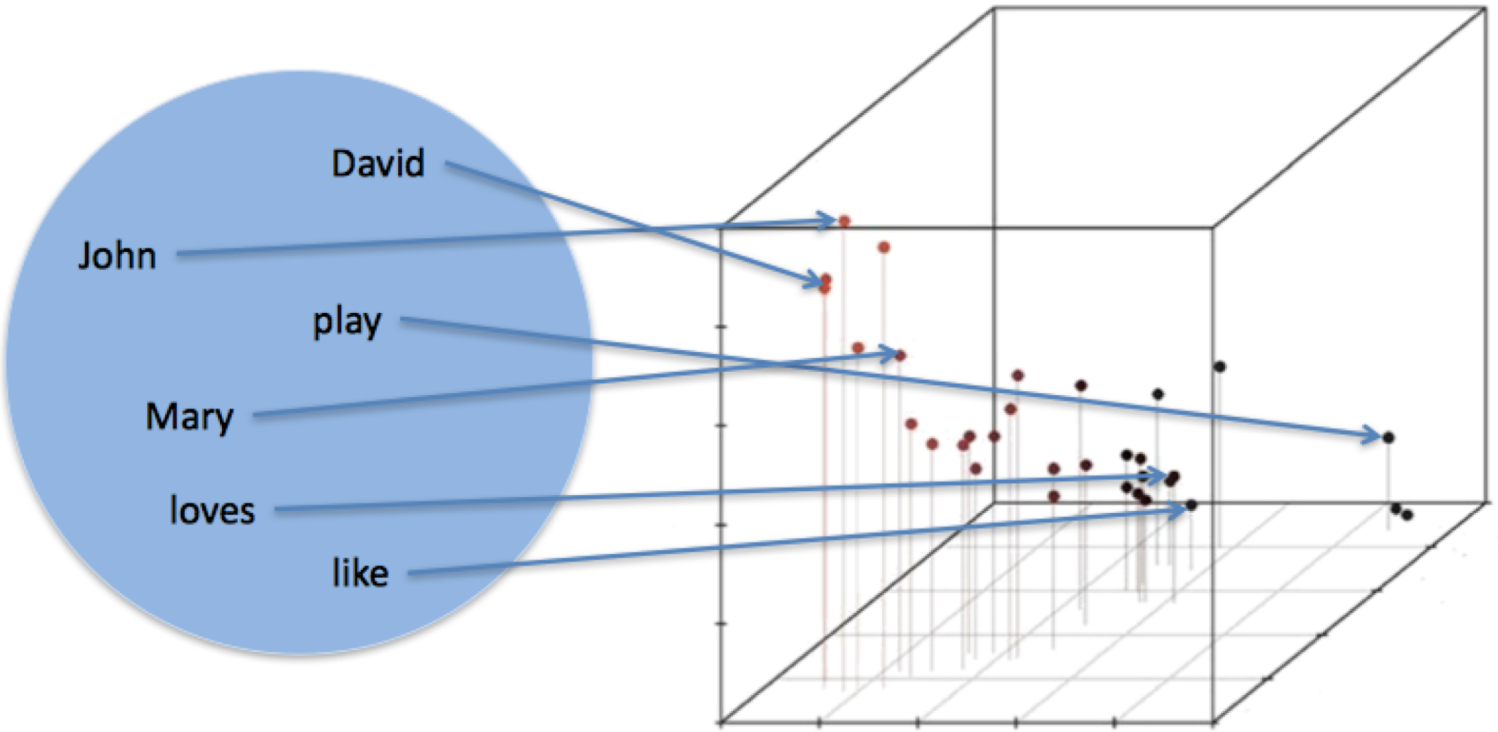


Figure extracted from Christopher Moody's slides

Neural NLP – Pure Numbers

- Word Embeddings



Neural NLP – Pure Numbers

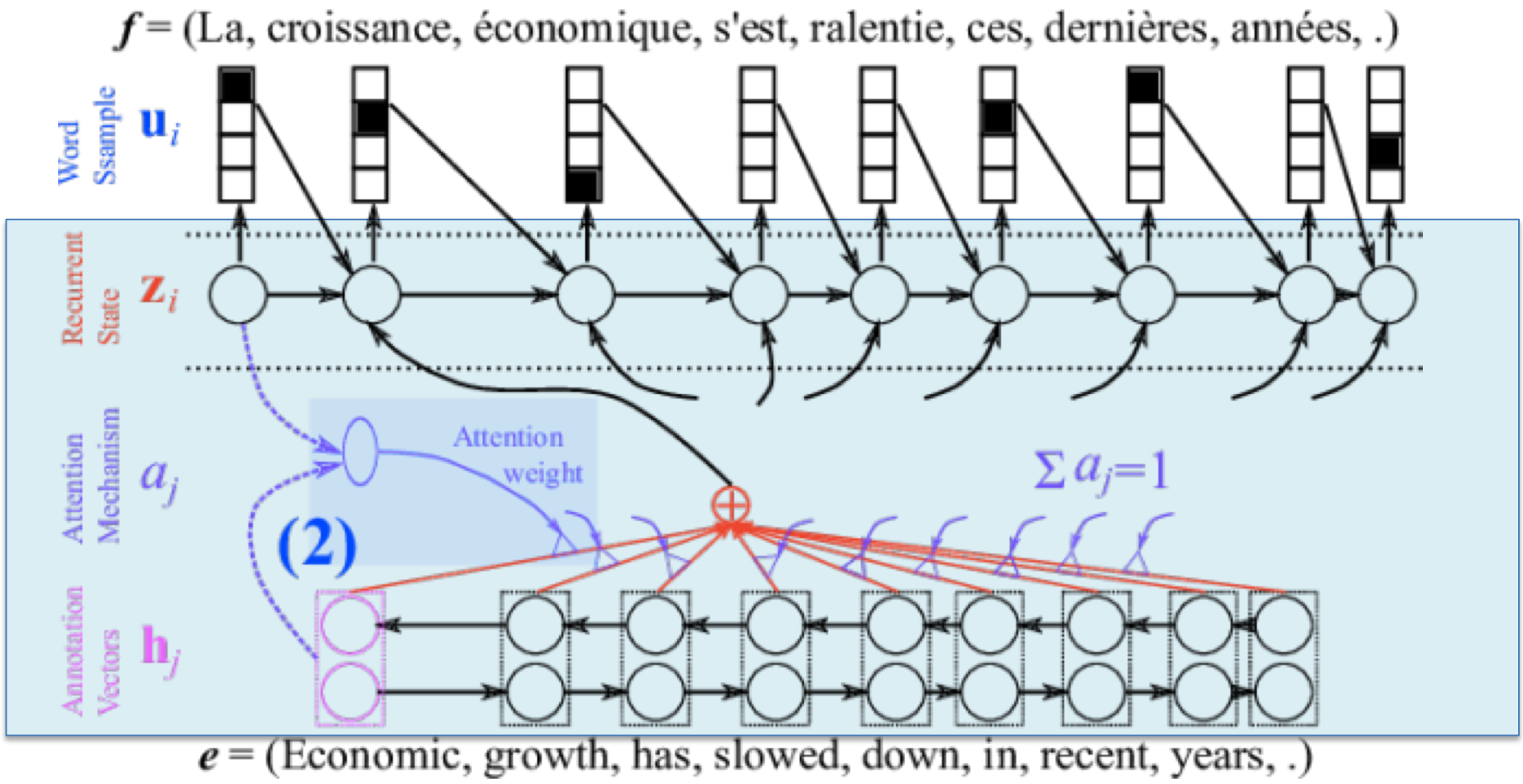
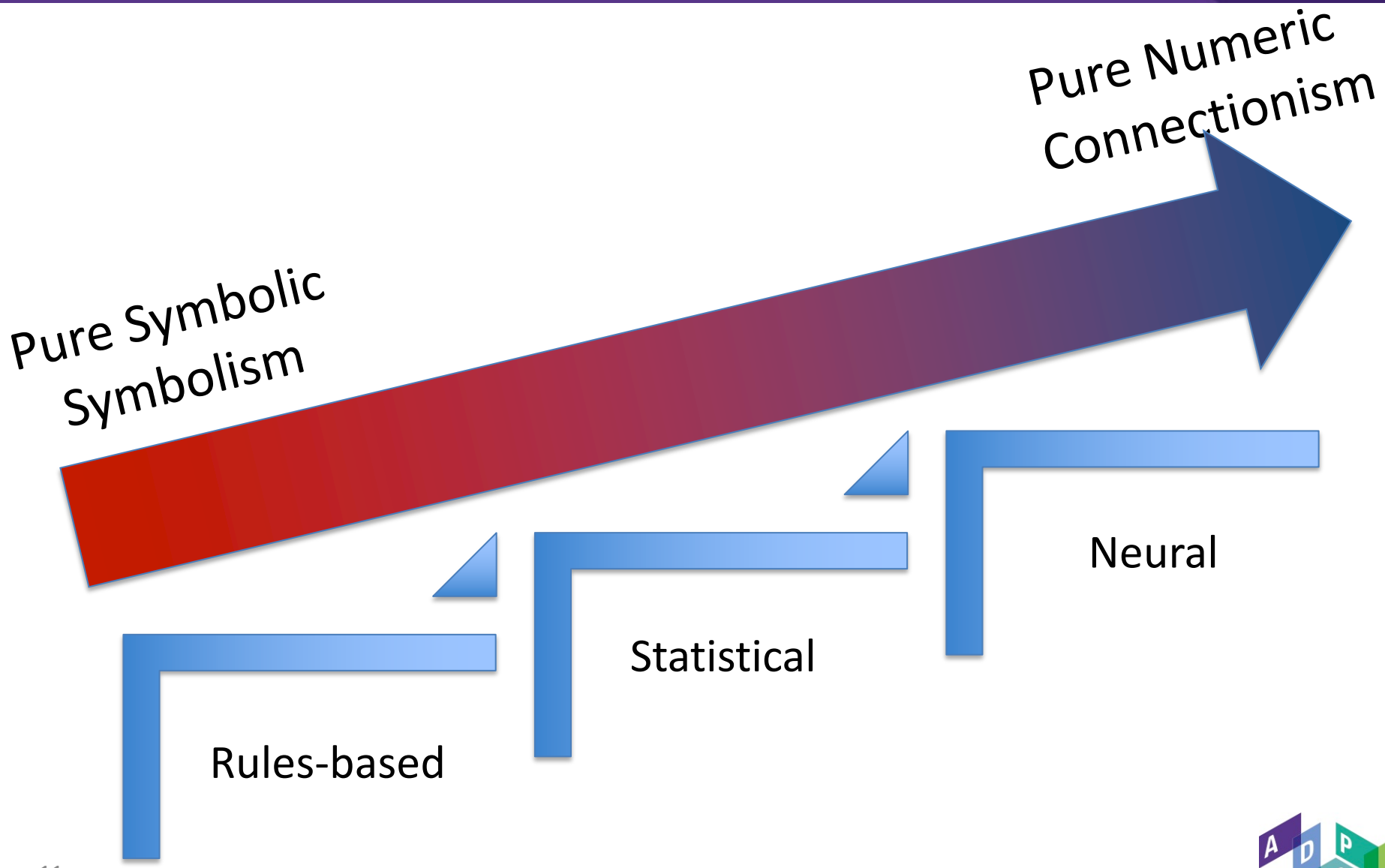


