A New General Deep Learning Approach for Natural Language Processing

Hui Jiang

Department of Electrical Engineering and Computer Science
York University, Toronto, Canada

Joint work with Shiliang Zhang (USTC), Quan Liu (USTC),
Mingbin Xu (York), Joseph Sanu (York)

Outline

iFLYTEK Laboratory for Neural Computing and Machine Learning (iNCML)

2054 LASSONDE SCHOOL OF ENGINEERING | YORK UNIVERSITY

- Introduction
  - Deep Learning for NLP
- A New General Approach: FOFE-net
  - FOFE: lossless encoding
  - DNNs: universal approximator
- NLP Applications
  - Word Embedding
  - Language Modelling
  - Named Entity Recognition and Mention Detection in KBP-EDL
  - Coreference Resolution in Winograd
- On-going work
  - Entity Linking
  - Syntactic/Semantic Parsing
- Conclusions

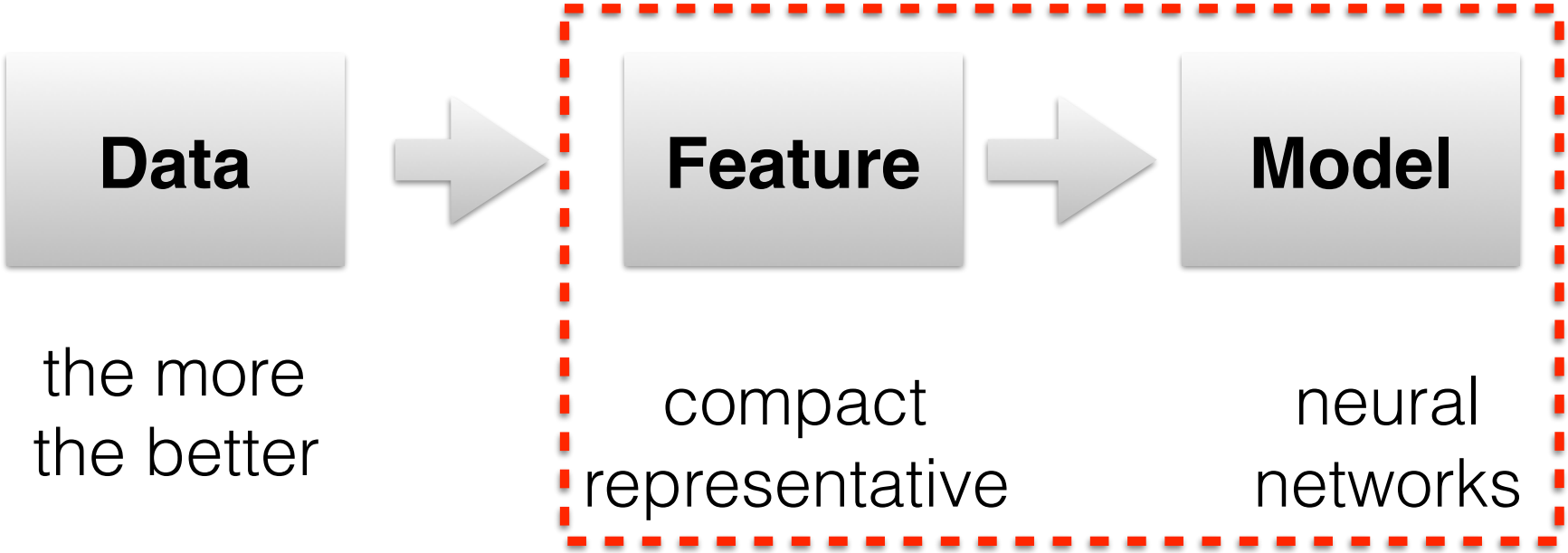
Deep Learning for NLP

the more
the better

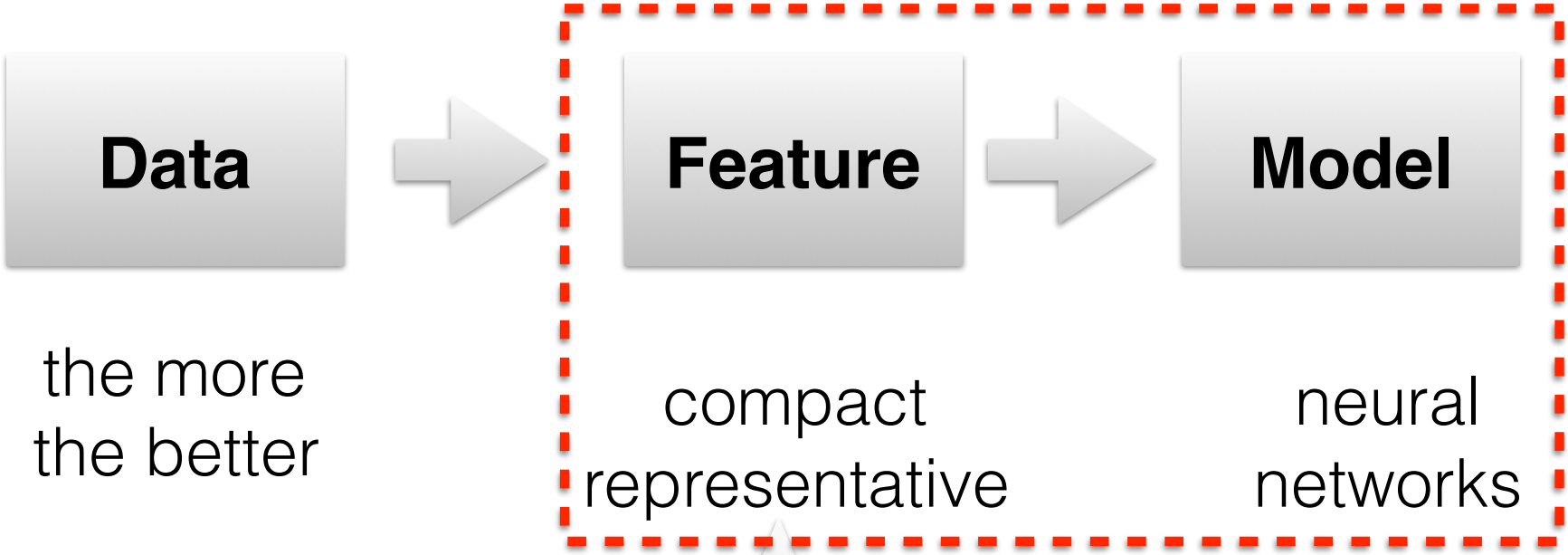
compact
representative

neural
networks

Deep Learning for NLP



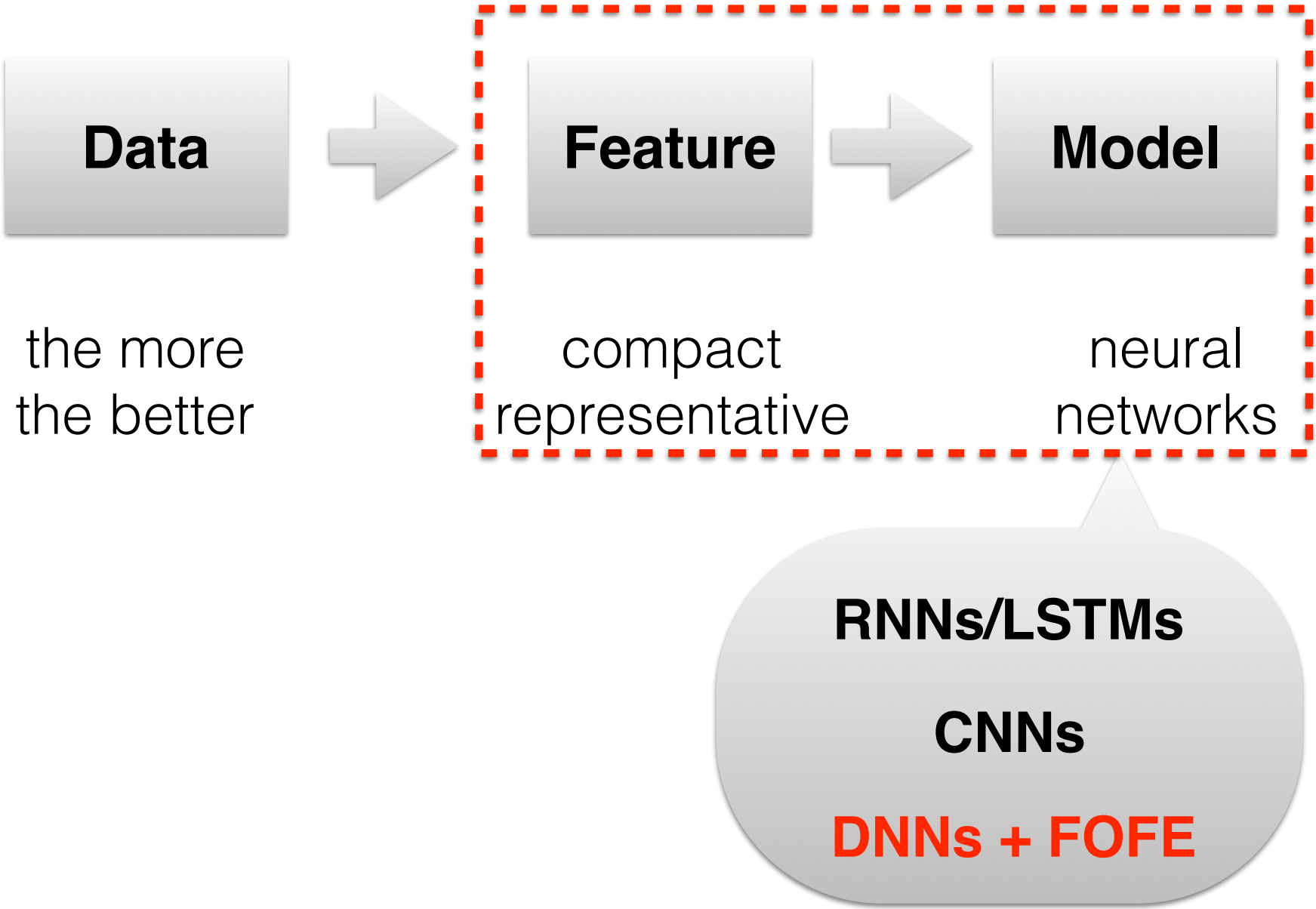
Deep Learning for NLP



Word: *word embedding*

sentence/paragraph/document:
variable-length *word sequences*

Deep Learning for NLP



Deep Learning for NLP: Difficulties

- How to handle ***variable-length sequences***?
  - Extracting fixed-size features —> information loss
  - Using CNN/RNN/LSTM as sequence labelling —> Not flexible in use; slow and hard to learn
- How to deal with ***data sparsity*** problem?
  - *Maybe* more training data + powerful models

