Lecture 5: Automatic Chinese Word Segmentation

Xu Ruifeng

Harbin Institute of Technology, Shenzhen

Today’s class – Automatic Chinese Word Segmentation

- Chinese Word Segmentation
- Problem Statement
  - Segmentation Ambiguity
  - Unknown Word Identification
- Automatic Word Segmentation Algorithm
- Benchmarks
- Open Resources

Automatic Chinese Word Segmentation

- Definition
  - A sentence *S* is a sequence of Chinese characters $(C_1, C_2, \ldots, C_M)$. An n-character token, namely word, of *S* beginning at the *i*th character is an ordered n-tuple of the form $(C_i, C_{i+1}, \ldots, C_{i+n-1})$

    A partition *W* of a sentence partitions all its characters into non-overlapping tokens of various length $(C_1, C_2, \ldots, C_{i_1\text{-}1})$, $(C_{i_1}, C_{i_1+1}, \ldots, C_{i_2\text{-}1})$, $(C_{i_{N-1}}, C_{i_{N-1}+1}, \ldots, C_M)$.

    Name the elements of W as respectively.

    $(W_1, W_2, \ldots, W_N)$,

Why We Need Word Segmentation

- Accurate word segmentation is the foundation of Chinese language processing
  - Information Retrial
    - 和服 | 务 | 于三日后裁制完毕，并呈送将军府中
    - 王府饭店的设施 | 和 | 服务 | 是一流的。
  - Text to Speech
    - 他们是来 | 查 | 金泰 | 撞人那件事的。“查” cha
    - 行侠仗义的 | 查金泰 | 远近闻名。 “查” zha
  - Machine Translation
    - 我看见周星驰同张学友打招呼
    - Transtar: *I see week star Chi open together study friend greet.*

Segmentation Difficulties

这样|的|人才|能|经受|住|考验

vs. 这样|的|人|才能|经受|住|考验

白|天鹅|在|水|中|游来游去

vs. 白天|鹅|在|水|中|游来游去

我|将来|要|上|大学

vs. 我|将|来|上 海

两|个|人|一起|去

vs. 这|是|个人|问题

Segmentation Difficulties

今天|学生|会面讨论|这个|问题

vs. 他|是|学生会|主席

这|篇|文章|太|平淡|了

vs. 难得一个|太平|盛世

莱温斯基|本来|就|很|不|爽

vs.莱温斯|基本|来|就|很|不|爽

夏尔·莫里斯·塔列朗|的|祖先|从|10世纪|卡佩王朝|建立|时|起|就|已经|是|宫廷|贵人|了。他|的|父亲|塔列朗伯爵|查理-达尼埃尔|同|国王|路易十六|还|是|表兄弟

Natural language generation/Retrieval

- 白痴
  - 小白痴痴地等着她回来
- 如果
  - 汽水不如果汁好喝
- 本来
  - 那书本来打头会很痛

Two Main Challenges

- Two main Challenges
  - Ambiguity
  - Unknown word (Out of Vocabulary words, OOV)
- Types of ambiguity
  - Overlapping ambiguity (交叉型歧义)

    “人才|能” and “人|才能”
  - Combinational ambiguity (组合型歧义)

    “个人” and “个|人”
  - Mixed ambiguity (混合型歧义)

    “太平|淡” and “太|平淡”

Overlap of common characters : Examples

- one: 和尚末
- two: 结合成分
- three: 为人民工作
- four: 中国产品质量
- six: 努力学习语法规则
- seven: 治理解放大道路面积水

Overlap of Common Characters: Status

链长	1	2	3	4	5	6	7	8	总计
词频	47402	28790	1217	608	29	19	2	1	78248
%	50.58	47.02	1.56	0.78	0.04	0.02	0	0	100
字段数	12686	10131	743	324	22	5	2	1	23914
%	53.05	42.36	3.11	1.35	0.09	0.02	0.01	0.01	100