Figure from : https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/

The Trend


Pure Numeric
Connectionism
Pure Symbolic
Symbolism
Neural
Statistical
Rules-based

Other Changes brought by DL to NLP

- **Finer granularity**
- Better generalization
- Breakdown of the boundary between modalities
- Monolithic NLP Models

Granularity Selection for Conventional NLP Models

Model	Granularity
Word-based SMT	words
Phrase-based SMT	words, phrases (grammatical & ungrammatical)
Syntax-based SMT	words, (grammatical) phrases, rules
Wordnet	synset

- A NLP model must be defined in a certain granularity
- Models of different granularities are
  - totally different
  - not compatible with each other: PBMT and Syntactic SMT
- It's hard to integrate knowledge of other granularities into a model which is defined in a different granularity
  - None of SMT models can translate morphologically rich languages well

Granularities for Neural NLP Models

- In Neural NLP models, all linguistic units are expressed as embeddings
  - Word embeddings, sub-word em, character em
  - Phrase embeddings, sentence embeedings, …
- Neural NLP models with different linguistic granularities are similar
- It's easy to incorporate linguistic data or knowledge in different granularities in neural NLP framework

Other Changes brought by DL to NLP

- Finer granularity
- **Better generalization**
- Breakdown of the boundary between modalities
- Monolithic NLP Models

Better Generalization

- Use Language Models (LMs) as an example
- In N-Gram LMs,

$$P_{bigram}(w_1 \dots w_n) = \prod_{i=1}^{n} p(w_i | w_{i-1})$$

if **"loves Mary" appears 1000 times** in the training corpus but **"loves John" appears only once**, they will obtain very different bigram LM probabilities

- However, in Neural LMs, because **"Mary" and "John" are very close in the embedding space**, "loves Mary" and "loves John" will **obtain very similar NLM probabilities**

Other Changes brought by DL to NLP

- Finer granularity
- Better generalization
- **Breakdown of the boundary between modalities**
- Monolithic NLP Models

Breaking down the boundary between Modalities

- Traditionally NLP and Image/Speech processing are very different techniques
  - NLP is working on symbolic language data
  - Image/Speech is working on continuous signal data
- By using NN, all language data are converted to numerical numbers, which is in the same form as image/speech data.
- Multimodal/cross-modal processing become straightforward and easy to implement

Image Caption Generation

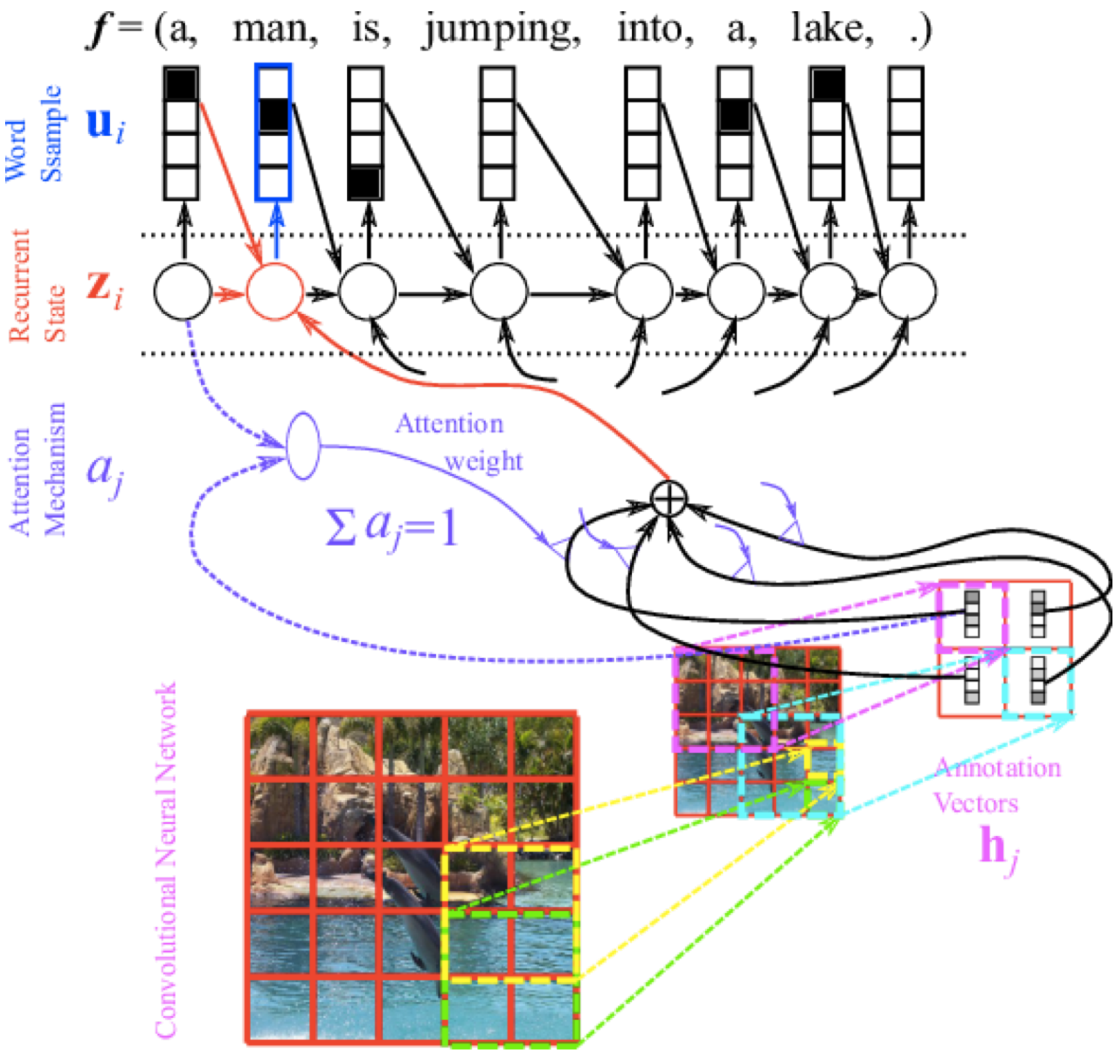


Figure from: https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/

Visual Question Answering (VQA)

Welcome to the VQA Challenge 2017!

What is the mustache made of?

AI System

bananas

http://visualqa.org/challenge.html

Other Changes brought by DL to NLP

- Finer granularity
- Better generalization
- Breakdown of the boundary between modalities
- **Monolithic NLP Models**

Monolithic NLP Models

- Traditionally, an complex NLP model is usually be divided to multiple submodels, where each submodel are trained with specific data, and optimized against its own object function. The submodels are connected in a certain structure, typically a pipeline structure.
  - The submodels are trained locally which may not lead to a global optimum.
  - Error propagations exist through the data flow.
- NN use monolithic models and can avoid the above problems ➔ so called “end-to-end model”.