Outline

- Introduction
  - Deep Learning for NLP
- **A New General Approach: FOFE-net**
  - **FOFE: lossless encoding**
  - **DNNs: universal approximator**
- NLP Applications
  - Word Embedding
  - Language Modelling
  - Named Entity Recognition and Mention Detection in KBP-EDL
  - Coreference Resolution in Winograd
- On-going work
  - Entity Linking
  - Syntactic/Semantic Parsing
- Conclusions

Fixed-size Ordinally-Forgetting Encoding (FOFE)

- **FOFE**: encode any ***variable-length sequence*** into a ***fixed-size representation***

  - Unique and invertible —> no information loss (theoretically guaranteed)
  - Elegant and automatic —> no feature engineering

WORD	1-OF-K
w_0	1000000
w_1	0100000
w_2	0010000
w_3	0001000
w_4	0000100
w_5	0000010
w_6	0000001

ANY SEQUENCE	FOFE
w_6	$0,0,0,0,0,0,1$
w_6, w_4	$0,0,0,0,1,0,\alpha$
w_6, w_4, w_5	$0,0,0,0,\alpha,1,\alpha^2$
w_6, w_4, w_5, w_0	$1,0,0,0,\alpha^2,\alpha,\alpha^3$
w_6, w_4, w_5, w_0, w_5	$\alpha,0,0,0,\alpha^3,1+\alpha^2,\alpha^4$
$w_6, w_4, w_5, w_0, w_5, w_4$	$\alpha^2,0,0,0,1+\alpha^4,\alpha+\alpha^3,\alpha^5$

$$z_t = \alpha \cdot z_{t-1} + e_t \; (1 \leq t \leq n)$$

FOFE: theoretical guarantee

- FOFE is ***almost unique*** for any ***α*** *(0<**α**<1)*:

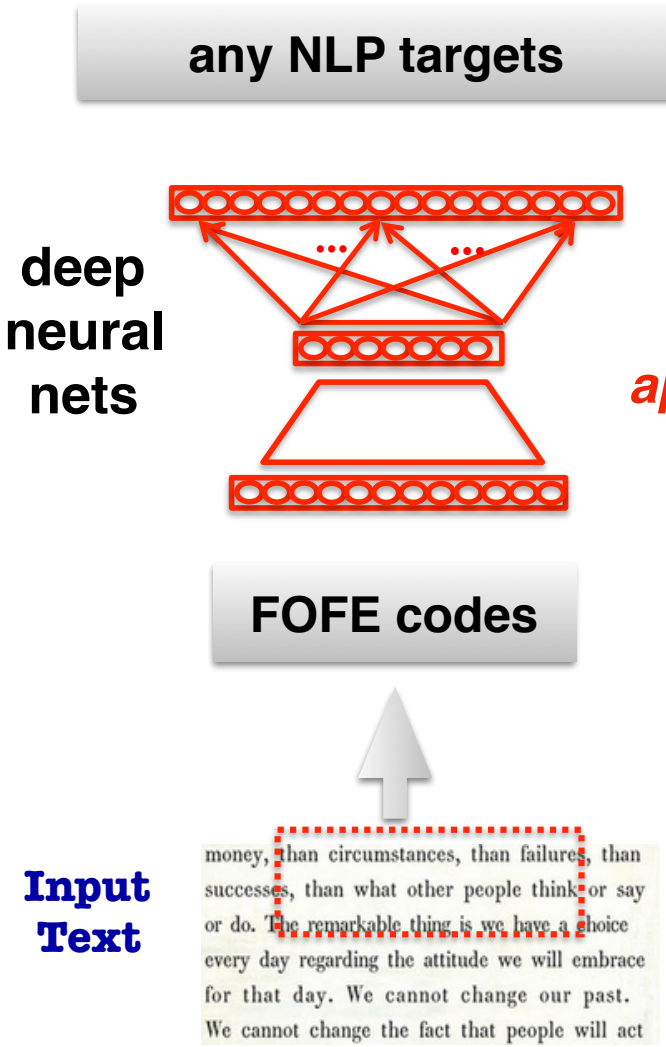
Theorem 1 *If the forgetting factor* α *satisfies* $0 < \alpha \leq 0.5$, FOFE *is unique for any* K *and* T.

Theorem 2 *For* $0.5 < \alpha < 1$, *given any finite values of* K *and* T, FOFE *is almost unique everywhere for* $\alpha \in (0.5, 1.0)$, *except only a finite set of countable choices of* α.

A New General Deep Learning Approach for Natural Language Processing

- **FOFE-based lossless encoding**
  - **any text input —> a fixed-size FOFE vector**
- **Deep neural networks (DNN):**
  - **universal approximator *(Cybenko, '89)*: learned to map to any NLP targets**
- **Labelled training data:**
  - **human labelling; semi-supervised; distant supervision**

FOFE-net for NLP



- Theoretically sound
- No feature engineering
- General methodology
  - (almost) all NLP tasks
- Enough labelled data
  - human performance (?)

Applicable NLP tasks

- Word Embedding
- Chunking
  - Tokenization
  - Segmentation
- Word Sense Disambiguation
- POS Tagging
- Named Entity Recognition
- Semantic Role Labelling
- Syntactic Parsing
- Semantic Parsing
- Language Modelling
- Text Categorization
- Coreference Resolution
- Information Extraction
- Relationship Extraction
- more …

Applicable NLP tasks

- **Word Embedding**
- **Chunking**
  - **Tokenization**
  - **Segmentation**
- Word Sense Disambiguation
- POS Tagging
- **Named Entity Recognition**
- Semantic Role Labelling
- **Syntactic Parsing**
- **Semantic Parsing**
- **Language Modelling**
- **Text Categorization**
- **Coreference Resolution**
- Relationship Extraction
- **Information Extraction**
- more …

Applicable NLP tasks

- **Word Embedding**
- **Chunking**
  - **Tokenization**
  - **Segmentation**
- Word Sense Disambiguation
- POS Tagging
- **Named Entity Recognition**
- Semantic Role Labelling
- **Syntactic Parsing**
- **Semantic Parsing**
- **Language Modelling**
- **Text Categorization**
- **Coreference Resolution**
- Relationship Extraction
- **Information Extraction**
- more …

Outline

iFLYTEK Laboratory for Neural Computing and Machine Learning (iNCML)

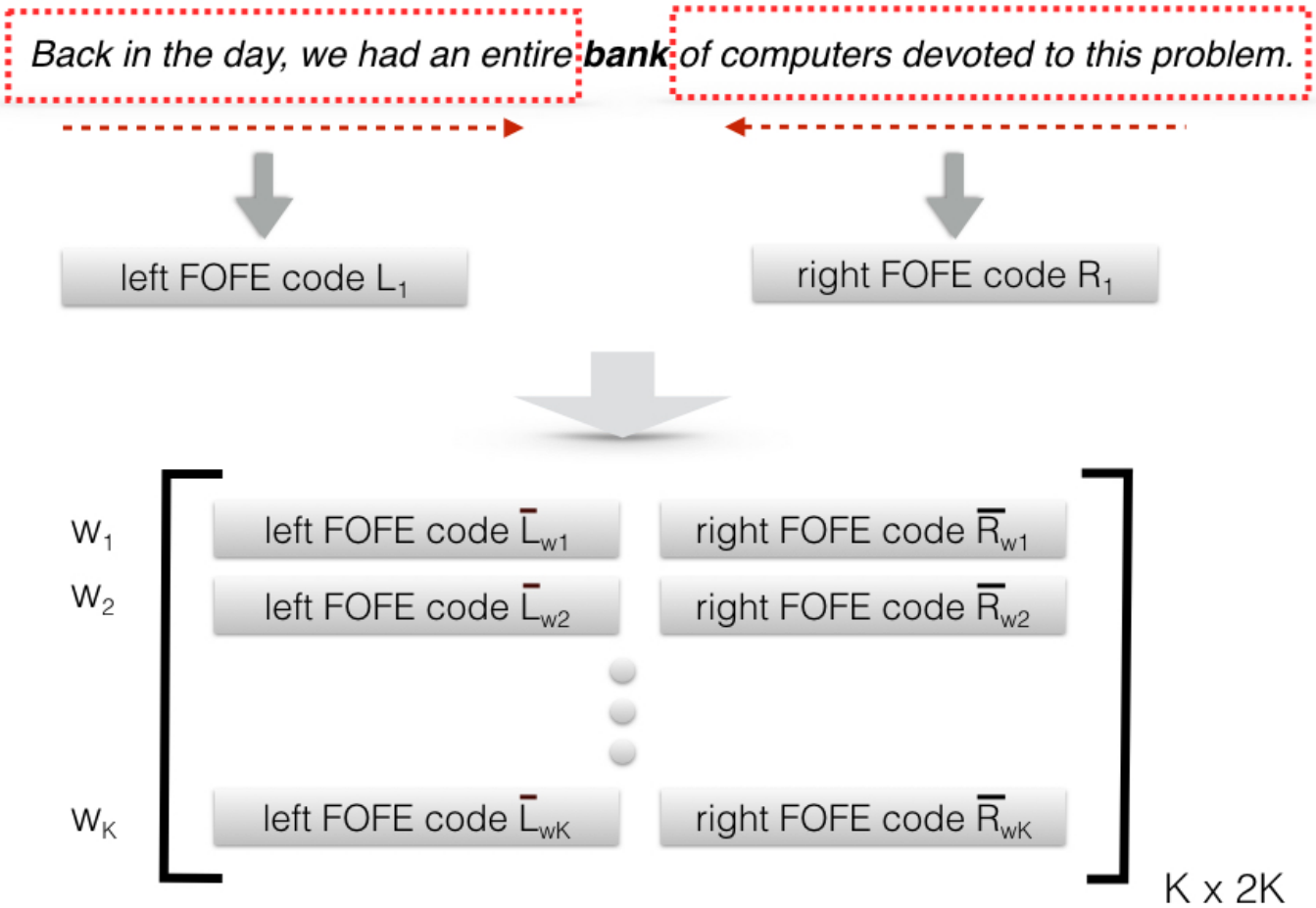
2054 LASSONDE SCHOOL OF ENGINEERING | YORK UNIVERSITY