- Table 1. Statistics on 5M-word news corpus

-- 刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，第65页

National Standard

- 信息处理用现代汉语分词规范
  - GBT 13715-1992
- 刘源,谭强,沈旭昆《信息处理用现代汉语分词规范及自动分词方法 》
- 《资讯处理用中文分词规范》台湾中研院
- 《人民日报》语料库词语切分规范

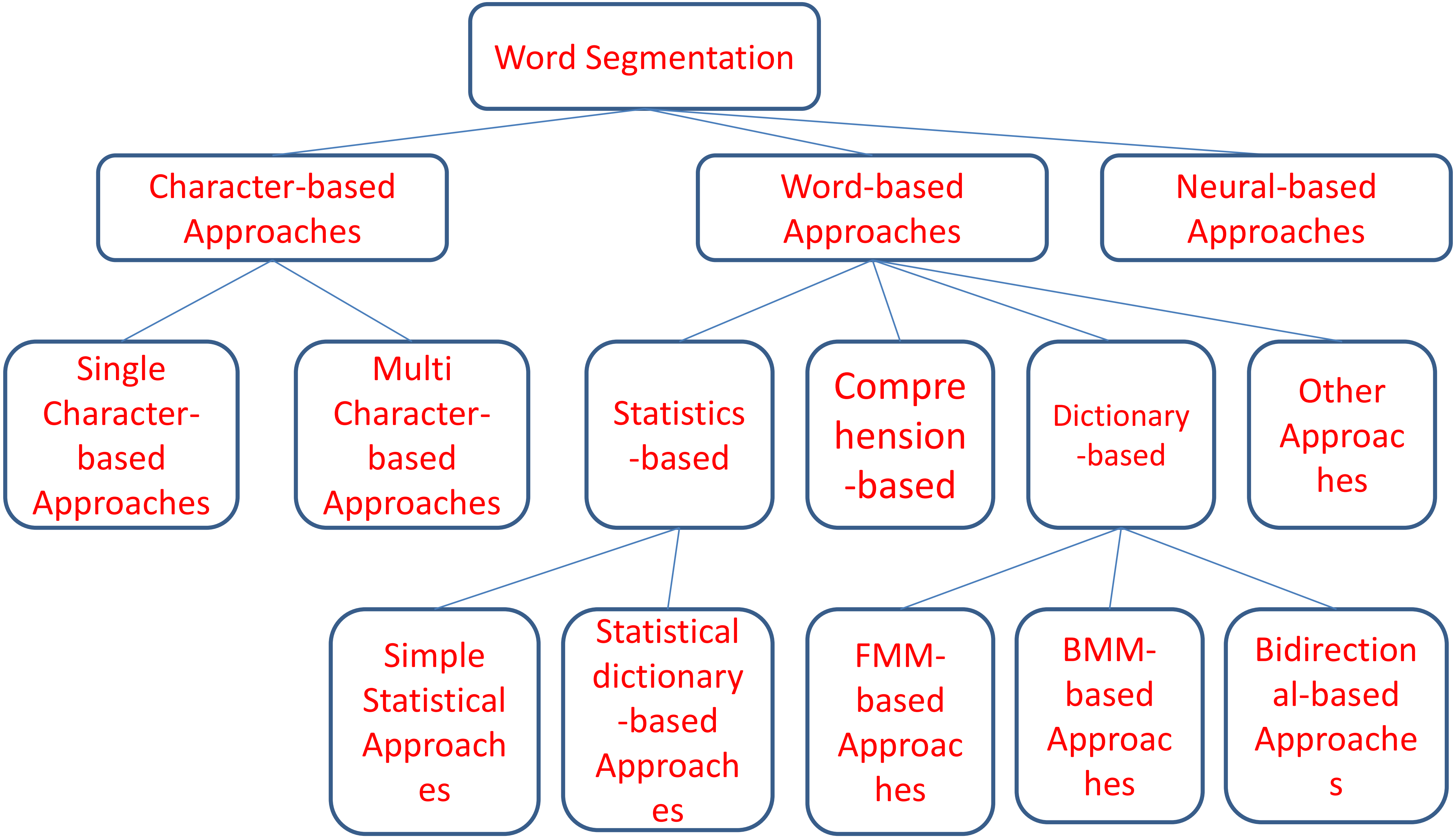
Out-of-vocabulary (OOV) problem

- Words unknown to dictionary “莱温斯基” “铁西”
- They are typically Named Entities 命名实体
  - Person name, e.g. “温家宝””查理-达尼埃尔”
  - Location name , e.g. “马甸”
  - Organization name , e.g. “微软”
  - Informal word or new word in Internet, e.g. “欺实马” “亲们” ” 酱油帝”
- The unknown words caused the word segmentation accuracy loss at least five times more than word ambiguity [Huang and Zhao 2007]

Tasks of Typical Word Segmentation Module

- Dictionary lookup to get words
- Recognition of 重叠词、离合词 and 词缀（SC*）
  - 重叠词：高兴 ➔ 高高兴兴，高兴高兴
  - 离合词：担心➔担什么心，洗澡➔洗了个热水澡
  - 词缀：标准➔超标准，科学➔科学家
- Disambiguation in word segmentation
- OOV word detection
  - Named entity recognition (命名实体识别)

Automatic Word Segmentation Approach



Word-based Approach: Dictionary-based

- Forward Maximum Matching method
  - Greedy search routine
  - Find the longest string starting from the very point in the sentence that matches a word entry
- Backward Maximum Matching method
  - Greedy search routine
  - Similar to FMM, but backward
  - More accurate

Forward Maximum Matching method

Algorithm:

1. I ← the number of Chinese characters in the longest word in the dictionary.
2. *segment*$_{match}$ ←one segment with length of I from the corpus.
3. If *segment*$_{match}$ in the dictionary:

   then match success, make segment$_{match}$ as a word, goto 6.
4. Else match failed, remove the last Chinese character in *segment*$_{match}$,
5. Repeat 3-4 until match success.
6. Goto 1 until all the words in the sentence are segmented.

Forward Maximum Matching method

Analyze :

- “市场/中国/有/企业/才能/发展/”
- Not a good solution to the overlapping ambiguity and combinational ambiguity.