Interlingua: experiences in 1989

- Makoto Nagao (Kyoto University) said: ".. when the pivot language [i.e. interlingua] is used, the results of the analytic stage must be in a form which can be utilized by all of the different languages into which translation is to take place. This level of subtlety is a practical impossibility." (Machine Translation, Oxford, 1989)
- Patel-Schneider (METAL system) said: "METAL employs a modified transfer approach rather than an interlingua. If a meta-language [an interlingua] were to be used for translation purposes, it would need to incorporate all possible features of many languages. That would not only be an endless task but probably a fruitless one as well. Such a system would soon become unmanageable and perhaps collapse under its own weight." (A four-valued semantics for terminological reasoning, Artificial Intelligence, 38, 1989)

Outline

The Changes brought by DL to NLP

Our work on DL-based NLP

Weakness of DL-based NLP and Future Direction

Our work on DL-based NLP

- **Multimodal Machine Translation**
- NMT with Morphological Information
- NMT with Discourse Information
- Domain Adaptation for NMT
- Lexical Constrained Decoding for NMT

Incorporate Visual Features into NMT

Figure 1: Image extracted from Wikipedia for which there are human created captions in English and Portuguese.

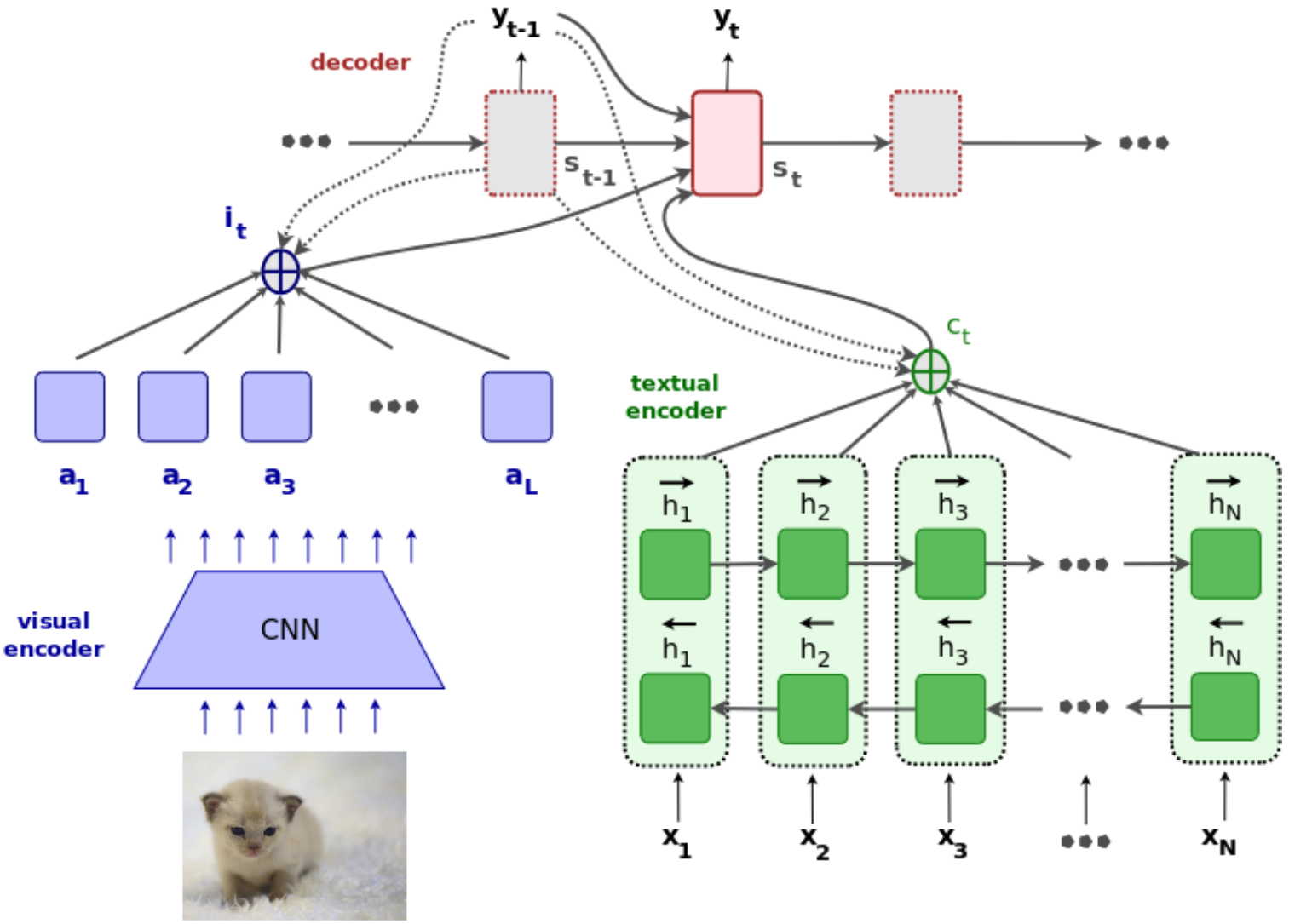
- `Human EN` – a **plant steem wrapping** around another one.
- `Human PT` – **caule de planta enrolando-se** em outro.
- `MT PT` – Uma **fábrica steem embalagem** em torno de um outro.

factory, industrial plant

packaging, wrapping a gift

out-of-vocabulary

Doubly-Attentive Decoder for Multi-Modal NMT



Calixto, Liu & Campbell (ACL 2017)

Local Image Features

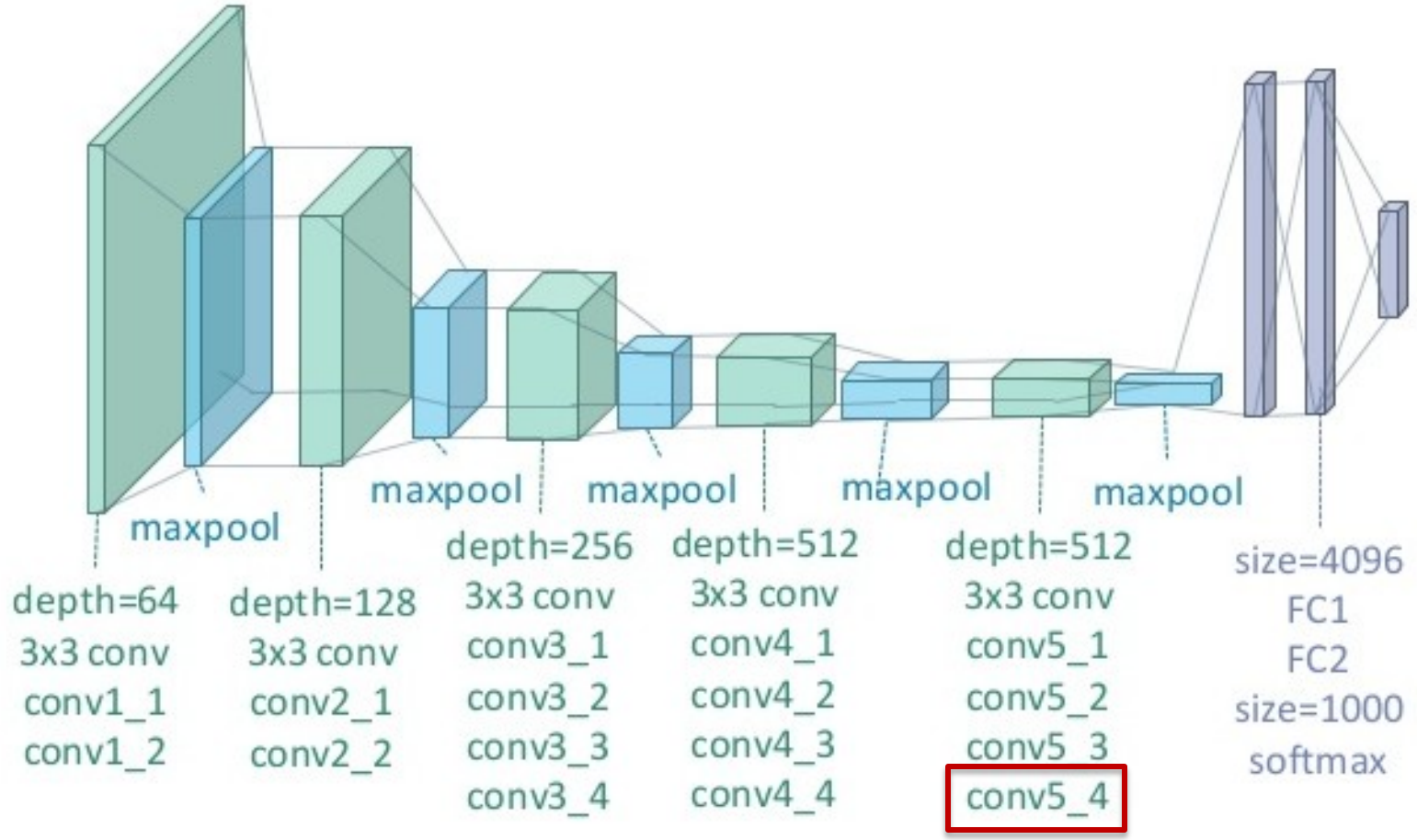


Figure 2.7: Illustration of the VGG19 network architecture.

