

# Journey of Keyphrase Extraction

Animesh Prasad  
PhD Candidate  
WING-NUS

DSO Data Science & AI Talk Series  
Nov 2, 2017

# What is Keyphrase Extraction?



Key Concept 1

Key Concept 2

.....



# Keyphrases everywhere!



Subscribe (Full Service) [Register](#) (Limited Service, Free) [Login](#)

Search: ☒ The ACM Digital Library ☐ The Guide

**THE ACM DIGITAL LIBRARY**

**Human evaluation of Kea, an automatic keyphrasing system**

**Full text**  (381 KB)

**Source** [International Conference on Digital Libraries](#) [archive](#)  
**Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries** [table of contents](#)  
 Roanoke, Virginia, United States  
 Pages: 148 - 156  
 Year of Publication: 2001  
 ISBN:1-58113-345-6

**Authors** [Steve Jones](#) Department of Computer Science, University of Waikato, Private Bag 3105, Hamilton, New Zealand  
[Gordon W. Paynter](#) Department of Computer Science, University of Waikato, Private Bag 3105, Hamilton, New Zealand

**Sponsor** [ACM](#): Association for Computing Machinery

**Publisher** ACM New York, NY, USA

**Keywords:**  
[author keyphrases](#), [digital libraries](#), [keyphrase extraction](#), [subjective evaluation](#), [user interface](#)

**Collaborative Colleagues:**  
 Gordon W. Paynter: David Bainbridge  
 Stefan J. Boddie  
 George Buchanan  
 Sally Jo Cunningham

**SEARCH INSIDE!™**

**Understanding Digital Libr (Paperback)**  
 by [Michael Lesk](#) (Author) "This book is a  
**Key Phrases:** [perfect hashing](#), [digital libr](#)  
 ★★★★★ (1 customer review)  
 List Price: \$51.95  
 Price: **\$51.95** & this item shi

**Inside This Book**  
**Citations:** This book cites 60 books  
**Explore:** Citations | Books on Related Topics |  
 Concordance | Text Stats

**IMDb** NOW PLAYING MOVIE/TV NEWS MY MOVIES

Internet Movie Database Home | Top Movies | Pl  
 search All

IMDb > [The Matrix \(1999\)](#) > Plot keywords

**Plot keywords for The Matrix (1999)**

★ [photos](#) [board](#)

- [Spoon](#)
- [Future](#)
- [Fall From Height](#)
- [Cat](#)
- [Altered Version Of Studio Logo](#)
- [555 Phone Number](#)
- [Exploding Body](#)
- [Parallel World](#)
- [Kung Fu](#)
- [Mentor](#)
- [No Opening Credits](#)
- [Blown To Pieces](#)
- [Shot Through Wall](#)
- [Prophecy](#)
- [Maya](#)
- [Computer](#)

**Quicklinks**  
 plot keywords

**Top Links**  
[trailers and videos](#)  
[full cast and crew](#)  
[trivia](#)  
[official sites](#)  
[memorable quotes](#)

*To think about: Are tags keyphrases?*

# Why do we care?

- information retrieval (IR) tasks,
  - such as text summarization,
  - text categorization,
  - opinion mining and
  - document indexing



# Corpora

Source	Dataset/Contributor	Statistics		
		Documents	Tokens/doc	Keys/doc
Paper abstracts	<i>Inspec</i> [20] *	2,000	< 200	10
Scientific papers	NUS corpus [42] *	211	≈ 8K	11
	citeulike.org [37] *	180	-	5
	SemEval-2010 [27] *	284	> 5K	15
Technical reports	NZDL [56] *	1,800	-	-
News articles	DUC-2001 [53] *	308	≈ 900	8
	<i>Reuters</i> corpus [19]	12,848	-	6
Web pages	Yih et al. (2006)	828	-	-
	Hammouda et al. (2005) *	312	≈ 500	-
	Blogs [13]	252	≈ 1K	8
Meeting transcripts	ICSI [30]	161	≈ 1.6K	4
Emails	Enron corpus [9] *	14,659	-	-
Live chats	Library of Congress [25]	15	-	10

# Approaches

KEA — 1999

- **Supervised**
  - **Binary Classification**
    - naïve Bayes,
    - decision trees,
    - bagging,
    - boosting,
    - maximum entropy,
    - multi-layer perceptron,
    - and support vector machines