- Introduction
  - Deep Learning for NLP
- A New General Approach: FOFE-net
  - FOFE: lossless encoding
  - DNNs: universal approximator
- **NLP Applications**
  - **Word Embedding**
  - **Language Modelling**
  - **Named Entity Recognition and Mention Detection in KBP-EDL**
  - **Coreference Resolution in Winograd**
- On-going work
  - Entity Linking
  - Syntactic/Semantic Parsing
- Conclusions

FOFE for Word Embedding

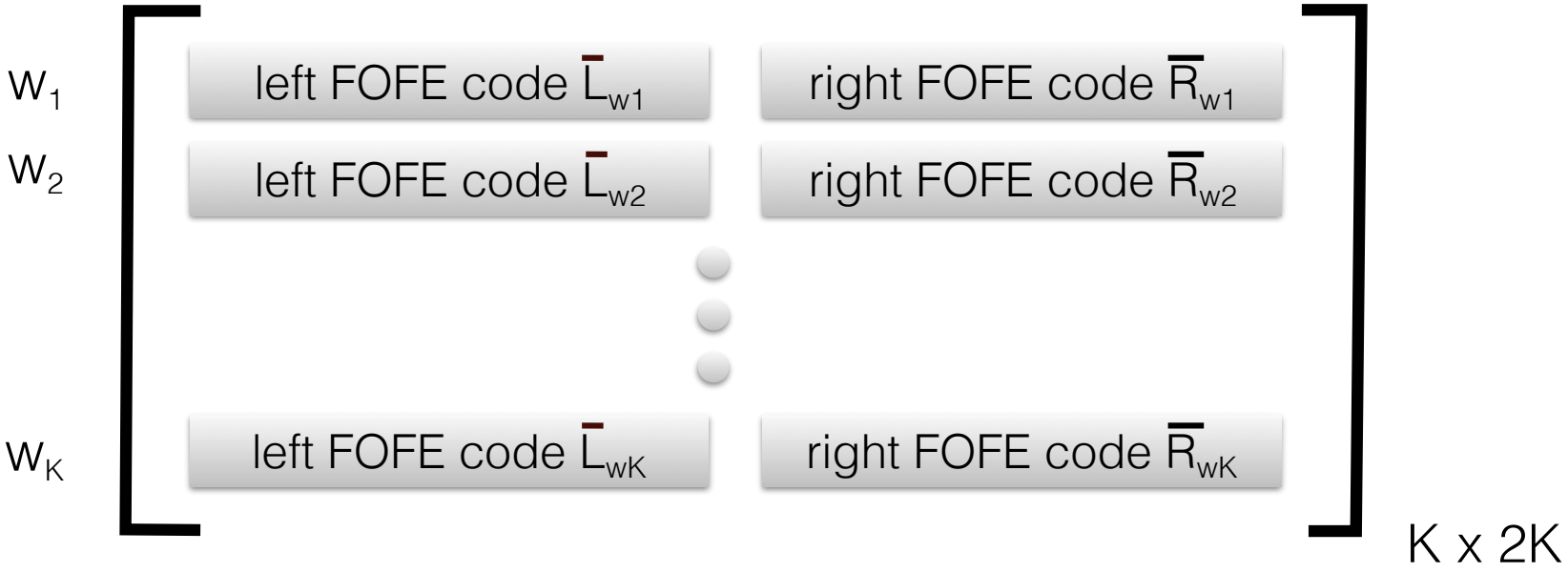
- **Distributional hypothesis** *(Harris, 1954)*: words that appear in similar **contexts** have similar meaning.

- ***Using FOFE to model contexts:***

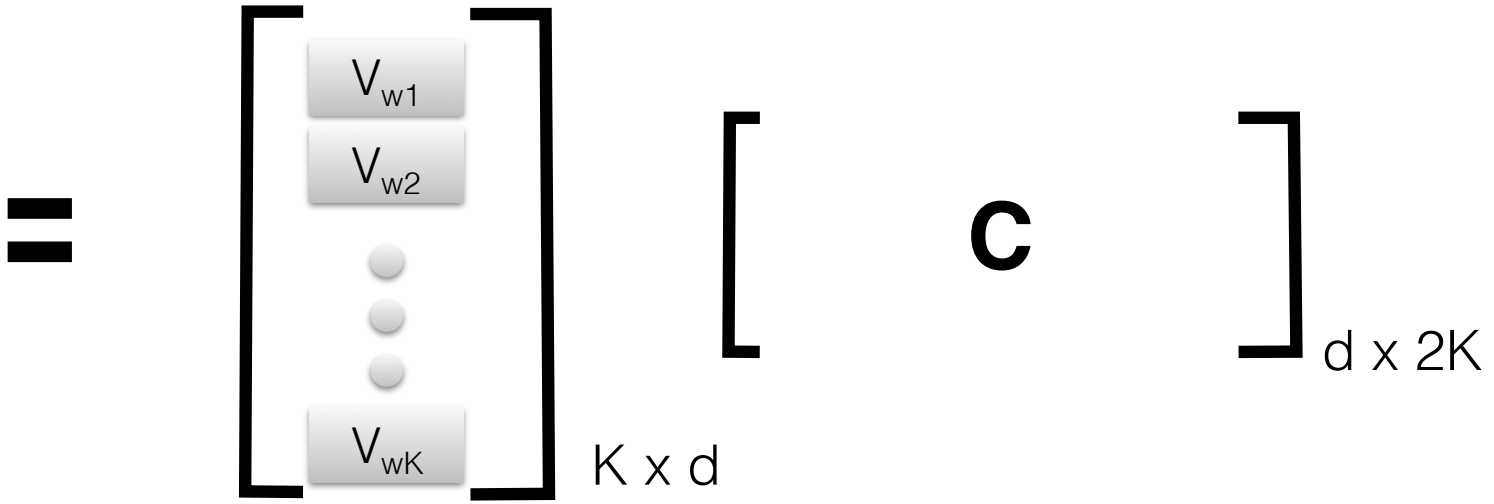


FOFE for Word Embedding

- Generate word-context (the averaged FOFE) matrix (one word per line):



- Stochastic (online) truncated SVD for sparse matrices



Experiments: FOFE for Word Embedding

- Training Corpus: ***enwiki9*** (130 million words)
- Vocabulary: 80,000 words; Truncated dimension: 300
- FOFE matrices are weighted by ***PMI***; Use ***scipy*** for truncated SVD
- Evaluated on five popular word similarity tasks:

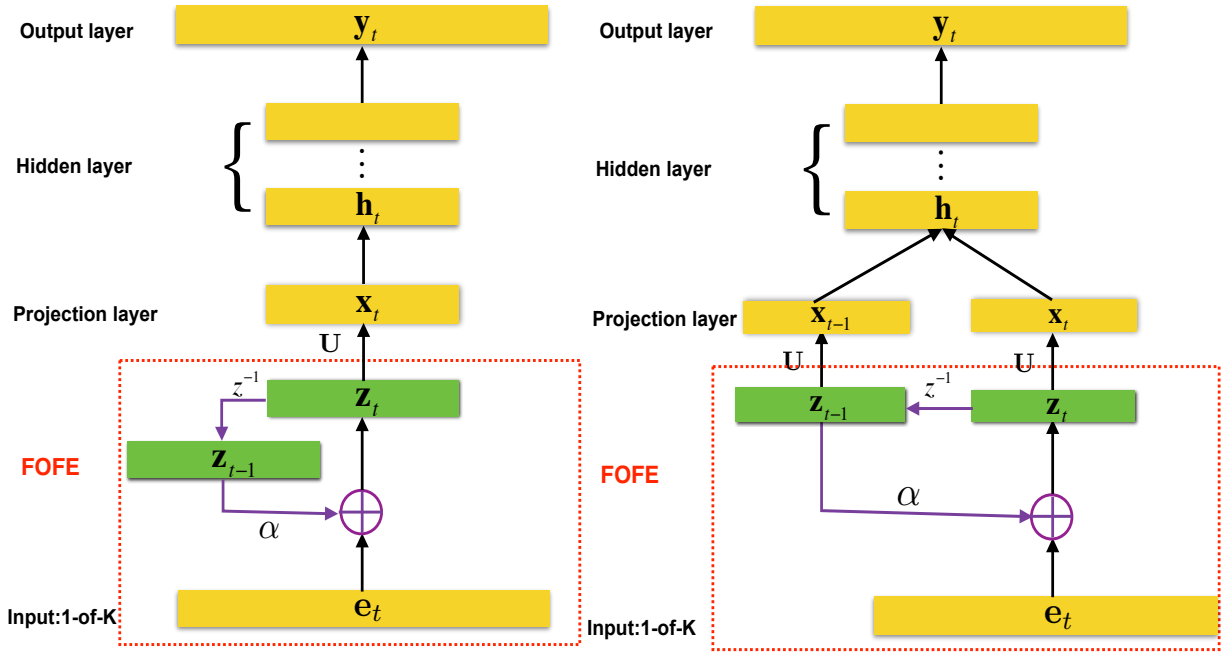
Method	**WordSim353**	**MEN**	**Mech Turk**	**Rare Words**	**SimLex-999**
VSM+SVD	0.71	0.71	0.63	0.48	0.39
CBOW	0.68	0.68	0.66	0.43	0.35
SGNS	0.70	0.67	0.62	0.44	0.37
GloVe	0.59	0.64	0.58	0.39	0.29
Swivel	0.73	0.72	0.70	0.44	0.33
FOFE+SVD	**0.76**	**0.76**	**0.65**	**0.50**	**0.39**

Pearson correlation coefficients with human scores

FOFE-net for Language Modelling

- Use FOFE to ***recursively*** encode ***partial sequence (history)*** to predict next word

$$\mathbf{z}_t = \alpha \cdot \mathbf{z}_{t-1} + \mathbf{e}_t \quad (1 \le t \le T)$$



FOFE-net for Language Modelling

- Formulate FOFE as efficient **matrix multiplications**
- Much faster than RNN based language models