Doubly-Attentive Decoder for Multi-Modal NMT

Training on $M30k_T$

		English→German				
Model	**Training data**	**BLEU4↑**	**METEOR↑**	**TER↓**	**chrF3↑** (prec. / recall)	
NMT	$M30k_T$	33.7	52.3	46.7	65.2	(67.7 / 65.0)
PBSMT	$M30k_T$	32.9	54.3†	45.1†	**67.4**	(66.5 / 67.5)
Huang et al. (2016)	$M30k_T$	35.1 (↑ 1.4)	52.2 (↓ 2.1)	—	—	—
	+ RCNN	**36.5** (↑ 2.8)	54.1 (↓ 0.2)	—	—	—
$NMT_{SRC+IMG}$	$M30k_T$	**36.5**†‡	**55.0**†	**43.7**†‡	67.3	(66.8 / 67.4)
Improvements						
$NMT_{SRC+IMG}$ vs. NMT		↑ **2.8**	↑ **2.7**	↓ **3.0**	↑ 2.1	↓ 0.9 / ↑ 2.4
$NMT_{SRC+IMG}$ vs. PBSMT		↑ **3.6**	↑ 0.7	↓ **1.4**	↓ 0.1	↑ 0.3 / ↓ 0.1
$NMT_{SRC+IMG}$ vs. Huang		↑ 1.4	↑ 2.8	—	—	—
$NMT_{SRC+IMG}$ vs. Huang (+RCNN)		↑ 0.0	↑ 0.9	—	—	—

Doubly-Attentive Decoder for Multi-Modal NMT

(a) Image–target word alignments.

(b) Source–target word alignments.

Figure 7.2: Visualisation of image– and source–target word alignments for the M30k_{T} test set.

Incorporate Global Image Features into NMT

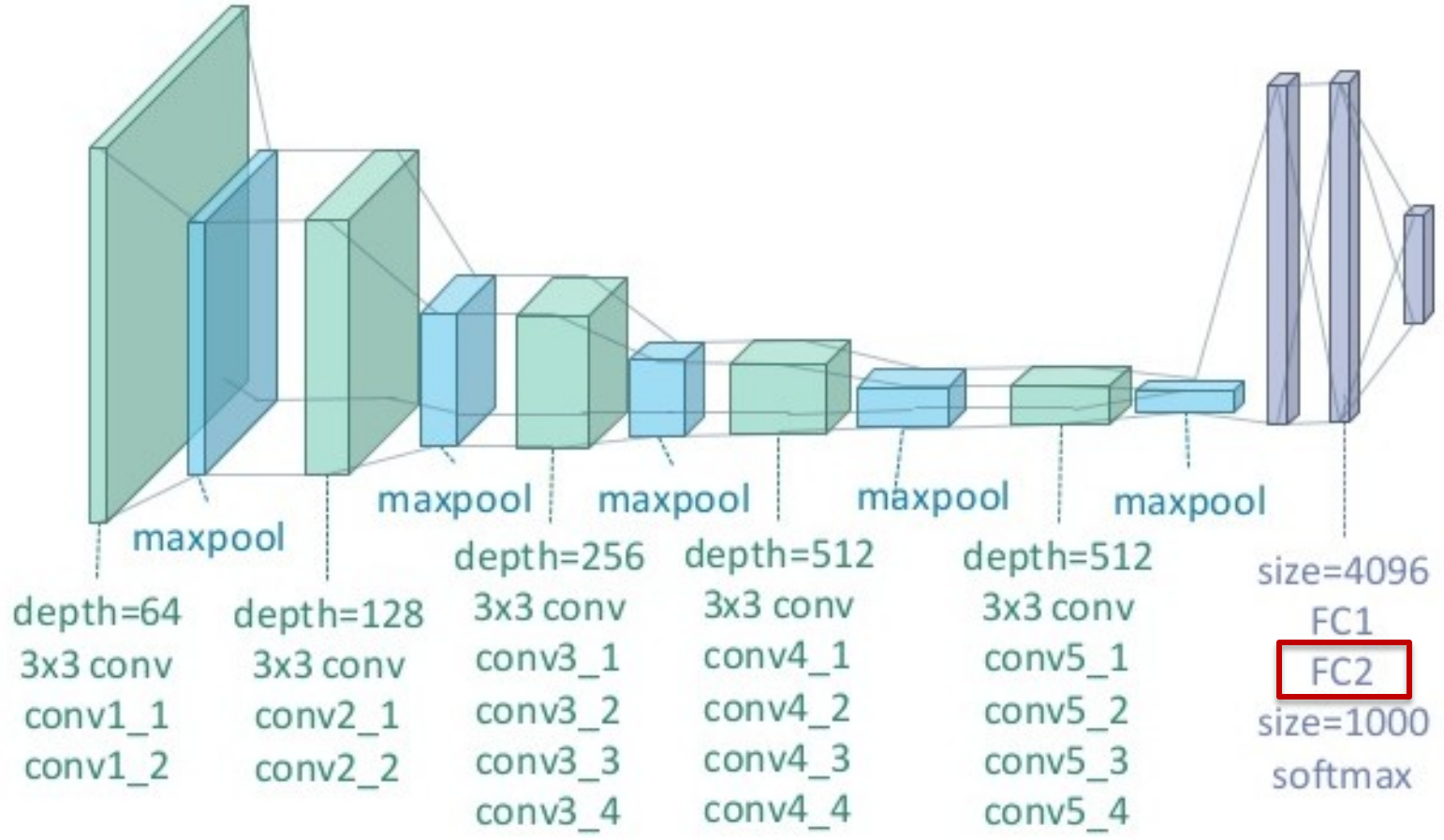
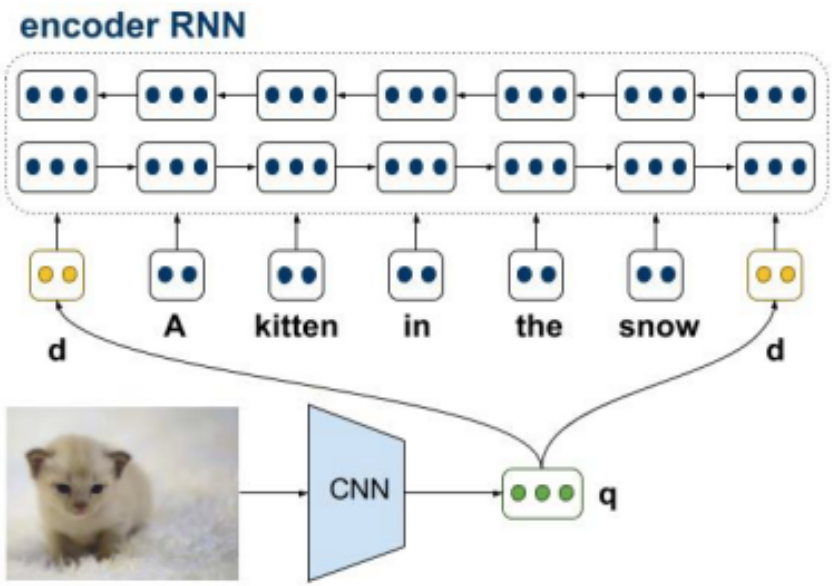
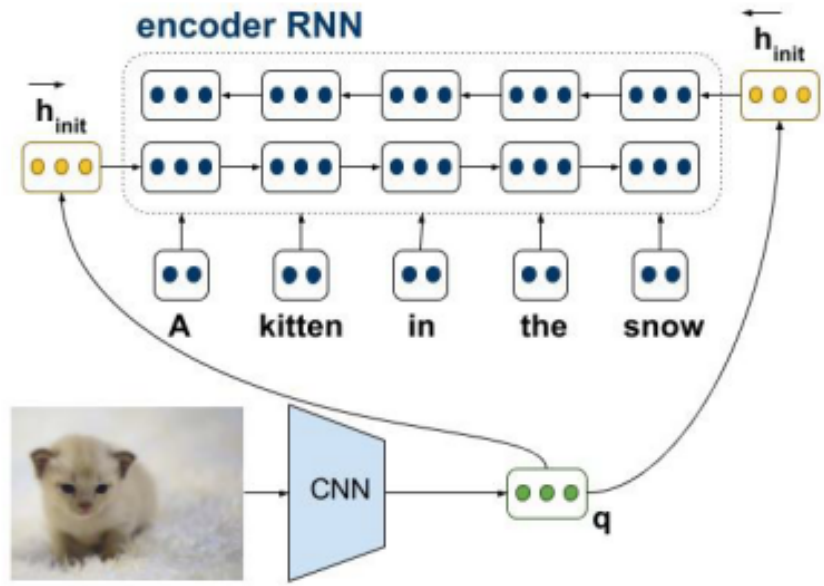


Figure 2.7: Illustration of the VGG19 network architecture.