# Approaches

KEA — 1999

- Supervised
  - Binary Classification
    - naïve Bayes,
    - decision trees,
    - bagging,
    - boosting,
    - maximum entropy,
    - multi-layer perceptron,
    - and support vector machines
  - **Problem? Classification is not a tournament!**

# Approaches

- Supervised
  - Binary Classification
  - **Problem? Classification is not a tournament!**
- Ranking

KEA — 1999

Jiang et al — 2009





# Apporaches

- **Supervised**
  - **Features**
    - **In document features**
      - **Statistical features**  
tf-idf, keyphraseness etc
      - **Structural features**  
document structure like section etc
      - **Syntactic features**  
POS tags etc

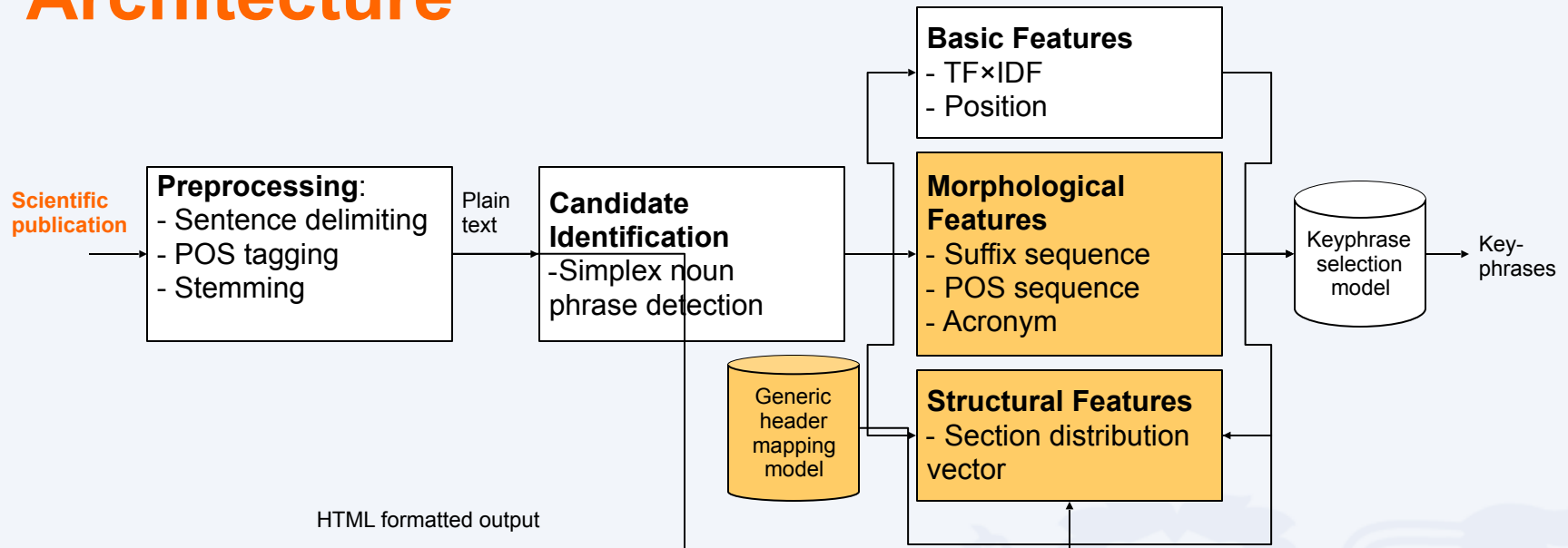


# Approaches

- Supervised
  - Features
    - In document features
    - Out of the document features
      - Wikipedia-based keyphraseness
      - Search Log based keyphraseness
      - *Web as a corpus for related terms*



# Architecture



# Approaches

- **Unsupervised**
  - **Graph-Based Ranking**
    - **TextRank (Page Rank)**
  - **Clustering with Wikipedia**
  - **Topical PageRank**
- **And many more complex graph based algorithm**

Text Rank — 2004

Liu et al — 2010

# Benchmarking

- **Unsupervised**
  - **Graph-Based Ranking**
    - **TextRank (Page Rank)**
  - **Clustering with Wikipedia**
  - **Topical PageRank**
- **Any many more complex graph based algorithm**

Text Rank — 2004

Liu et al — 2010

# Benchmarking

Dataset	Author	Reader	Combined
Trial	149	526	621
Training	559	1824	2223
Test	387	1217	1482