- segmentation error: 1/169.
- It’s often used with other methods.

Backward Maximum Matching method

- BMM is similar to FMM, it just starts from the end of the corpus, and removes the first Chinese character not the last one when match fails.
- “市场/中/国有/企业/才能/发展/”
- segmentation error: 1/245
- can recognize the overlapping ambiguity

Word-based Approach: Dictionary-based

- Optimized Maximum Matching
  - Optimize the entries in the dictionary according to their frequency
  - Speed matching
  - No performance improvement
- Bi-directional Maximum Matching approach
  - applies FMM and then BMM
  - compares the two segmentation results to resolve any inconsistency and ambiguity
- Dictionary-based approach is difficult to process combinational ambiguities

Word-based Approach: Statistical-based

- Simple statistical approaches
  - [Sproat et al 2000] estimate the mutual information of two adjacent characters to determine whether they form a two-character word
  - [Sun et al 1998] further consider mutual information and the difference of *t*-score between characters

Sun HL. Distributional properties of Chinese collocations in text.

Word-based Approach: Statistical-based

- [Ge et al 1999]: a probabilistic model based on the Expectation Maximization (EM) algorithm
  - H1:There are a finite number of words of length *1* to *k*
  - H2:Each word has an unknown probability of occurrence
  - H3:Words are independent of each other
- Method: Use EM algorithm to multi pass estimation
  - Word are the candidate multi-grams from training corpus
  - Word probabilities are randomly assigned initially
  - They are used to segment the text
  - The word probabilities are re-estimated based on segmented results
  - The text are re-segmented using re-estimated probabilities
  - Iterates until convergence

Ge X, Pratt W., Smyth P. Discovering Chinese words from unsegmented text.