Incorporate Global Image Features into NMT



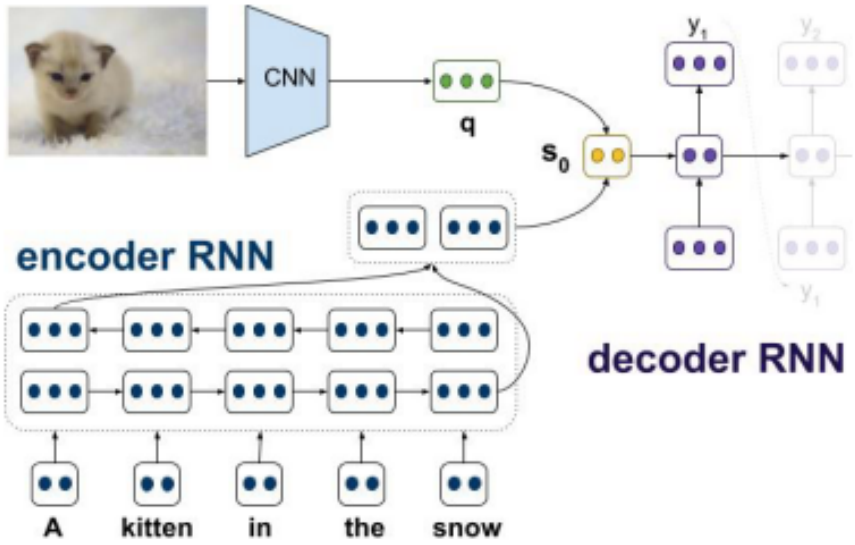
(a) $\mathbf{IMG_W}$: An encoder RNN that uses image features as words in the source sequence.



(b) $\mathbf{IMG_E}$: Using an image to initialise the encoder hidden states.

Calixto and Liu (EMNLP 2017)

Incorporate Global Image Features into NMT



(c) $\mathbf{IMG_D}$: Image as additional data to initialise the decoder hidden state s_0.

Calixto and Liu (EMNLP 2017)

Incorporate Global Image Features into NMT

English→German

	BLEU4↑	METEOR↑	TER↓	chrF3↑
	training data: translated M30k			
PBSMT	32.9	54.1	45.1	67.4
NMT	33.7	52.3	46.7	64.5
Huang	35.1	52.2	—	—
+ RCNN	36.5	54.1	—	—
IMG_{1W}	$37.1^{\dagger\ddagger}$ (↑ 3.4)	$54.5^{\ddagger}$ (↑ 0.4)	$42.7^{\dagger\ddagger}$ (↓ 2.4)	66.9 (↓ 0.5)
IMG_{2W}	$36.9^{\dagger\ddagger}$ (↑ 3.2)	$54.3^{\ddagger}$ (↑ 0.2)	$\mathbf{41.9}^{\dagger\ddagger}$ (↓ 3.2)	66.8 (↓ 0.6)
IMG_{E}	$37.1^{\dagger\ddagger}$ (↑ 3.4)	$55.0^{\dagger\ddagger}$ (↑ 0.9)	$43.1^{\dagger\ddagger}$ (↓ 2.0)	67.6 (↑ 0.2)
IMG_{D}	$\mathbf{37.3}^{\dagger\ddagger}$ (↑ 3.6)	$\mathbf{55.1}^{\dagger\ddagger}$ (↑ 1.0)	$42.8^{\dagger\ddagger}$ (↓ 2.3)	**67.7** (↑ 0.3)
IMG_{2W+D}	$35.7^{\dagger\ddagger}$ (↑ 2.0)	$53.6^{\ddagger}$ (↓ 0.5)	$43.3^{\dagger\ddagger}$ (↓ 1.8)	66.2 (↓ 1.2)
IMG_{E+D}	$37.0^{\dagger\ddagger}$ (↑ 3.3)	$54.7^{\ddagger}$ (↑ 0.6)	$42.6^{\dagger\ddagger}$ (↓ 2.5)	67.2 (↓ 0.2)
	+ back-translated comparable M30k			
PBSMT	34.0	55.0	44.7	68.0
NMT	35.5	53.4	43.3	65.3
IMG_{2W}	$36.7^{\dagger\ddagger}$ (↑ 1.2)	$54.6^{\ddagger}$ (↓ 0.4)	$42.0^{\dagger\ddagger}$ (↓ 1.3)	66.8 (↓ 1.2)
IMG_{E}	$\mathbf{38.5}^{\dagger\ddagger}$ (↑ 3.0)	$55.7^{\dagger\ddagger}$ (↑ 0.9)	$\mathbf{41.4}^{\dagger\ddagger}$ (↓ 1.9)	68.3 (↑ 0.3)
IMG_{D}	$\mathbf{38.5}^{\dagger\ddagger}$ (↑ 3.0)	$\mathbf{55.9}^{\dagger\ddagger}$ (↑ 1.1)	$41.6^{\dagger\ddagger}$ (↓ 1.7)	**68.4** (↑ 0.4)

Incorporate Global Image Features into NMT

German→English

	BLEU4↑	METEOR↑	TER↓	chrF3↑
	training data: translated M30k			
PBSMT	32.8	34.8	43.9	61.8
NMT	38.2	35.8	40.2	62.8
IMG_{2W}	39.5 ‡ (↑ 1.3)	37.1†‡ (↑ 1.3)	37.1†‡ (↓ 3.1)	63.8 (↑ 1.0)
IMG_{E}	41.1†‡ (↑ 2.9)	37.7†‡ (↑ 1.9)	37.9†‡ (↓ 2.3)	65.7 (↑ 2.9)
IMG_{D}	41.3†‡ (↑ 3.1)	37.8†‡ (↑ 2.0)	37.9†‡ (↓ 2.3)	65.7 (↑ 2.9)
IMG_{2W+D}	39.9†‡ (↑ 1.7)	37.2†‡ (↑ 1.4)	**37.0**†‡ (↓ 3.2)	64.4 (↑ 1.6)
IMG_{E+D}	**41.9**†‡ (↑ 3.7)	**37.9**†‡ (↑ 2.1)	37.1†‡ (↓ 3.1)	**66.0** (↑ 3.2)
$PBSMT^{+}$	42.5	39.5	35.6	**68.7**
	+ back-translated comparable M30k			
NMT	42.6	38.9	36.1	67.6
IMG_{2W}	42.4 ‡ (↓ 0.2)	39.0 ‡ (↑ 0.1)	**34.7**†‡ (↓ 1.4)	67.6 (↑ 0.0)
IMG_{E}	**43.9**†‡ (↑ 1.3)	**39.7** ‡ (↑ 0.8)	34.8†‡ (↓ 1.3)	**68.7** (↑ 1.1)
IMG_{D}	43.4 ‡ (↑ 0.8)	39.3 ‡ (↑ 0.4)	35.2 ‡ (↓ 0.9)	67.8 (↑ 0.2)

Our work on DL-based NLP

- Multimodal Machine Translation
- **NMT with Morphological Information**
- NMT with Discourse Information
- Domain Adaptation for NMT
- Lexical Constrained Decoding for NMT

NMT with Morphological Information

- Chinese and English are morphologically poor languages
- Unlike Chinese and English, a number of languages have very rich morphology:

Word	Translation
Turkish:	
terbiye	good manners
terbiye+siz	rude
terbiye+siz+lik	rudeness
terbiye+siz+lik+leri	their rudeness
terbiye+siz+lik+leri+nden	from their rudeness
terbiye+siz+lik+leri+nden+mis	it was because of their rudeness

NMT with Morphological Information

- Morphologically rich languages (MRLs) are hard to model in SMT because (i) a single word may have a large number of variations (>1000); (ii) the vocabulary size is very big; (iii) the data scarcity / OOV problem becomes severe.
- In SMT framework, although a lot of ideas have been proposed for MRL translation, none of them can provide a successful solution.

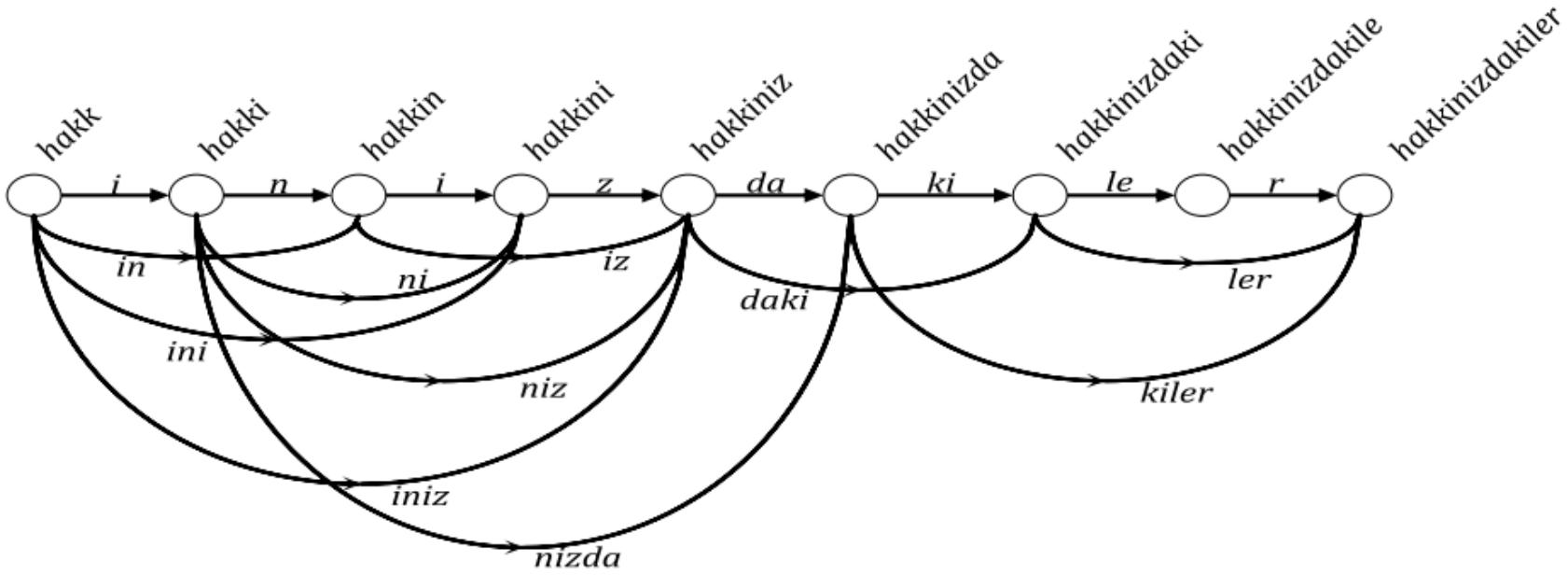
NMT with Morphological Information

- In NMT framework, two solutions are proposed to deal with MRLs both of which are elegant and obtained preliminary success:
  - Character level models
  - Subword level models
- Character level models are promising but also computationally high cost
- Subword level models can reach a comprise between performance and complexity