SemEval — 2010

# Benchmarking

System	Rank	Top 5 candidates			Top 10 candidates			Top 15 candidates		
		P	R	F	P	R	F	P	R	F
HUMB	1	39.0%	13.3%	19.8%	32.0%	21.8%	26.0%	27.2%	27.8%	27.5%
WINGNUS	2	40.2%	13.7%	20.5%	30.5%	20.8%	24.7%	24.9%	25.5%	25.2%
KP-Miner	3	36.0%	12.3%	18.3%	28.6%	19.5%	23.2%	24.9%	25.5%	25.2%
SZTERGAK	4	34.2%	11.7%	17.4%	28.5%	19.4%	23.1%	24.8%	25.4%	25.1%
ICL	5	34.4%	11.7%	17.5%	29.2%	19.9%	23.7%	24.6%	25.2%	24.9%
SEERLAB	6	39.0%	13.3%	19.8%	29.7%	20.3%	24.1%	24.1%	24.6%	24.3%
KX_FBK	7	34.2%	11.7%	17.4%	27.0%	18.4%	21.9%	23.6%	24.2%	23.9%
DERIUNLP	8	27.4%	9.4%	13.9%	23.0%	15.7%	18.7%	22.0%	22.5%	22.3%
Maui	9	35.0%	11.9%	17.8%	25.2%	17.2%	20.4%	20.3%	20.8%	20.6%
DFKI	10	29.2%	10.0%	14.9%	23.3%	15.9%	18.9%	20.3%	20.7%	20.5%
BUAP	11	13.6%	4.6%	6.9%	17.6%	12.0%	14.3%	19.0%	19.4%	19.2%
SJTULTLAB	12	30.2%	10.3%	15.4%	22.7%	15.5%	18.4%	18.4%	18.8%	18.6%
UNICE	13	27.4%	9.4%	13.9%	22.4%	15.3%	18.2%	18.3%	18.8%	18.5%
UNPMC	14	18.0%	6.1%	9.2%	19.0%	13.0%	15.4%	18.1%	18.6%	18.3%
JU_CSE	15	28.4%	9.7%	14.5%	21.5%	14.7%	17.4%	17.8%	18.2%	18.0%
LIKEY	16	29.2%	10.0%	14.9%	21.1%	14.4%	17.1%	16.3%	16.7%	16.5%
UvT	17	24.8%	8.5%	12.6%	18.6%	12.7%	15.1%	14.6%	14.9%	14.8%
POLYU	18	15.6%	5.3%	7.9%	14.6%	10.0%	11.8%	13.9%	14.2%	14.0%
UKP	19	9.4%	3.2%	4.8%	5.9%	4.0%	4.8%	5.3%	5.4%	5.3%

# Benchmarking

System	Rank	Top 5 candidates			Top 10 candidates			Top 15 candidates		
		P	R	F	P	R	F	P	R	F
HUMB	1	39.0%	13.3%	19.8%	32.0%	21.8%	26.0%	27.2%	27.8%	27.5%
WINGNUS	2	40.2%	13.7%	20.5%	33.0%	25.2%	24.7%	24.9%	25.5%	25.2%
KP-Miner	3	36.0%	12.3%	18.3%	28.6%	19.5%	23.2%	24.9%	25.5%	25.2%
SZTERGAR	4	34.2%	11.7%	17.4%	28.5%	19.4%	23.1%	24.8%	25.4%	25.1%
ICL	5	34.4%	11.7%	17.5%	29.2%	19.9%	23.7%	24.6%	25.2%	24.9%
SEERLAB	6	39.0%	13.3%	19.8%	29.7%	20.3%	24.1%	24.1%	24.6%	24.3%
KX-FBK	7	34.2%	11.7%	17.4%	27.0%	18.4%	21.9%	23.6%	24.2%	23.9%
DERIUNLP	8	27.4%	9.4%	13.9%	23.0%	15.7%	18.7%	22.0%	22.5%	22.3%
Maui	9	35.0%	11.9%	17.8%	25.2%	17.2%	20.4%	20.3%	20.8%	20.6%
DFKI	10	29.2%	10.0%	14.9%	23.3%	15.9%	18.9%	20.3%	20.7%	20.5%
BUAP	11	13.6%	4.6%	6.9%	17.6%	12.0%	14.3%	19.0%	19.4%	19.2%
SJTULTLAB	12	30.2%	10.3%	15.4%	22.7%	15.5%	18.4%	18.4%	18.8%	18.6%
UNICE	13	27.4%	9.4%	13.9%	22.4%	15.3%	18.2%	18.3%	18.8%	18.5%
UNPMC	14	18.0%	6.1%	9.2%	19.0%	13.0%	15.4%	18.1%	18.6%	18.3%
JU_CSE	15	28.4%	9.7%	14.5%	21.5%	14.7%	17.4%	17.8%	18.2%	18.0%
LIKEY	16	29.2%	10.0%	14.9%	21.1%	14.4%	17.1%	16.3%	16.7%	16.5%
UvT	17	24.8%	8.5%	12.6%	18.6%	12.7%	15.1%	14.6%	14.9%	14.8%
POLYU	18	15.6%	5.3%	7.9%	14.6%	10.0%	11.8%	13.9%	14.2%	14.0%
UKP	19	9.4%	3.2%	4.8%	5.9%	4.0%	4.8%	5.3%	5.4%	5.3%