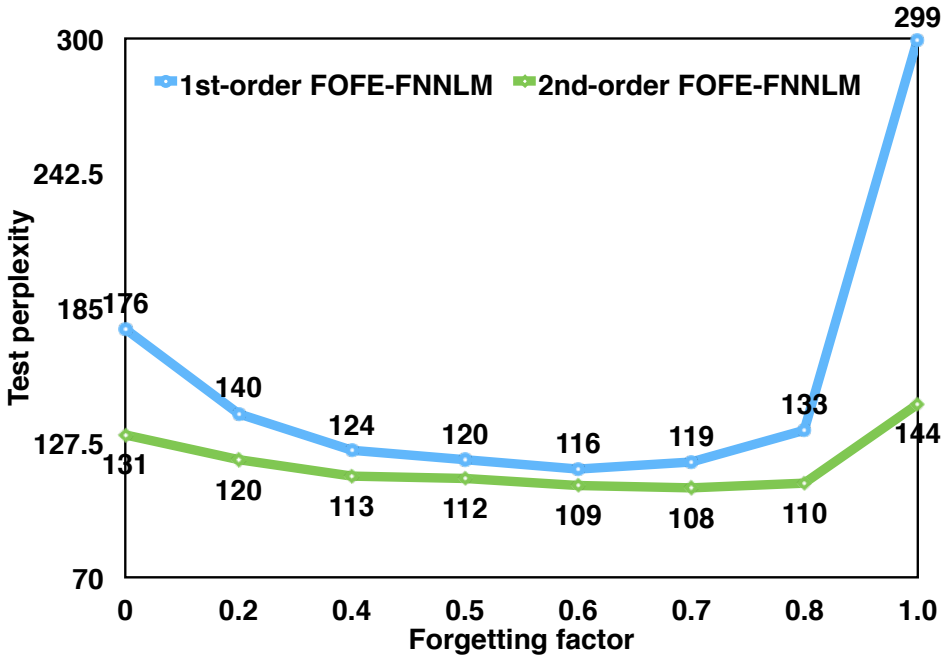
$$\mathbf{S} = \begin{bmatrix} 1 & & & \\ \alpha & 1 & & \\ \alpha^2 & \alpha & 1 & \\ \vdots & & \ddots & 1 \\ \alpha^{T-1} & \cdots & \alpha & 1 \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ \vdots \\ \mathbf{e}_T \end{bmatrix} = \mathbf{MV}$$

$$\bar{\mathbf{S}} = \begin{bmatrix} \mathbf{M}_1 & & & \\ & \mathbf{M}_2 & & \\ & & \ddots & \\ & & & \mathbf{M}_N \end{bmatrix} \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \vdots \\ \mathbf{V}_N \end{bmatrix} = \bar{\mathbf{M}}\bar{\mathbf{V}}.$$

$$\mathbf{H} = f\Big((\bar{\mathbf{M}}\bar{\mathbf{V}})\mathbf{U}\mathbf{W} + \mathbf{b}\Big) = f\Big(\bar{\mathbf{M}}(\bar{\mathbf{V}}\mathbf{U})\mathbf{W} + \mathbf{b}\Big)$$

Experiments: FOFE for Language Modelling

- **Penn Treebank (PTB)** data set:
  - Training 930K words; valid 74K words; test 82K words
  - Vocabulary 10K words; ***no drop-out***



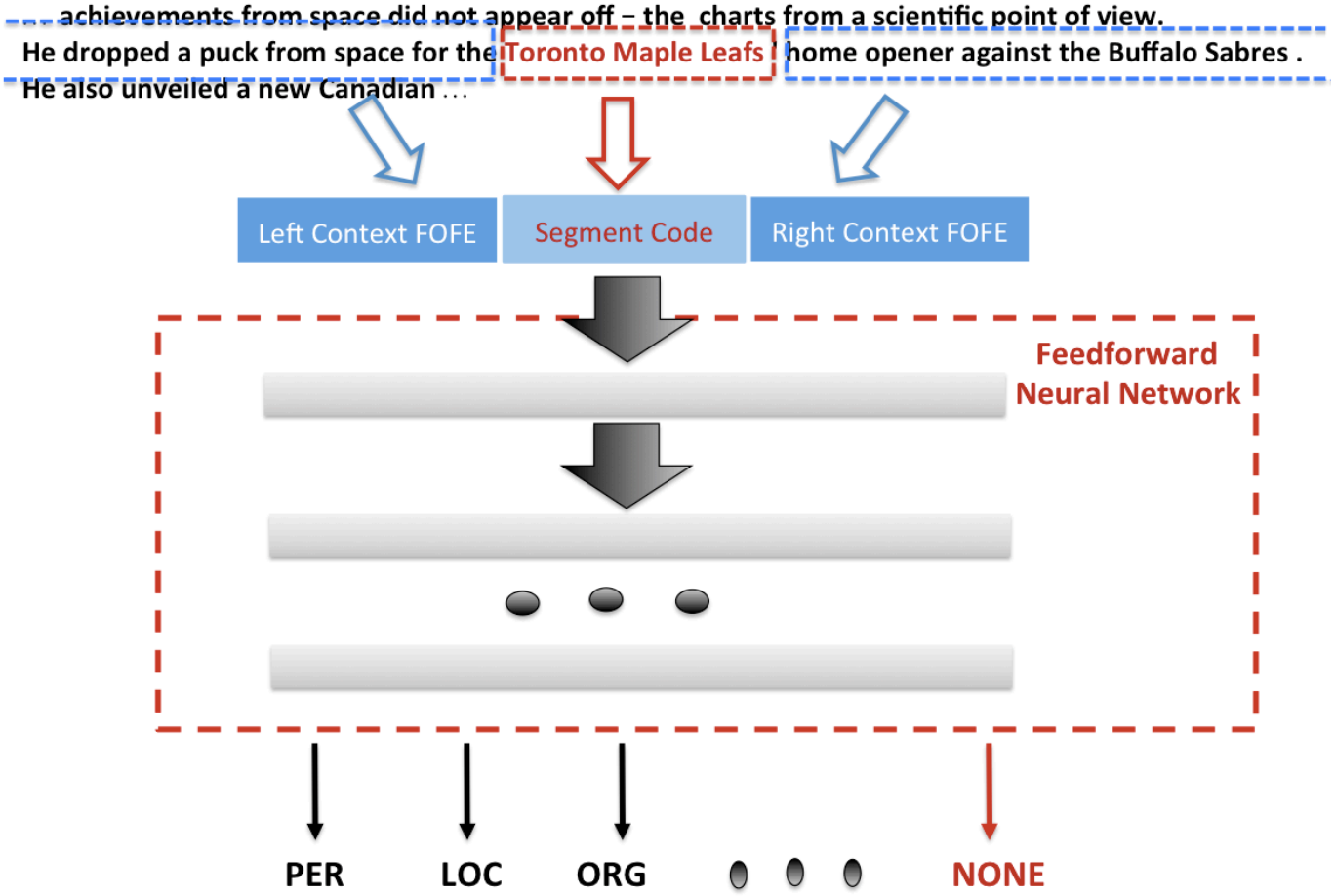
Model	Test PPL
KN 5-gram (Mikolov, 2011)	141
FNNLM (Mikolov, 2011)	140
RNNLM (Mikolov, 2011)	123
LSTM (Graves, 2013)	117
Bigram FNNLM	176
Trigram FNNLM	131
4-gram FNNLM	118
5-gram FNNLM	114
6-gram FNNLM	113
1st-order FOFE-FNNLM	116
2nd-order FOFE-FNNLM	106

Experiments: FOFE for Language Modelling

- **Enwiki9** data set:
  - Training 153M words; valid 8.9M words; test 8.9M words
  - Vocabulary 80K words; ***no drop-out***

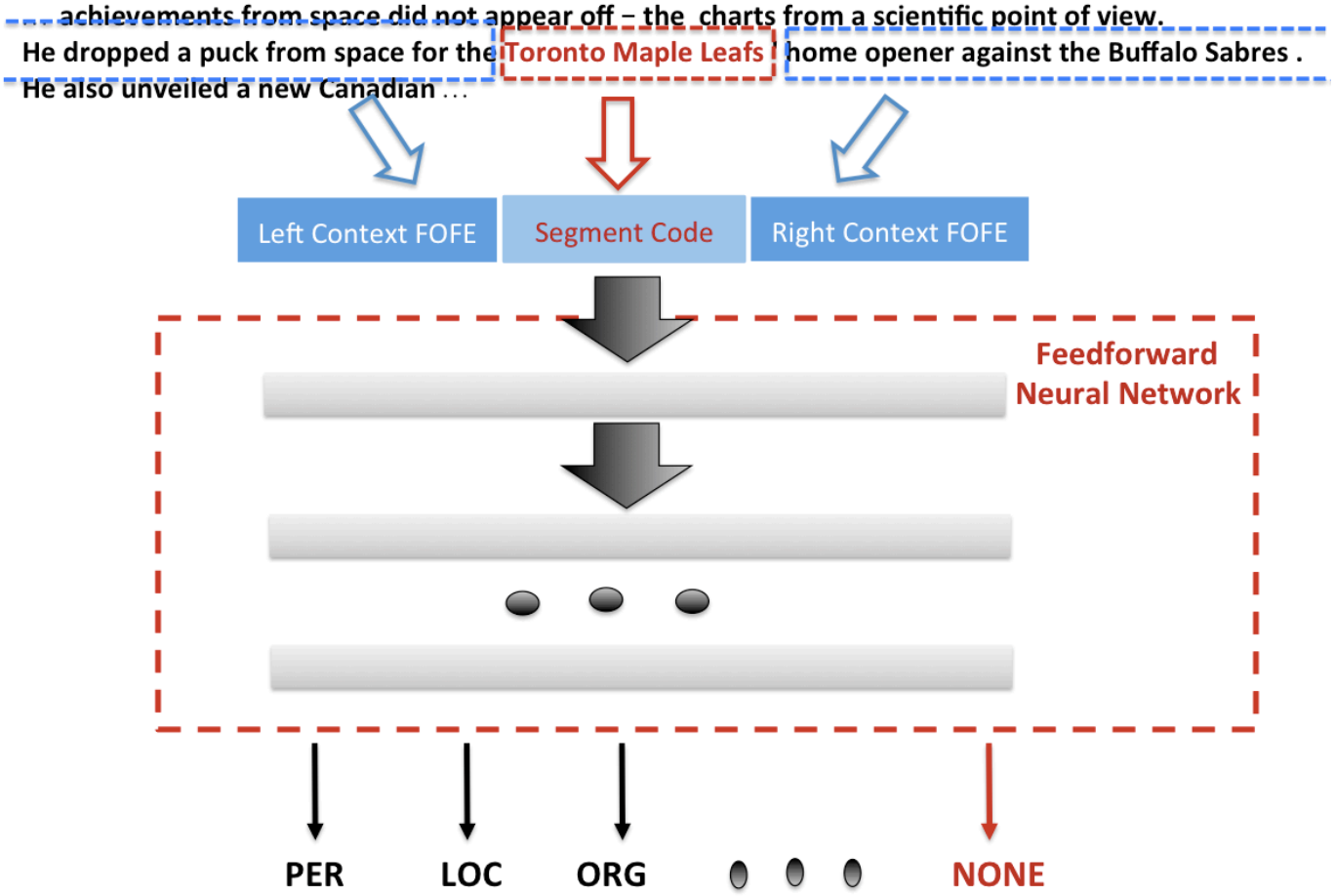
Model	Architecture	Test PPL
KN 3-gram	-	156
KN 5-gram	-	132
FNN-LMs	[1*200]-400-400-80k	241
	[2*200]-400-400-80k	155
	[2*200]-600-600-80k	150
	[3*200]-400-400-80k	131
	[4*200]-400-400-80k	125
RNN-LM	[1*600]-80k	112
FOFE FNN-LMs	[1*200]-400-400-80k	120
	[1*200]-600-600-80k	115
	[2*200]-400-400-80k	112
	[2*200]-600-600-80k	107