NMT with Morphological Information

- The most popular subword level model are Byte Pair Encoding (BPE) proposed by Rico et al. (2016)
- Google researchers used a WordPiece model in Google NMT system
- Both BPE and WordPiece generate many mistakes in their subword segmentation
- We proposed a novel (almost) unsupervised model to generate better morpheme (subword) segmentation and archived better MT quality

Dynamic Programming Morpheme Segmentation



$$score = \sum_{s} imp(n) \times freq_{n\_gram} freq(s), imp(n) = \frac{count(n\_gram)}{\sum_{n'=1}^{N}(n'\_gram)}$$

Ref: Peyman Passban's PhD thesis

NMT with Morphological Information

Scheme	Segmented Sequence
Turkish Seq.	[görünüşe]$_1$ [göre]$_2$ [söylemeyeceğinden]$_3$ [çok]$_4$ [eminsin]$_5$
bpe-5K	[görünüş•e]$_1$ [göre]$_2$ [söylemey•eceğ•inden]$_3$ [çok]$_4$ [emin•sin]$_5$
bpe-30K	[görünüşe]$_1$ [göre]$_2$ [söylemey•eceğinden]$_3$ [çok]$_4$ [eminsin]$_5$
bpe-50K	[görünüşe]$_2$ [göre]$_2$ [söylemey•eceğinden]$_3$ [çokeminsin]$_{4,5}$
Our model	[görü•nüş•e]$_1$ [gör•e]$_2$ [söyle•mey•eceği•nden]$_3$ [çok]$_4$ [e•m•in•sin]$_5$
Translation	[seems like]$_{1,2}$ [you've made sure]$_{4,5}$ [to not tell]$_3$
Turkish Seq.	[fırtınayi]$_1$ [buraya]$_2$ [getirecek]$_3$
bpe-5K	[fırt•ın•ayı]$_1$ [buraya]$_2$ [getir•ecek]$_3$
bpe-30K	[fırtın•ayı]$_1$ [buraya]$_2$ [getirecek]$_3$
bpe-50K	[fırtın•ayı]$_1$ [buraya]$_2$ [getirecek]$_3$
Our model	[fırt•ın•ayı]$_1$ [bura•y•a]$_2$ [get•ir•ecek]$_3$
Translation	[it's gonna bring]$_3$ [that storm]$_1$ [here]$_2$

NMT with Morphological Information

	Model	Source	Target	Direction	BLEU
En→MRL	Chung et al. (2016)	*bp*	*bp*		20.47
	Chung et al. (2016)	*bp*	*char*		21.33
	Firat et al. (2016a)	*bp*	*bp*	En→De	20.59
	Sennrich et al. (2016b)	C2/50K	C2/50K		22.8
	Our model	*dp*	*dp*		**23.41**
	Chung et al. (2016)	*bp*	*bp*		25.30
	Chung et al. (2016)	*bp*	*char*		**26.00**
	Firat et al. (2016a)	*bp*	*bp*	En→Ru	19.39
	Sennrich et al. (2016b)	C2/50K	C2/50K		20.90
	Our model	*dp*	*dp*		24.71
	Chung et al. (2016)*	*bp*	*char*		18.01
	Our model	*bp*	*bp*	En→Tr	16.76
	Our model	*dp*	*dp*		**21.05**
MRL→En	Costa-jussà and Fonollosa (2016)	*word*	*word*		18.83
	Costa-jussà and Fonollosa (2016)	*char*	*word*		21.40
	Firat et al. (2016a)	*bp*	*bp*	De→En	24.00
	Lee et al. (2016)	*bp*	*char*		25.27
	Lee et al. (2016)	*bp*	*char*		25.83
	Our model	*dp*	*dp*		**27.13**
	Firat et al. (2016a)	*bp*	*bp*		22.40
	Lee et al. (2016)	*bp*	*char*	Ru→En	22.83
	Lee et al. (2016)	*char*	*char*		22.73
	Our model	*dp*	*dp*		**23.07**
	Chung et al. (2016)*	*bp*	*char*		23.11
	Our model	*bp*	*bp*	Tr→En	23.17
	Our model	*dp*	*dp*		**23.46**

NMT with different subword segmentation methods

Our work on DL-based NLP

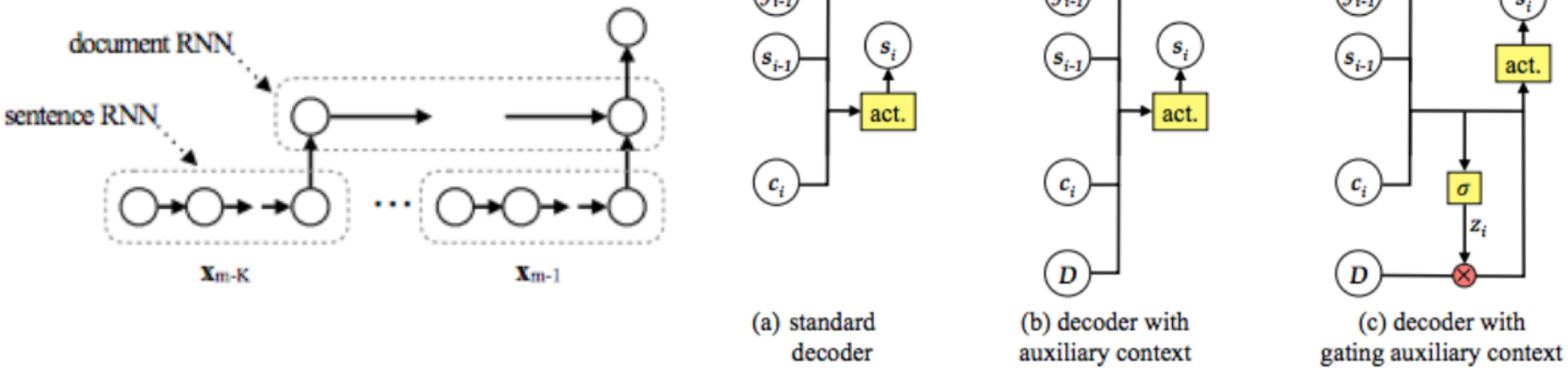
- Multimodal Machine Translation
- NMT with Morphological Information
- **NMT with Discourse Information**
- Domain Adaptation for NMT
- Lexical Constrained Decoding for NMT

NMT with Discourse Information

- Most MT systems translate sentences separately without considering relations between sentences (discourse information)
  - Coherence problem
  - Cohesion problem
  - Consistency problem
- It is very hard and complex to incorporate discourse information into an SMT system
- A small number of work has been conducted discourse level translation under SMT framework but not very successful

NMT with Discourse Information

- We recent proposed an idea:
  Exploiting Cross-Sentence Context for NMT



Wang et al. (EMNLP 2017)