Unsupervised



## Shift in Focus

- Tens of work on incorporating domain specific, semantically rich feature for extraction algorithm
- Meanwhile DARPA's MUC7 saw approx. 88% results and leading to subsequent research in KB using entities
- This lead to shift in focus for keyphrases from indexing component to an upstream task for KBC

SemEval — 2017

# Shift in Focus

- Named Entity Recognition



- SemEval 2017 Keyphrase Extraction

**Task**  
Information extraction is the process of extracting structured data from unstructured text, which is relevant for several end-to-end tasks, including **Task** question answering. This paper addresses the tasks of **Task** named entity recognition (NER), a subtask of **Task** information extraction, using **Process** conditional random fields (CRF). Our method is evaluated on the **Material** ConLL-2003 NER corpus.

**Task** same-as **Task** is-a **Task**

**Process** same-as **Process**

SemEval — 2017

# Benchmarking SemEval 2017 Task 10: Science IE

- **Subtask (A): Identification of keyphrases**
- Given a scientific publication, the goal of this task is to identify all the keyphrases in the document.
  
- **Subtask (B): Classification of identified keyphrases**
- In this task, each keyphrase needs to be labelled by one of three types: (i) PROCESS, (ii) TASK, and (iii) MATERIAL.
- **PROCESS:** Keyphrases relating to some scientific model, algorithm or process should be labelled by PROCESS.
- **TASK:** Keyphrases those denote the application, end goal, problem, task should be labelled by TASK.
- **MATERIAL:** MATERIAL keyphrases identify the resources used in the paper.
  
- **Subtask (C): Extraction of relationships between two identified keyphrases**
- Every pair of keyphrases need to be labelled by one of three types: (i) HYPONYM-OF, (ii) SYNONYM-OF, and (iii) NONE.
- **HYPONYM-OF:** The relationship between two keyphrases A and B is HYPONYM-OF if semantic field of A is included within that of B. One example is Red HYPONYM-OF Color.
- **SYNONYM-OF:** The relationship between two keyphrases A and B is SYNONYM-OF if they both denote the same semantic field, for example Machine Learning SYNONYM-OF ML.