FOFE-net for Named Entity Recognition



- **Local detection**: no Viterbi decoding; Nested/Embedded entities
- No feature engineering: FOFE codes
- Easy and fast to train; make use of partial labels

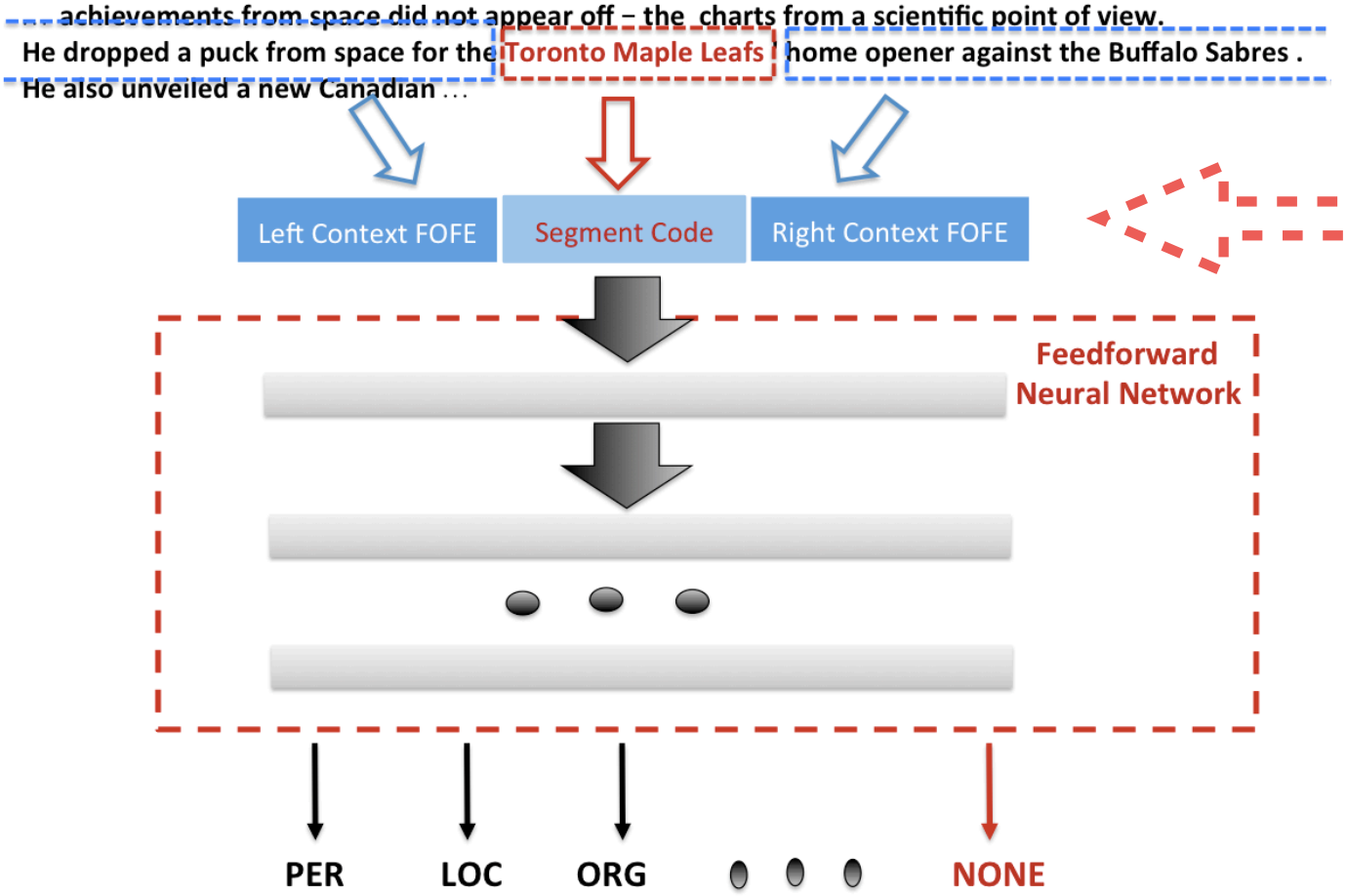
FOFE-net for Named Entity Recognition



- FOFE code for left context
- FOFE code for right context
- BoW vector
- Char FOFE code

- **Local detection**: no Viterbi decoding; Nested/Embedded entities
- No feature engineering: FOFE codes
- Easy and fast to train; make use of partial labels

FOFE-net for Named Entity Recognition

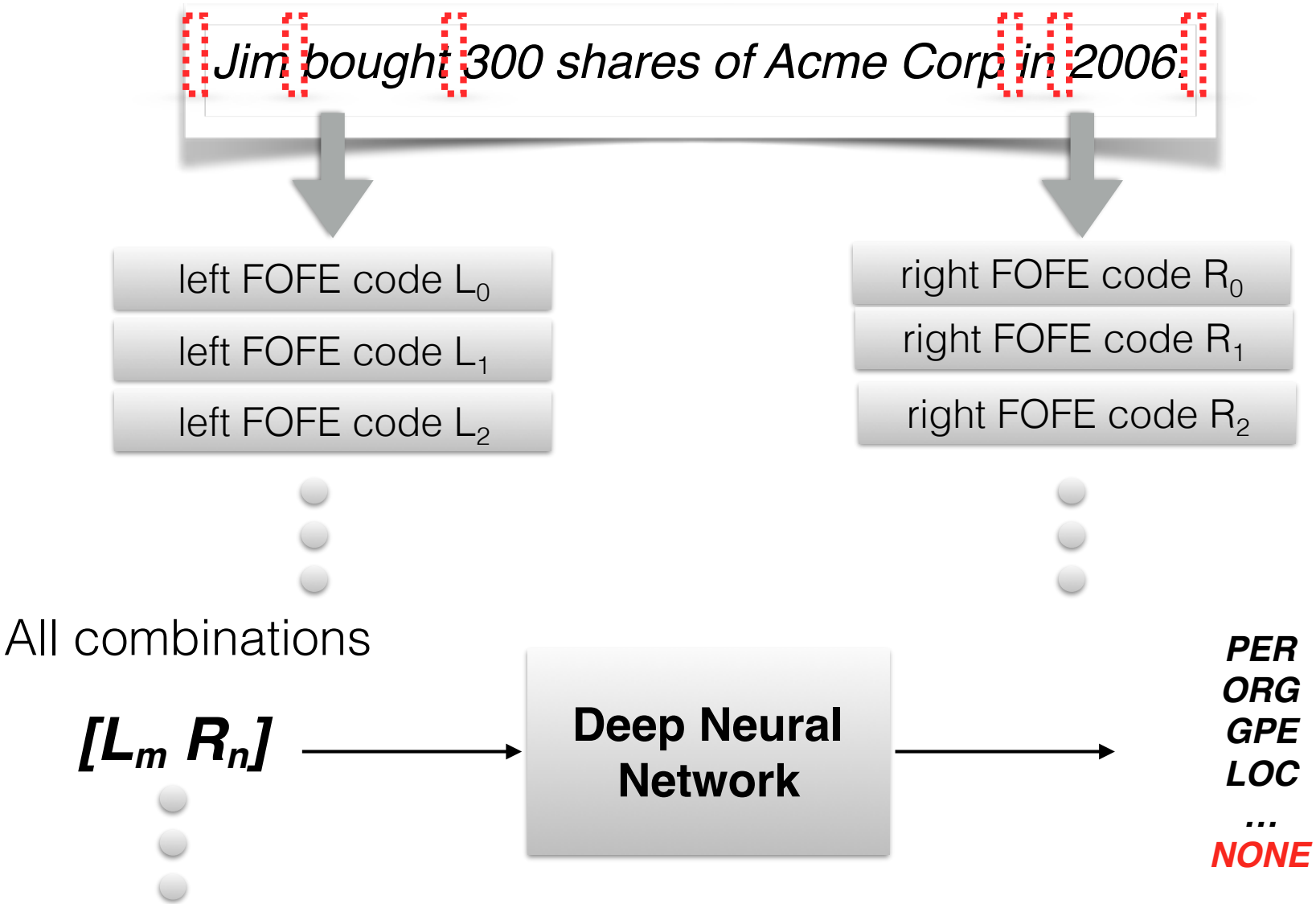


- FOFE code for left context
- FOFE code for right context
- BoW vector
- Char FOFE code

- **Local detection**: no Viterbi decoding; Nested/Embedded entities
- No feature engineering: FOFE codes
- Easy and fast to train; make use of partial labels

FOFE-net for Named Entity Recognition

- **In training: labelled entities + negative sampling**
- **In test: consider all spans up to 7-8 words in a sentence**