Statistical dictionary-based Approach

- Combine statistical and dictionary-based approaches
  - [Peng et al 2001]
  - a variant of the EM algorithm for Chinese segmentation
  - keeps a core lexicon which contains real words and a candidate lexicon that contains all other multi-grams not in the core lexicon
  - EM algorithm is used to maximize the likelihood of the training Corpus given the two lexicons and suggest new words as candidates for the core lexicon
  - Once a new word is added to core lexicon, the EM algorithm is reinitialized by giving half of the total probability mass to the core lexicon
  - Iterates until convergence

Peng F, Schuurmans D. Self-supervised Chinese word segmentation.

Comprehension-based Approach

- Taking into account the syntactic structure of sentences vs. ignore previous sentence
  - [Chen et al 1997]
    - used a segmented training Corpus to learn a set of rules to discriminate monosyllabic words from monosyllabic morphemes that may be parts of unknown words
    - monosyllabic words as instances of lexical units
    - examine the instances lexical units and non-lexical units as well as their contexts in the corpus
    - derive a set of context-dependent rules
    - The rules are sequentially applied to distinguish proper and improper characters

Chen C, Bai M, Chen K. Category guessing for Chinese unknown words.

– [Chen et al 2002] extend
    - used a set of context free morphological rules to model the structure of unknown words
    - a bottom-up merging algorithm that consults the morphological rules to extract unknown words

– [Wu et al 1988]
    - applied the technology of sentence understanding to word segmentation
    - segment words based on the syntactic parser

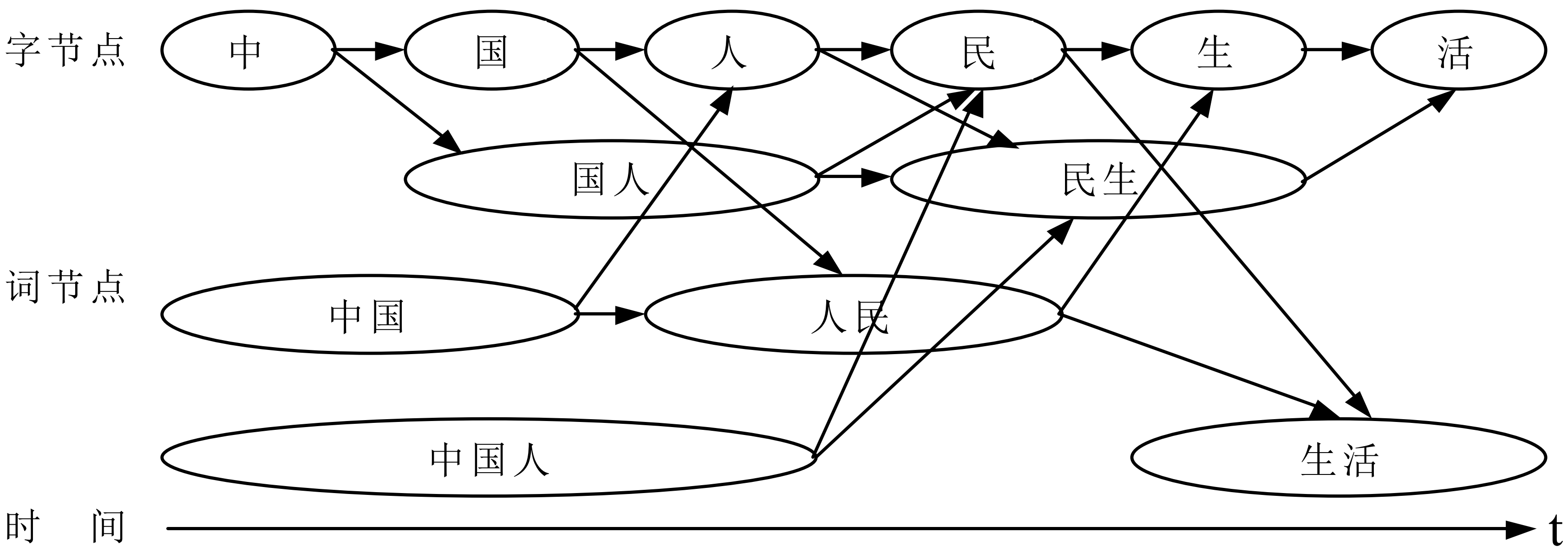
Chen K, Ma W. Unknown word extraction for Chinese documents.
Wu A, Jiang Z. Word segmentation in sentence analysis.