## Benchmarking SemEval 2017 Task 10: Science IE

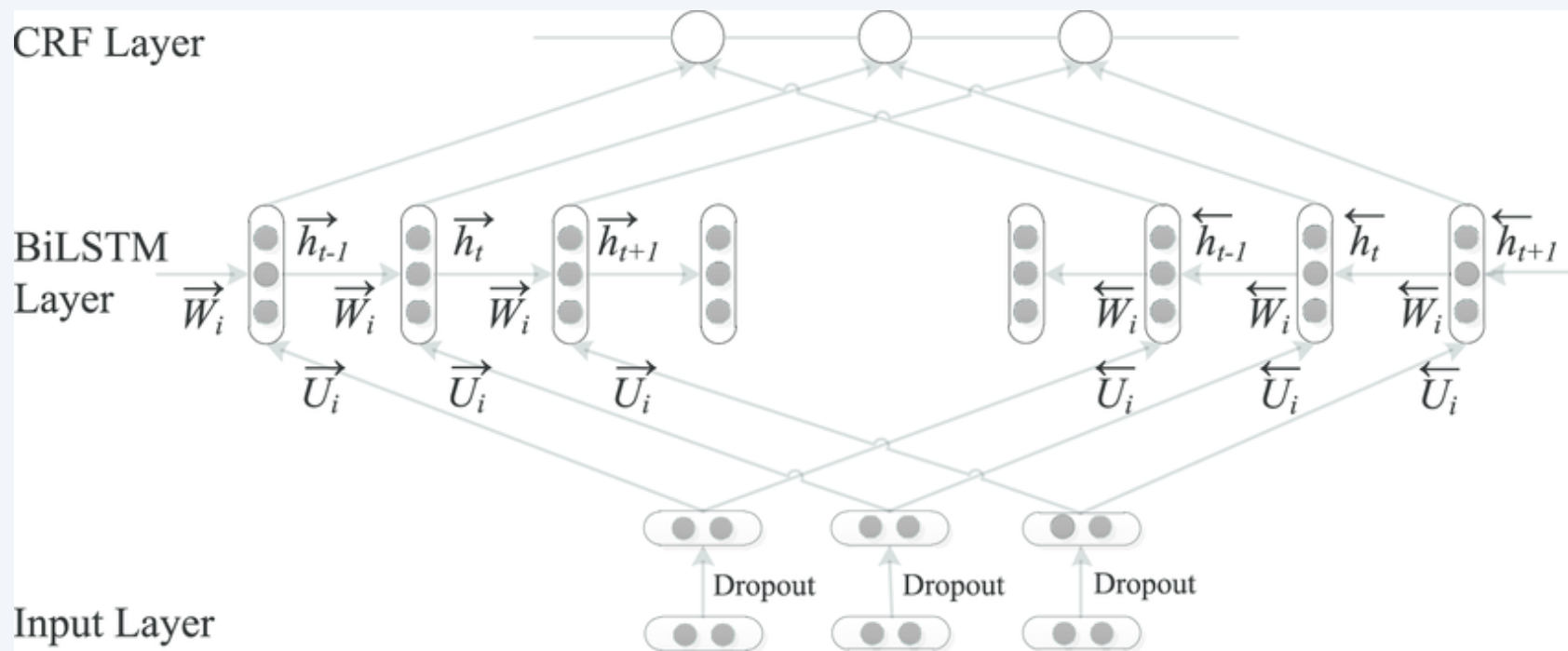
Teams	Overall	A	B	C
s2_end2end (Ammar et al., 2017)	<b>0.43</b>	0.55	<b>0.44</b>	<b>0.28</b>
TIAL_UW	0.42	<b>0.56</b>	<b>0.44</b>	
TTL_COIN (Tsujimura et al., 2017)	0.38	0.5	0.39	0.21
PKU_ICL (Wang and Li, 2017)	0.37	0.51	0.38	0.19
NTNU-1 (Marsi et al., 2017)	0.33	0.47	0.34	0.2
WING-NUS (Prasad and Kan, 2017)	0.27	0.46	0.33	0.04
Know-Center (Kern et al., 2017)	0.27	0.39	0.28	
SZTE-NLP (Berend, 2017)	0.26	0.35	0.28	
NTNU (Lee et al., 2017b)	0.23	0.3	0.24	0.08
LABDA (Segura-Bedmar et al., 2017)	0.23	0.33	0.23	
LIPN (Hernandez et al., 2017)	0.21	0.38	0.21	0.05
SciX	0.2	0.42	0.21	
IHS-RD-BELARUS	0.19	0.41	0.19	
HCC-NLP	0.16	0.24	0.16	
NITK_IT_PG	0.14	0.3	0.15	
Surukam	0.1	0.24	0.1	0.13
GMBUAP (Flores et al., 2017)	0.04	0.08	0.04	
<i>upper bound</i>	0.84	0.85	0.85	0.77
<i>random</i>	0.00	0.03	0.01	0.00