NMT with Discourse Information

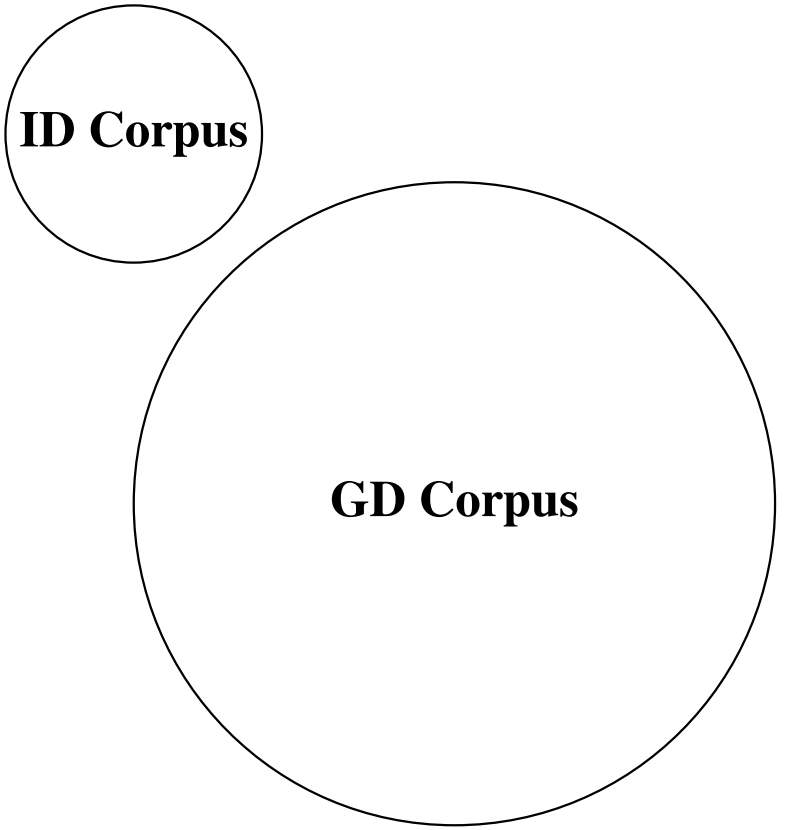
#	System	MT05	MT06	MT08	Ave.	△
1	MOSES	33.08	32.69	23.78	28.24	–
2	NEMATUS	34.35	35.75	25.39	30.57	–
3	+Init_{enc}	36.05	36.44†	26.65†	31.55	+0.98
4	+Init_{dec}	36.27	36.69†	27.11†	31.90	+1.33
5	+$\text{Init}_{\text{enc+dec}}$	36.34	36.82†	27.18†	32.00	+1.43
6	+Auxi	35.26	36.47†	26.12†	31.30	+0.73
7	+Gating Auxi	36.64	37.63†	26.85†	32.24	+1.67
8	+$\text{Init}_{\text{enc+dec}}$+Gating Auxi	**36.89**	**37.76**†	**27.57**†	**32.67**	+2.10

Our work on DL-based NLP

- Multimodal Machine Translation
- NMT with Morphological Information
- NMT with Discourse Information
- **Domain Adaptation for NMT**
- Lexical Constrained Decoding for NMT

Domain Adaptation for NMT

- Domain Adaptation is a common problem in MT application scenario where we have a large amount of general domain (GD) data but only a limited amount of in-domain (ID) data

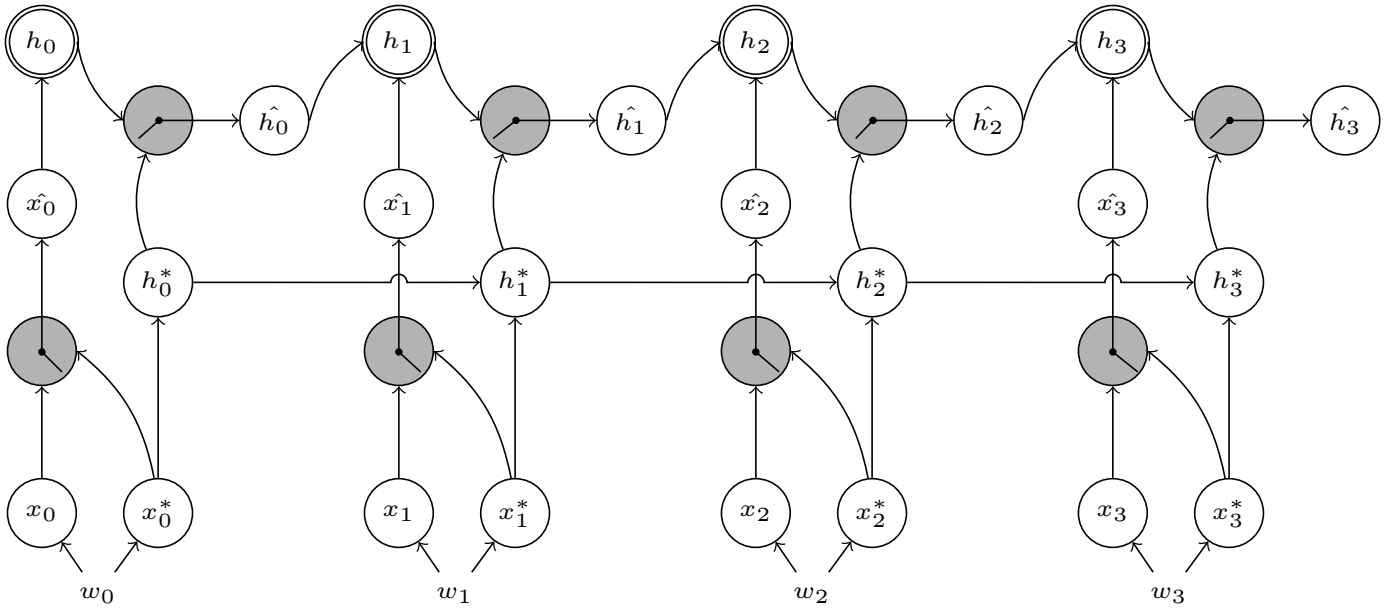


Domain Adaptation for NMT

- Many techniques have been proposed for domain adaptation in SMT
- However in NMT domain adaptation becomes not so difficulty: simply training an GD NMT system on ID corpus plus ensemble seems work well
- We proposed domain adaptation ideas for NMT in two specific use scenarios.

Domain Adaptation for NMT

- Domain Adaptation for NMT with pre-trained large scale word embedding



Zhang, et al. (COLING 2016a)
Jian Zhang's PhD Thesis

Domain Adaptation for NMT

	Validation Set	Test Set
Baselines		
Baseline (KN5)	148.007	141.186
Baseline (*word2vec*)	121.871	117.730
Baseline (Standard)	92.983	89.295
Adaptation on Word Representations		
WVC	95.149	91.414
WVS	88.398	85.231
Adaptation on Context Representations		
CVC	90.337	86.168
WCVC	88.551	85.067
CVS	88.244	84.721
WCVS	90.293	86.679
Gated Adaptation		
WVG	90.937	87.853
CVG	90.301	86.832
DAGRU	**86.247**	**81.900**

Table 2: LM perplexity on Penn Treebank corpus.

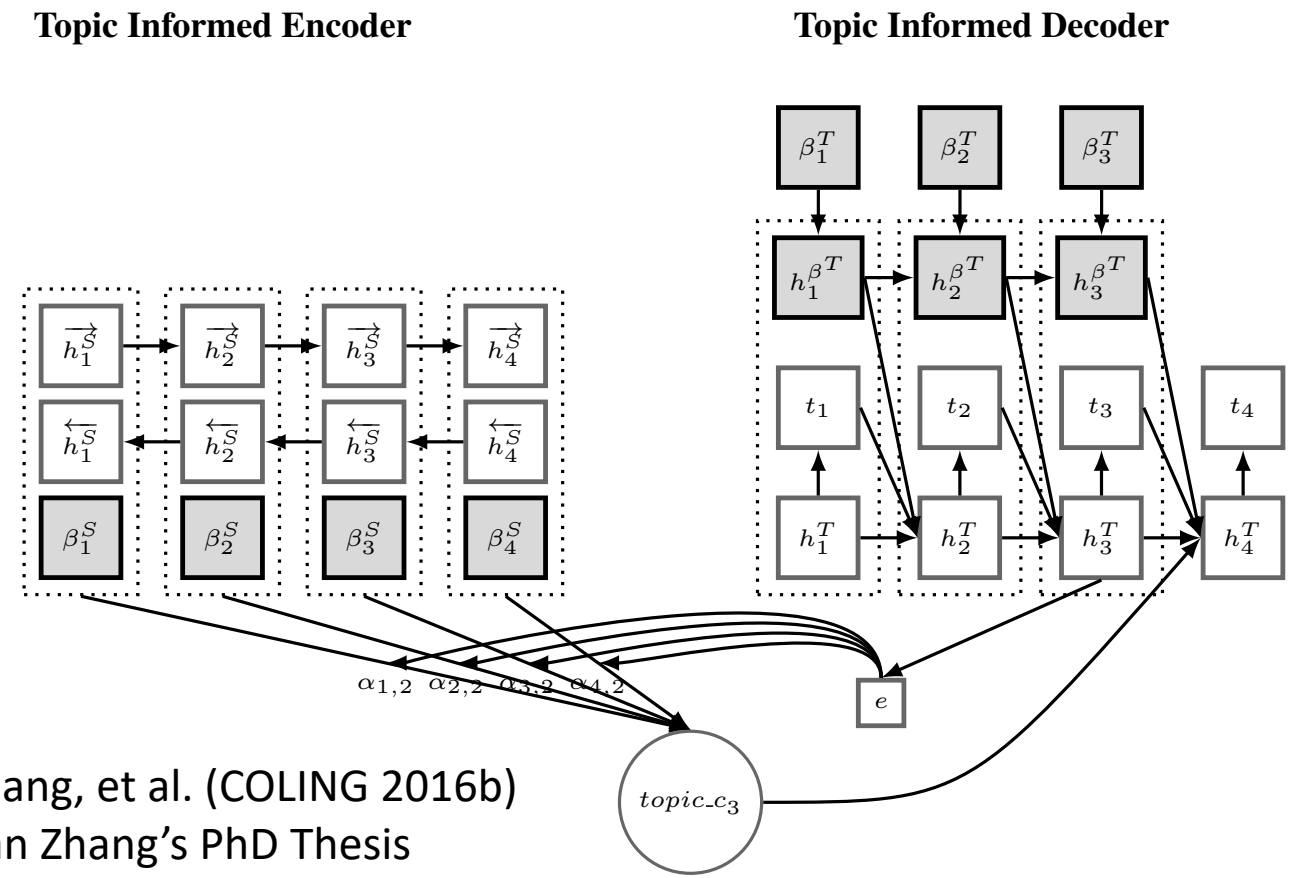
Domain Adaptation for NMT

System	NIST 2002	NIST 2004	NIST 2005
PBSMT	33.42	32.36	30.11
NMT	34.51	35.02	31.46
GDA NMT (glove_840b)	36.07	**35.99**‡	31.73
GDA NMT (*word2vec*)	35.63	35.84‡	**31.88**‡

Table 4.8: BLEU scores for NMT adaptation. We use ‡ to indicate statistically significant (Koehn 2004) improvements upon the NMT baseline model. The significance testing uses bootstrapping method at the p = 0.01 level with 1,000 iterations.