Experiments (I): FOFE-net for NER

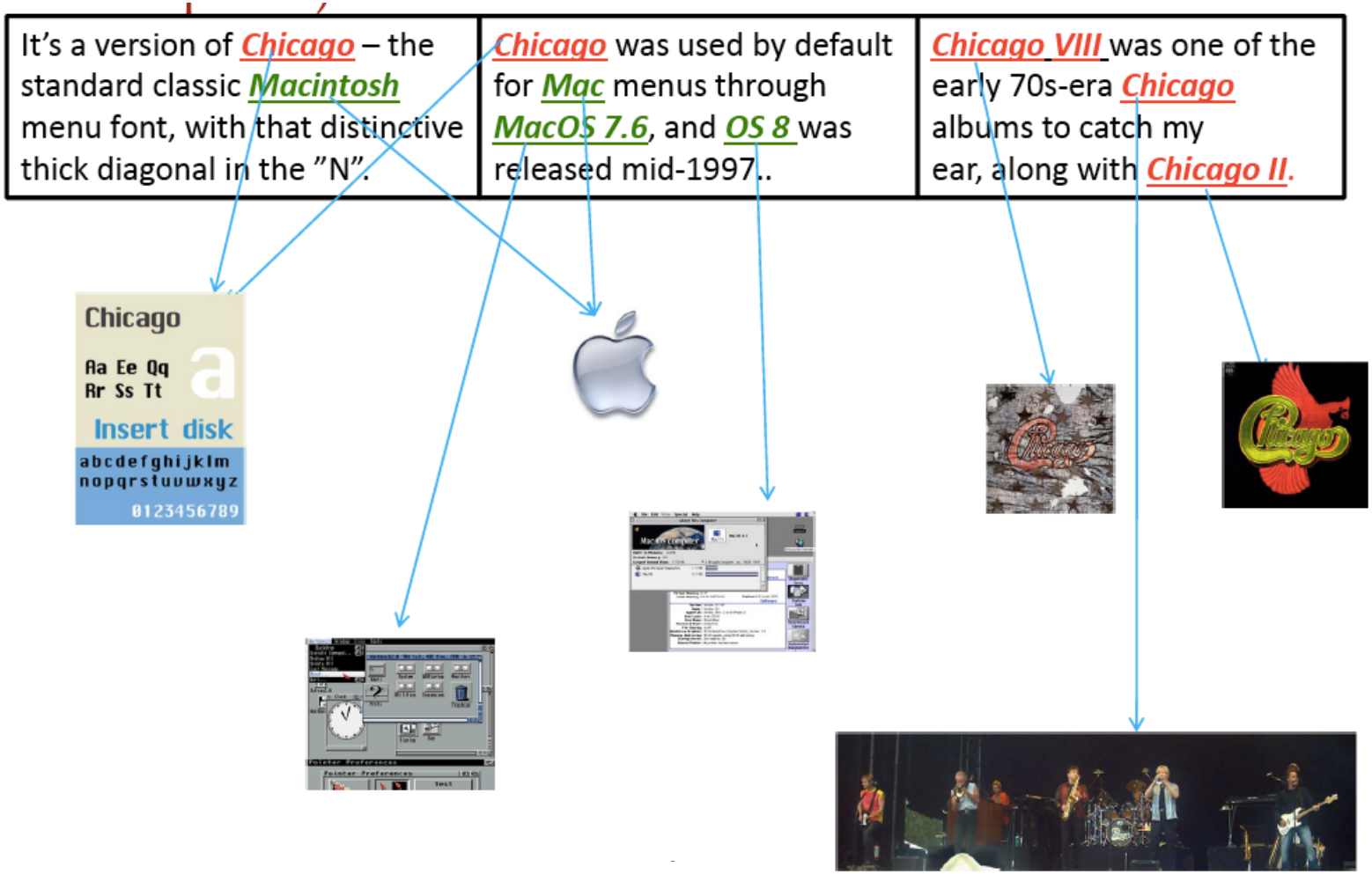
- **Performance comparison on the CoNLL03 test set**

FEATURE			P	R	F1
word-level	case-insensitive	context FOFE incl. focus word(s)	86.64	77.04	81.56
		context FOFE excl. focus word(s)	53.98	42.17	47.35
		BoW of focus word(s)	82.92	71.85	76.99
	case-sensitive	context FOFE incl. focus word(s)	88.88	79.83	84.12
		context FOFE excl. focus word(s)	50.91	42.46	46.30
		BoW of focus word(s)	85.41	74.95	79.84
char-level	Char FOFE of focus word(s)		67.67	52.78	59.31
	Char CNN of focus word(s)		78.93	69.49	73.91
all case-insensitive features			90.11	82.75	86.28
all case-sensitive features			90.26	86.63	88.41
all word-level features			92.03	86.08	88.96
all word-level & Char FOFE features			91.68	88.54	**90.08**
all word-level & Char CNN features			91.80	88.58	**90.16**
all word-level & all char-level features			93.29	88.27	**90.71**

	word	char	gaz	cap	pos	F1
(Collobert et al., 2011)	✓	✗	✓	✓	✗	89.59
(Huang et al., 2015)	✓	✓	✓	✓	✓	90.10
(Rondeau and Su, 2016)	✓	✗	✓	✓	✓	89.28
(Chiu and Nichols, 2016)	✓	✓	✓	✗	✗	**91.62**
this work	✓	✓	✗	✗	✗	**90.71**

TAC-KBP EDL Contest

- Information extraction from unstructured text (Wikification)
- Use ***FOFE-net*** for entity discovery and mention detection



Experiments (II): 2015 KBP-EDL

- **Pre-contest results: entity discovery performance comparison on 2015 KBP-EDL data set**
- **Trilingual: English, Chinese, Spanish**
- **5 named entity types (PER, GPE, LOC, ORG, FAC) + 1 nominal mention (PER)**

	2015 track best			ours		
	P	R	F_1	P	R	F_1
Trilingual	75.9	69.3	72.4	78.3	69.9	**73.9**
English	79.2	66.7	**72.4**	77.1	67.8	72.2
Chinese	79.2	74.8	**76.9**	79.3	71.7	75.3
Spanish	78.4	72.2	75.2	79.9	71.8	**75.6**

Experiments (III): 2016 KBP-EDL official results

- **2016 KBP Trilingual EDL track: FOFE-net ranked No.2 among all participating teams**
- **5 named entity types (PER, GPE, LOC, ORG, FAC) + all 5 nominal mentions**

LANG	NAME			NOMINAL			OVERALL		
	P	R	F1	P	R	F1	P	R	F1
	RUN1 (our official ED result in KBP2016 EDL2)								
ENG	0.898	0.789	0.840	0.554	0.336	0.418	0.836	0.680	0.750
CMN	0.848	0.702	0.768	0.414	0.258	0.318	0.789	0.625	0.698
SPA	0.835	0.778	0.806	0.000	0.000	0.000	0.835	0.602	0.700
ALL	0.893	0.759	0.821	0.541	0.315	0.398	0.819	0.639	**0.718**

FOFE-net for Coreference Resolution

[Blumenthal (D)]$_{per}$ is a candidate for [the U.S.]$_{org}$ Senate seat now held by [Christopher Dodd (D)]$_{org}$, and he has held a commanding lead in the race since he entered it.

FOFE code $[L_x, R_x]$

Entity buffer

[Christoper Dodd (D)]$_{per}$ — FOFE code $[L_2, R_2]$

[the U.S.]$_{org}$

[Blumenthal (D)]$_{per}$ — FOFE code $[L_1, R_1]$

$[L_x, R_x, L_1, R_1]$

$[L_x, R_x, L_2, R_2]$

→ **Deep Neural Network** →

y_1

y_2

sigmoid matching score

Winograd Schema Challenge (WSC)

- **WSC**: alternative Turing test for machine intelligence

The customer walked into the bank and stabbed one of the tellers.

He *was immediately taken to the* ***emergency***.

Who was taken to the emergency room? The customer / the teller.

The customer walked into the bank and stabbed one of the tellers.

He *was immediately taken to the* ***police***.

Who was taken to the emergency room? The customer / the teller.

2016 Winograd Schema Challenge

- **Pronoun Disambiguation Problems (PDP)**: the first round qualifying test in 2016 WS challenge

*Tom handed over the blueprints he had grabbed and, while his companion spread them out on **his** knee, walked toward the yard.*

Tom / companion

*One chilly May evening the English tutor invited Marjorie and myself into **her** room.*

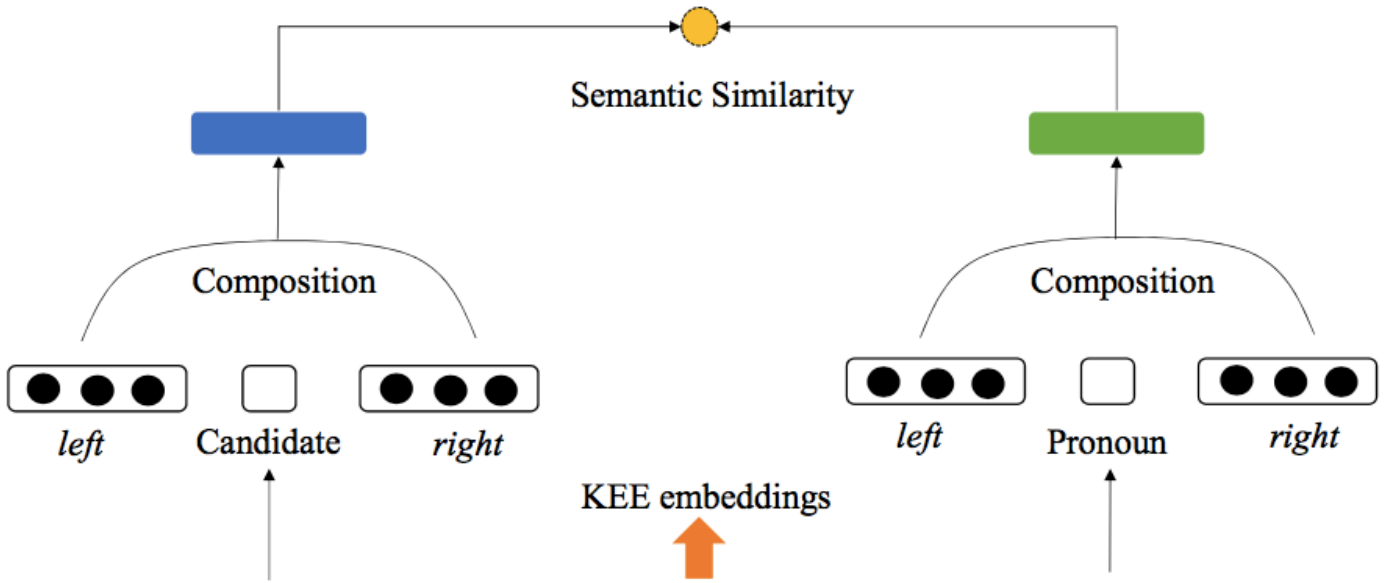
the English tutor / Marjorie

*Mariano fell with a crash and lay stunned on the ground. Castello instantly kneeled by his side and raised **his** head.*