## BiLSTM CRF Benchmark

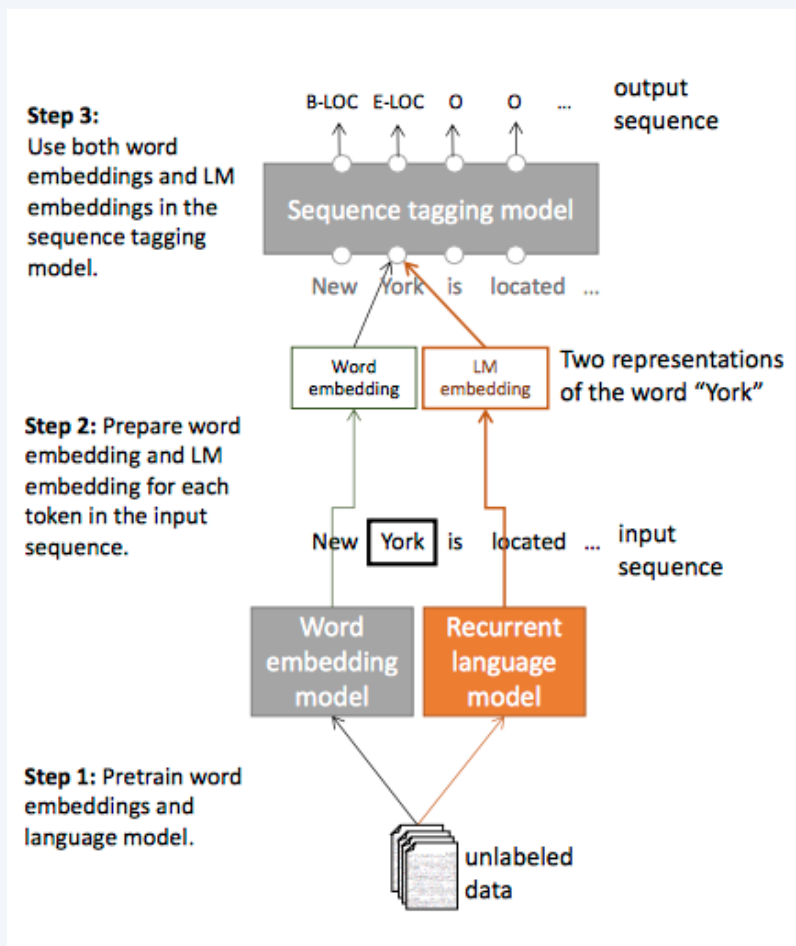
- It beats traditional models on Tagging, Chunking, Semantic Role Labelling and NER

System	accuracy
Combination of HMM, Maxent etc. (Florian et al., 2003)	88.76
MaxEnt classifier (Chieu., 2003)	88.31
Semi-supervised model combination (Ando and Zhang., 2005)	89.31
Conv-CRF (Collobert et al., 2011)	81.47
Conv-CRF (Senna + Gazetteer) (Collobert et al., 2011)	89.59
CRF with Lexicon Infused Embeddings (Passos et al., 2014)	<b>90.90</b>
BI-LSTM-CRF	84.26
BI-LSTM-CRF (Senna + Gazetteer)	90.10