Domain Adaptation for NMT

- Topic-Informed Neural Machine Translation



Zhang, et al. (COLING 2016b)
Jian Zhang's PhD Thesis

Domain Adaptation for NMT

Systems	NIST 2002 (dev)	NIST 2004 (test)	NIST 2005 (test)
SMT	33.42	32.36	30.11
NMT	34.33	34.76	31.12
Source Topic-Informed NMT (40)	35.39	35.17†	31.95‡
Target Topic-Informed NMT (10)	36.31	35.43‡	32.50‡
Topic-Informed NMT (40,10)	34.86	35.91‡	32.79‡

Table 2: BLEU scores of the trained SMT and NMT models. We use ‡ and † to indicate significant (Koehn, 2004) improvements upon the baseline NMT using bootstrapping method at the level p = 0.01 and p = 0.05 level, respectively (with 1000 iterations).

	Baseline NMT	Topic-Informed NMT
NIST 2004	2.3%	1.9%
NIST 2005	2.7%	2.3%

Table 3: The percentage of *UNK* tokens produced in translation outputs by baseline NMT and topic-informed NMT systems.

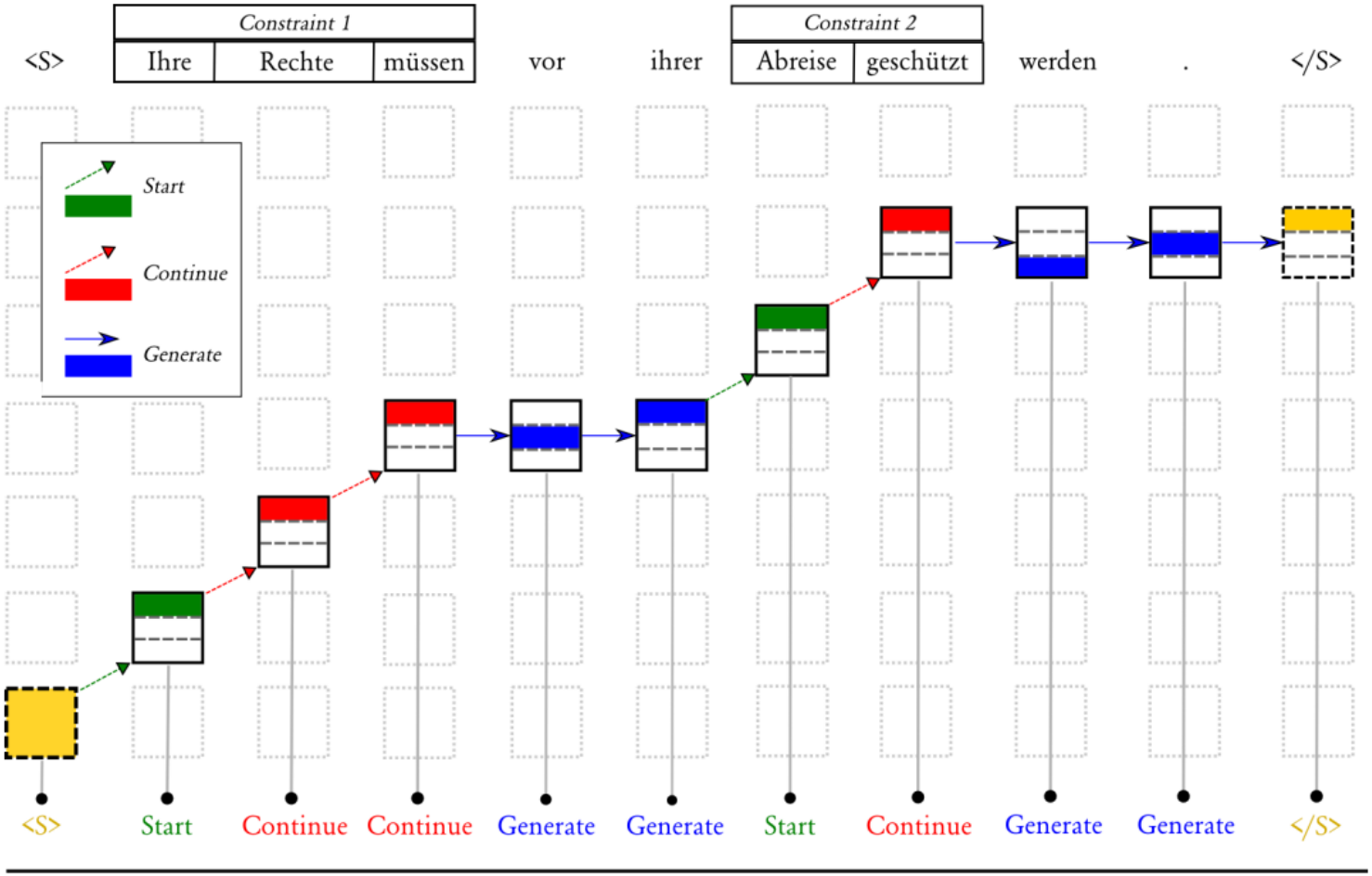
Our work on DL-based NLP

- Multimodal Machine Translation
- NMT with Morphological Information
- NMT with Discourse Information
- Domain Adaptation for NMT
- **Lexical Constrained Decoding for NMT**

Lexical Constrained Decoding for NMT

- In some scenario, we expect the MT system to generate translations with specific constraints:
  - Term translation
  - Pick-Revise-Translation loop for post-editing
- It is quite complex to implement this idea in SMT
- We proposed a simple and effective approach to implement this in NMT framework

Lexical Constrained Decoding for NMT



Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search
Hokamp & Liu (ACL 2017)

Lexical Constrained Decoding for NMT

ITERATION	0	1	2	3
Strict Constraints				
EN-DE	18.44	27.64 (+9.20)	36.66 (+9.01)	**43.92** (+7.26)
EN-FR	28.07	36.71 (+8.64)	44.84 (+8.13)	**45.48** +(0.63)
EN-PT*	15.41	23.54 (+8.25)	31.14 (+7.60)	**35.89** (+4.75)
Relaxed Constraints				
EN-DE	18.44	26.43 (+7.98)	34.48 (+8.04)	**41.82** (+7.34)
EN-FR	28.07	33.8 (+5.72)	40.33 (+6.53)	**47.0** (+6.67)
EN-PT*	15.41	23.22 (+7.80)	33.82 (+10.6)	**40.75** (+6.93)

Table 1: Results for four simulated editing cycles using WMT test data. EN-DE uses *newstest2013*, EN-FR uses *newstest2014*, and EN-PT uses the Autodesk corpus discussed in Section 4.2. Improvement in BLEU score over the previous cycle is shown in parentheses. * indicates use of our test corpus created from Autodesk post-editing data.

Lexical Constrained Decoding for NMT

System	BLEU
EN-DE	
Baseline	26.17
Random	25.18 (-0.99)
Beginning	26.44 (+0.26)
GBS	**27.99** (+1.82)
EN-FR	
Baseline	32.45
Random	31.48 (-0.97)
Beginning	34.51 (+2.05)
GBS	**35.05** (+2.59)
EN-PT	
Baseline	15.41
Random	18.26 (+2.85)
Beginning	20.43 (+5.02)
GBS	**29.15** (+13.73)

Table 2: BLEU Results for EN-DE, EN-FR, and EN-PT terminology experiments using the Autodesk Post-Editing Corpus. "Random' indicates inserting terminology constraints at random positions in the baseline translation. "Beginning" indicates prepending constraints to baseline translations.

Outline

The Changes brought by DL to NLP

Our work on DL-based NLP

Weakness of DL-based NLP and Future Direction

Weakness of DL-based NLP and Future Direction

- DL has totally changed the face of NLP research and provide powerful driving force for NLP
- In my personal opinion, DL-based NLP still has huge potential and is far from its plateau
- Weakness of DL-based NLP is also obvious:
  - Does not work with small data
  - Unawareness of semantic:
    - Most NLP system has no idea of what the language is talking about
    - There is no general techniques for task-oriented NLP

Weakness of DL-based NLP and Future Direction

- Future direction:
  - Apply more powerful DL technologies to solve NLP tasks
    - Reinforcement Learning
    - Adversarial Learning
  - General technologies to incorporate human knowledge to DL-based NLP
  - General technologies to incorporate semantics into DL-based NLP

Engaging Content
Engaging People

Thanks you for your attention

qun.liu@dcu.ie