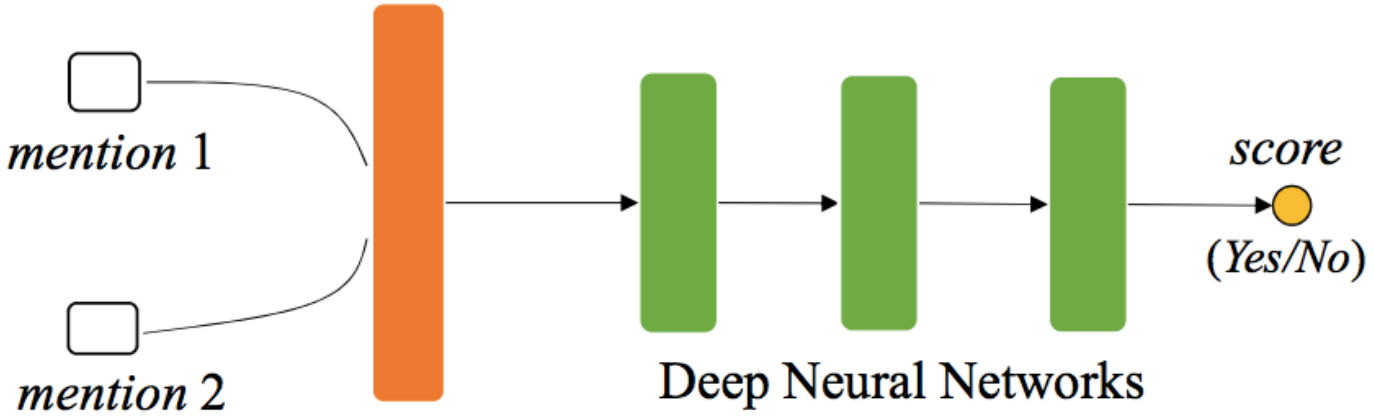
Mariano / Castello

FOFE-net for PDP

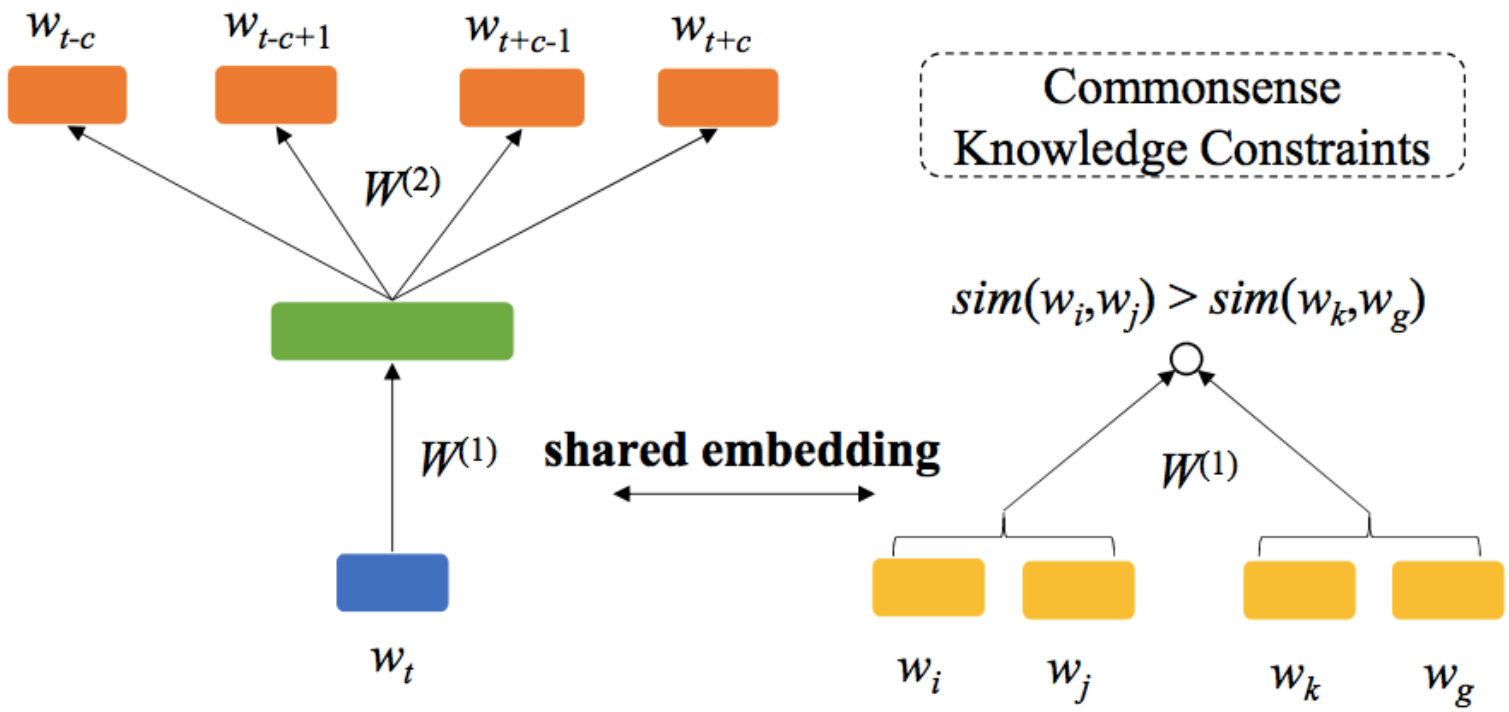
USSM:



NKAM:



Incorporating Common-sense Knowledge



1. For each triple in ConceptNet:

$$(w_h, r, w_t) \Rightarrow \text{sim}(w_h, w_t) > \text{sim}(w_h, w_k) \quad w_k \in V \text{ and } w_k \text{ is not linked with } w_h$$

2. For each cause-effect pair:

$$(w_i, w_j) \Rightarrow \text{sim}(w_i, w_j) > \text{sim}(w_i, w_k) \quad w_k \in V \text{ and } w_k \text{ is not the effect of } w_i$$

Experiments (I): FOFE-net for PDP

- Official results of the top systems on the PDP test set of the 2016 Winograd Schema Challenge

Systems	Accuracy
Random guess	45%
Denis Robert	31.7%
Patrick Dhondt	45.0%
Nicos Issak	48.3%
Quan Liu	**58.3%**

Experiments (II): FOFE-net for PDP

- Post-contest results: more training data; more tuning

Text Corpus	Problem Solver	KEE settings Accuracy		
		KEE training sources	Accuracy (%)	Improvements (%)
Wikipedia	USSM	Context	48.3	
		Context + ConcepNet	55.0	+13.9
		Context + WordNet	53.3	+10.4
		Context + CauseCom	55.0	+13.9
		Context + All KBs	**56.7**	**+17.4**
	NKAM	Context	51.7	
		Context + ConcepNet	60.0	+16.0
		Context + WordNet	60.0	+16.0
		Context + CauseCom	61.7	+19.3
		Context + All KBs	**63.3**	**+22.4**
	USSM+NKAM	Context	53.3	
		Context + ConcepNet	63.3	+18.7
		Context + WordNet	61.7	+15.7
		Context + CauseCom	65.0	+21.9
		Context + All KBs	**66.7**	**+25.1**

Experiments (II): FOFE-net for PDP

- Post-contest results: more training data; more tuning

Text Corpus	Problem Solver	KEE settings Accuracy		
		KEE training sources	Accuracy (%)	Improvements (%)
Wikipedia	USSM	Context	48.3	
		Context + ConcepNet	55.0	+13.9
		Context + WordNet	53.3	+10.4
		Context + CauseCom	55.0	+13.9
		Context + All KBs	**56.7**	**+17.4**
	NKAM	Context	51.7	
		Context + ConcepNet	60.0	+16.0
		Context + WordNet	60.0	+16.0
		Context + CauseCom	61.7	+19.3
		Context + All KBs	**63.3**	**+22.4**
	USSM+NKAM	Context	53.3	
		Context + ConcepNet	63.3	+18.7
		Context + WordNet	61.7	+15.7
		Context + CauseCom	65.0	+21.9
		Context + All KBs	**66.7**	**+25.1**

Outline

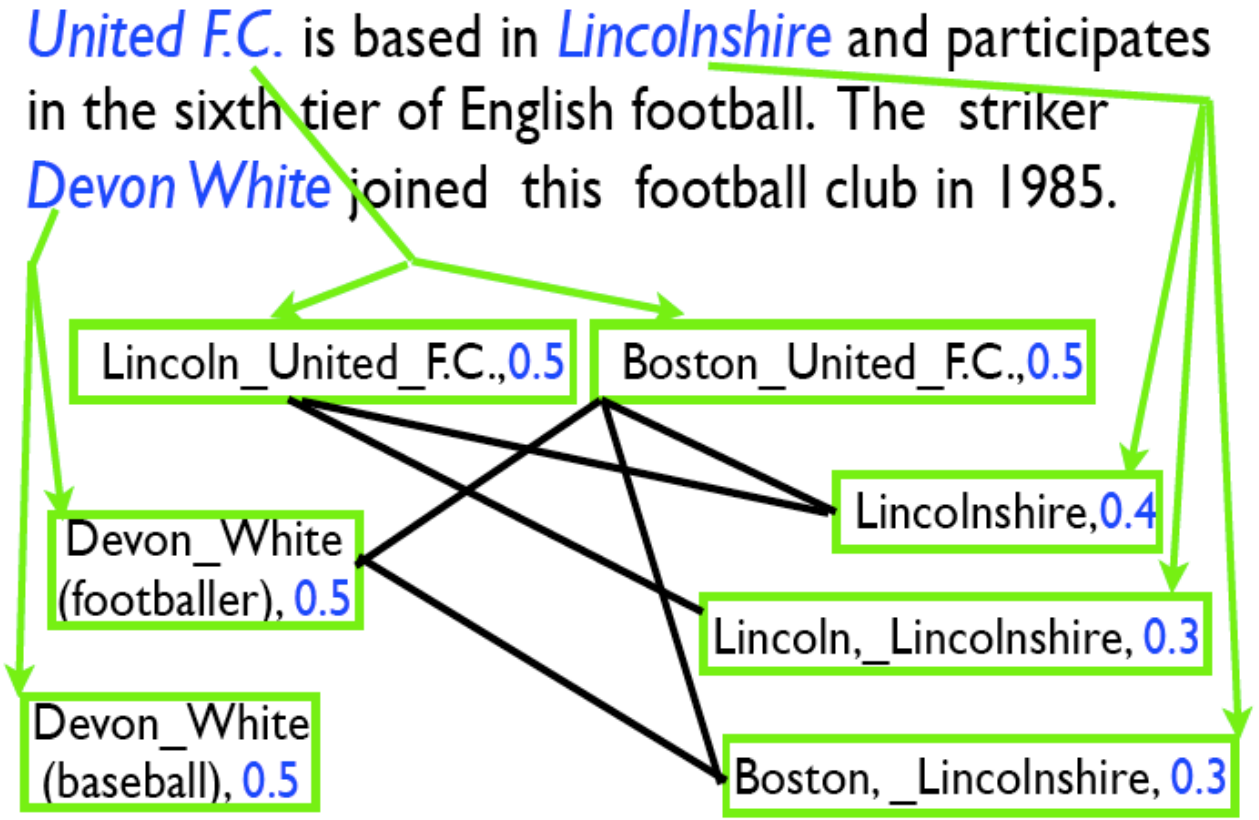
iFLYTEK Laboratory for Neural Computing and Machine Learning (iNCML)