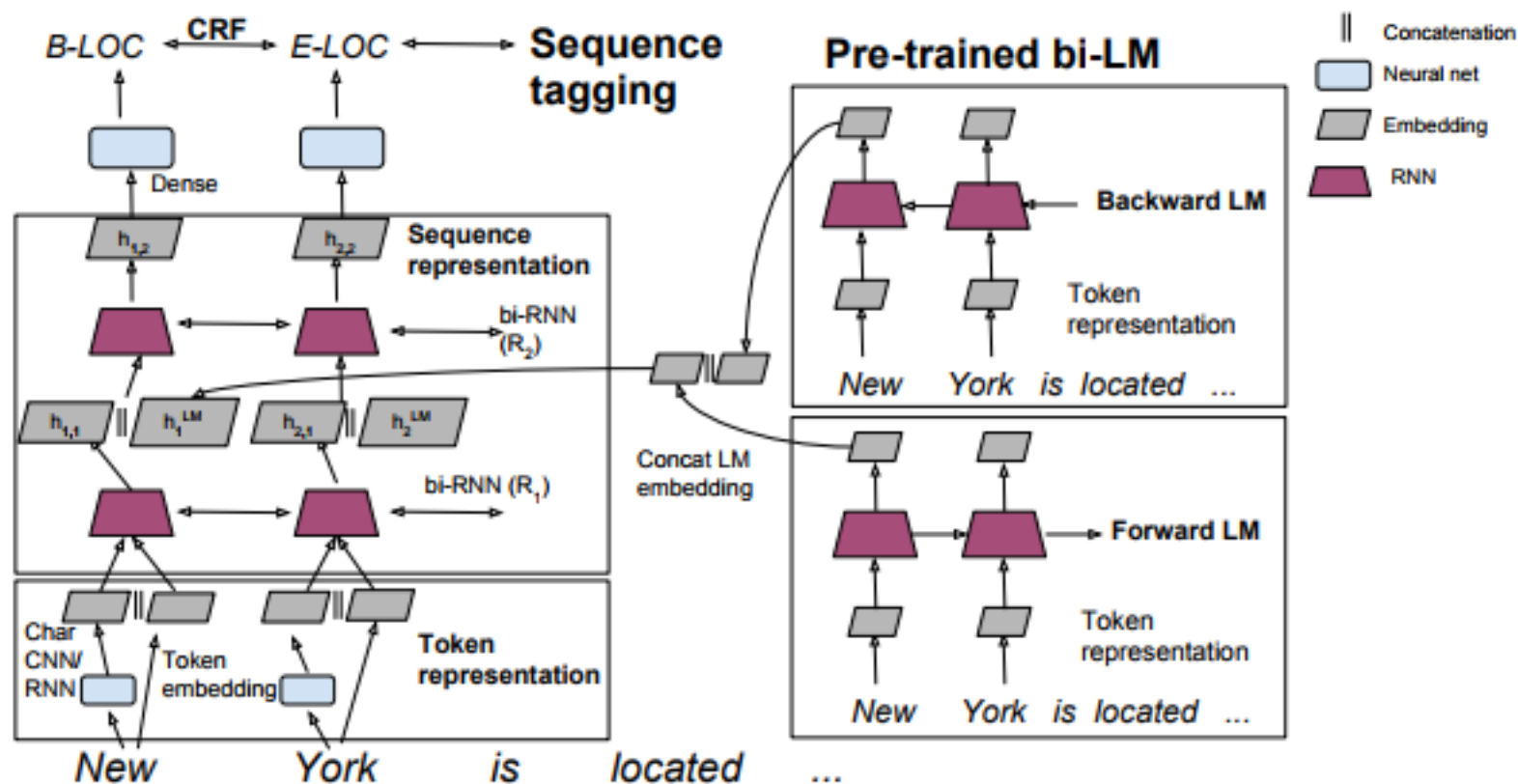
# BiLSTM CRF



# BiLSTM CRF with a touch of LM



# BiLSTM CRF with a touch of LM





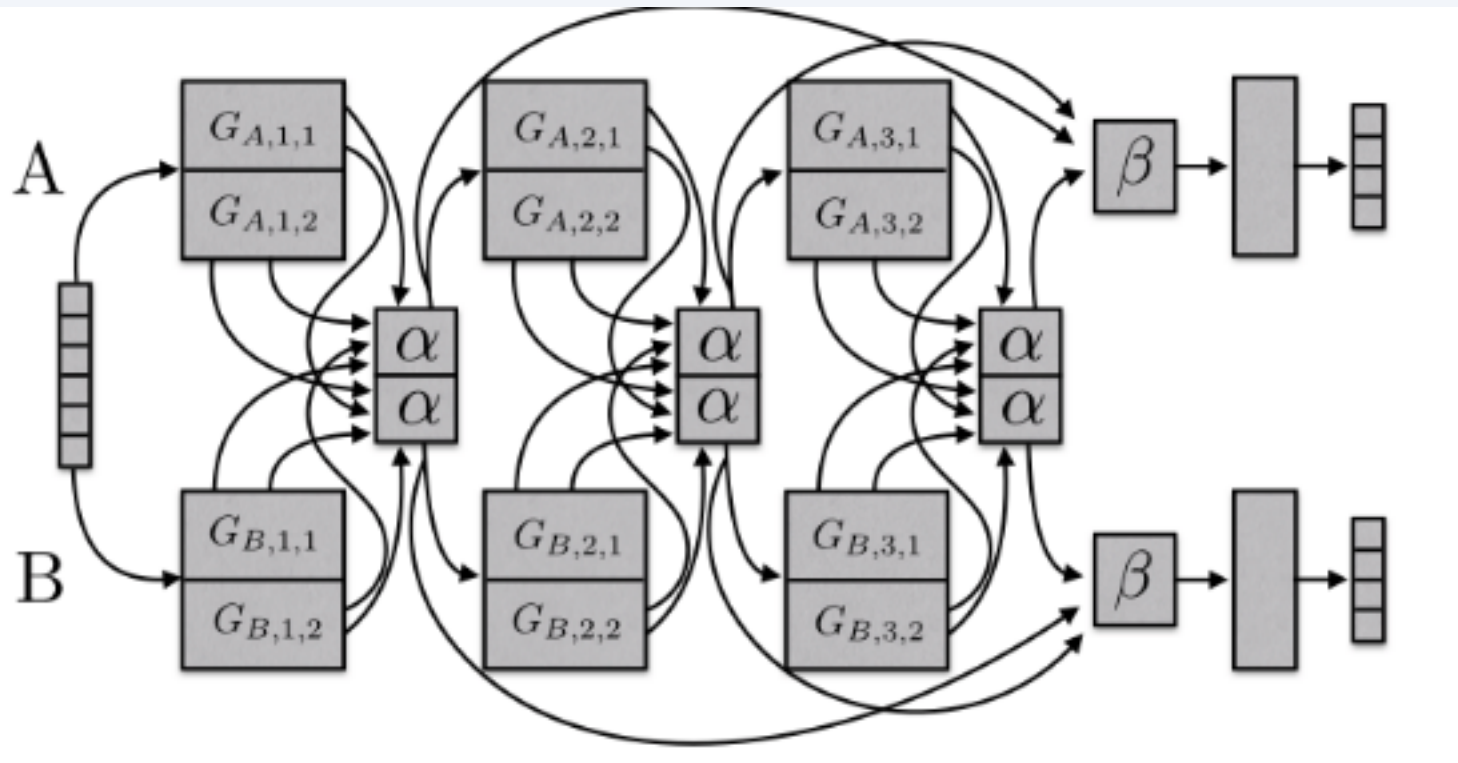
# BiLSTM CRF with a touch of LM

Model	External resources	$F_1$ Without	$F_1$ With	$\Delta$
Yang et al. (2017)	transfer from CoNLL 2000/PTB-POS	91.2	91.26	+0.06
Chiu and Nichols (2016)	with gazetteers	90.91	91.62	+0.71
Collobert et al. (2011)	with gazetteers	88.67	89.59	+0.92
Luo et al. (2015)	joint with entity linking	89.9	91.2	+1.3
	no LM vs TagLM <i>unlabeled data only</i>	90.87	<b>91.93</b>	+1.06

# Multitasking

Method	Unlabelled			Labelled		
	Precision	Recall	F1	Precision	Recall	F1
Finkel et al. (2005)	77.89	50.27	61.10	49.90	27.97	35.85
Lample et al. (2016)	71.92	49.37	58.55	41.36	28.47	33.72
BiLSTM	81.58	57.86	67.71	45.80	32.48	38.01
BiLSTM + Chunking	82.88	52.08	63.96	55.54	34.90	42.86
BiLSTM + Framenet	77.86	56.05	65.18	54.04	38.91	45.24
BiLSTM + Hyperlinks	76.59	60.53	67.62	46.99	44.09	41.13
BiLSTM + Multi-word	74.80	70.18	<b>72.42</b>	46.99	44.09	<b>45.49</b>
BiLSTM + Super-sense	83.70	51.76	63.93	56.94	35.25	43.54

## And yet another Multitasking: Sluice Network



## And yet another Multitasking: Sluice Network

Named entity recognition								
System	nw (ID)	bc	bn	mz	pt	tc	wb	OOD Avg
Single task	95.04	93.42	93.81	93.25	94.29	94.27	92.52	93.59
Hard parameter sharing	94.16	91.36	93.18	93.37	<b>95.17</b>	93.23	<b>92.99</b>	93.22
Low supervision	94.94	91.97	93.69	92.83	94.26	93.51	92.51	93.13
Cross-stitch network	95.09	92.39	93.79	93.05	94.14	93.60	92.59	93.26
Sluice network	<b>95.52</b>	<b>93.50</b>	<b>94.16</b>	<b>93.49</b>	93.61	<b>94.33</b>	92.48	<b>93.60</b>

# Challenges

- Long Documents vs Short Excerpts
  - Overgeneration
  - Infrequency
  - Redundancy
  - Evaluation



**Thank You**  
[animesh@comp.nus.edu.sg](mailto:animesh@comp.nus.edu.sg)