2054 LASSONDE SCHOOL OF ENGINEERING | YORK UNIVERSITY

- Introduction
  - Deep Learning for NLP
- A New General Approach
  - FOFE: lossless encoding
  - DNNs: universal approximator
- NLP Applications
  - Word Embedding
  - Language Modelling
  - Named Entity Recognition and Mention Detection in KBP-EDL
  - Coreference Resolution in Winograd
- **On-going work**
  - **Entity Linking**
  - **Syntactic/Semantic Parsing**
- Conclusions

FOFE-net for Entity Linking

- Each named entity matches multiple items in knowledge bases
  - Name-Item matching score (matched char; FOFE-net)
  - Coherence between entities (max graph edges, personalized PageRank)



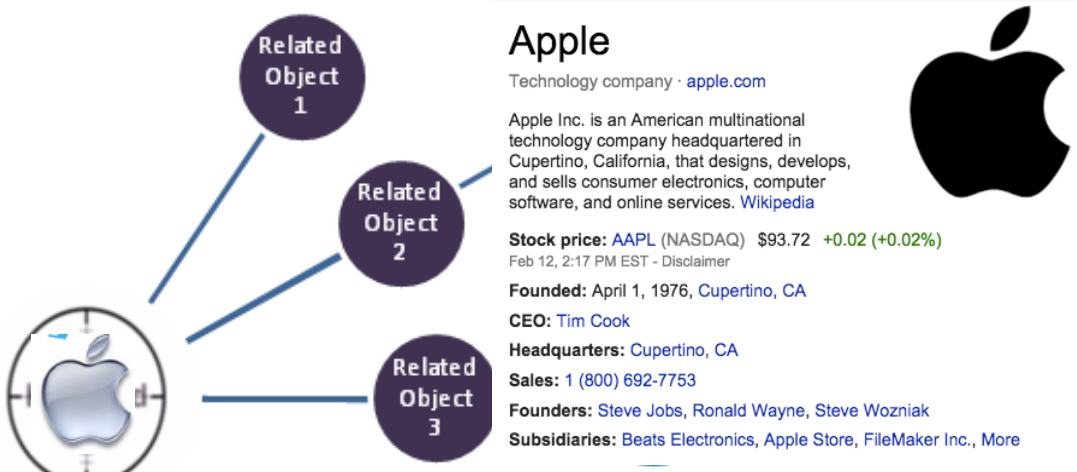
FOFE-net for Entity Linking

*The new **Apple** TV was unveiled by **Apple** at the iPhone launch event on 9 September, at which point **Apple** said it would ship in October.*

left FOFE L_1 | right FOFE R_1

left FOFE L_2 | right FOFE R_2

left FOFE L_3 | right FOFE R_3



node feature vectors N_{134}

bag-of-words
sum-of-FOFEs
...

$[L_1\ R_1\ N_{134}]$
$[L_2\ R_2\ N_{134}]$
$[L_3\ R_3\ N_{134}]$

→ **Deep Neural Network** → ***sigmoid matching score***

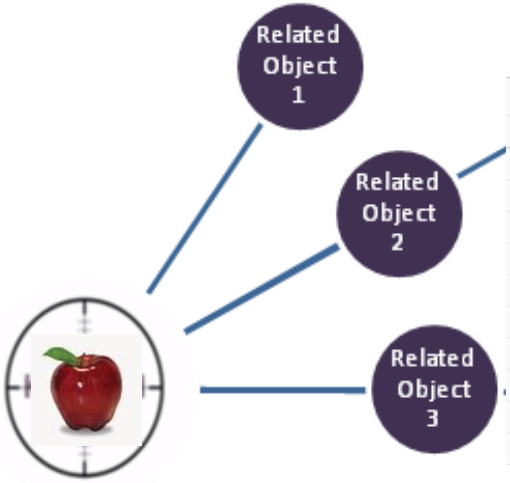
FOFE-net for Entity Linking

*The new **Apple** TV was unveiled by **Apple** at the iPhone launch event on 9 September, at which point **Apple** said it would ship in October.*

left FOFE L_1 | right FOFE R_1

left FOFE L_2 | right FOFE R_2

left FOFE L_3 | right FOFE R_3



Apple

Fruit

The apple tree is a deciduous tree in the rose family best known for its sweet, pomaceous fruit, the apple. It is cultivated worldwide as a fruit tree, and is the most widely grown species in the genus Malus. Wikipedia

Nutrition Facts

Apple

Amount Per 1 medium (3" dia) (182 g)	
Calories 95	
	% Daily Value*
Total Fat 0.3 g	0%
Saturated fat 0.1 g	0%
Polyunsaturated fat 0.1 g	
Monounsaturated fat 0 g	
Cholesterol 0 mg	0%
Sodium 2 mg	0%
Potassium 195 mg	5%
Total Carbohydrate 25 g	8%
Dietary fiber 4.4 g	17%

node feature vector N_{389}

Bag-of-words
Sum-of-FOFEs
...

[L_1 R_1 N_{389}]
[L_2 R_2 N_{389}]
[L_3 R_3 N_{389}]

→ **Deep Neural Network** → ***sigmoid matching score***

FOFE-net for Relation Extraction (Slot Filling)

[Dana Hull]$_{PER}$ reported that [Solyndra]$_{ORG}$ CEO [Brian Harrisis]$_{PER}$ a registered Republican

[Dana Hull]$_{PER}$ reported that *[Solyndra]$_{ORG}$* *CEO [Brian Harrisis]$_{PER}$ a registered Republican*

left FOFE code L_1

right FOFE code R_1

[Dana Hull]$_{PER}$ reported that [Solyndra]$_{ORG}$ CEO *[Brian Harrisis]$_{PER}$* *a registered Republican*

left FOFE code L_2

right FOFE code R_2

[L_1 R_1]
[L_2 R_2]

→

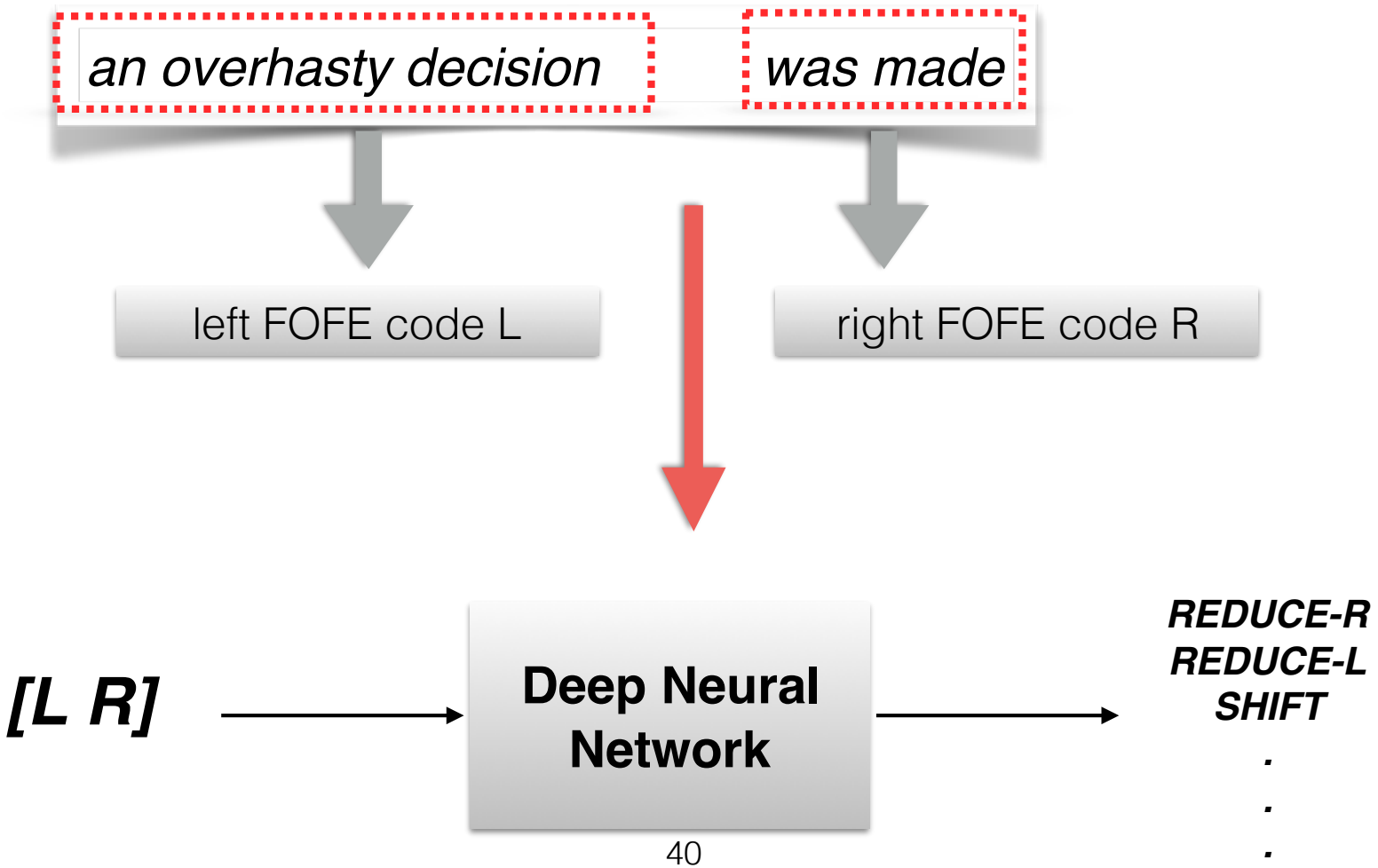
Deep Neural Network (ORG)

→

ORG:alternate_names
ORG:top_member_employees
ORG:date_founded
ORG:city_of_headquarters
...
NONE

FOFE-net for Syntactic Parsing

- Fit well with the transition-based parsing algorithm.
- No feature engineering; Fast to train from a very large corpus



Conclusions

- Proposed a new deep learning method (FOFE-net) for NLP
- FOFE-net:
  - Lossless fixed-size encoding + universal approximate
  - FOFE + neural network
  - Elegant and theoretically guaranteed
  - No feature engineering; Fast and easy to train
- FOFE-net is flexible: applicable to a wide range of NLP tasks
  - Word embedding
  - Language modelling
  - Named entity recognition
  - Coreference resolution
  - more and more …
- Achieved promising results on all examined tasks