Machine learning-based Approaches

- Transformation-based Algorithm
  - transformation-based learning (TBL) algorithm [Brill 1995]
  - requires a pre-segmented reference Corpus and an initial segmenter
  - the learning algorithm compares the initial segmentation with the reference Corpus
  - identifies a rule to correct the segmentation errors
  - initial segmentation is updated to outputs a set of ranked rules

Brill E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging.

Word Lattice



– Apply different machine learning based algorithms to identify the most likely path in the word lattice

The word lattice based Approaches

- The shortest path
- The maximum probability path
- The maximum probability sum path
- The maximum probability sum with shorter path
- …

Hidden Markov Model based Method

– Needs the information of POS tagging
– deals with word segmentation and POS tagging at the same time
– The goal is to find the POS sequence *T* and word sequence *W* that maximize

$$
\begin{aligned}
W,T &= arg \max_{W,T,W(S)=S} P(T,W \mid S) \\
&= arg \max_{W,T,W(S)=S} P(W,T) \\
&= arg \max_{W,T,W(S)=S} P(W \mid T)P(T)
\end{aligned}
$$

$$P(w_i \mid t_i) = \frac{F(< w_i, t_i >)}{F(t_i)} \qquad P(t_i \mid t_{i-1}) = \frac{F(t_i, t_{i-1})}{F(t_{i-1})}$$

- Source Channel Model [Gao et al 2005]
  – Five word classes: namely, lexicon words, morphologically derived words, factoids, named entities, and new words
  – Each character sequence was segmented into a word class sequence using Source Channel Model

$$w^* = \underset{w \in GEN(s)}{argmax} P(w \mid s) = \underset{w \in GEN(s)}{argmax} P(w)P(s \mid w)$$

  – Generalized as linear mixture models to incorporate a very large number of linguistic and statistical features

$$Score(w, s, \lambda) = \sum_{d=0}^{D} \lambda_d f_d(w, s) \qquad w^* = \underset{w \in GEN(s)}{argmax}\, Score(\mathrm{w}, \mathrm{s}, \lambda)$$

Gao J, Li M，Wu A, Huang CN. Chinese word segmentation and named entity recognition: A pragmatic approach.

Word-based Approach

- Dictionary-based approach
  - Easy to implement
  - High precision for known word
  - Low to detect new words
- Statistics-based approach
  - Requires large annotated corpus for training
- Comprehension approaches
  - Theoretically high precision but perform poor

Character-based Approach: Single Character-based

- Purely mechanical processes that extract certain number of characters (to form a string) from texts
- Divides Chinese texts into single characters
- Improved in Chinese Word Segmentation and Automated Indexing System (CWSAIS)
- Easy to implement
- Precision is low

Character-based Approach: Multi Character based

- Status of Chinese words

One Character	Two Character	Three Character	Four or More Character
5%	75%	14%	6%

- [Wu et al 1984]
  – segment texts into strings containing two (bigram), three or more characters
  – the bigram approach that segments a linear string of characters ABCDEF into AB, CD, EF and generates most of the correct Chinese word

Wu F, Zhou M. A comparator for Chinese text information retrieval.

Conditional Random Fields based

- Transfer word segmentation to Character labeling

  上海 / 计划 / 到 / 本 / 世纪 / 末 / 实现 / 人均 / 国内 / 生产 / 总值 / 五千美元 / 。

  上 / B海 / E计 / g划 / E到 / S 本 / s 世 / B纪 / E 末 / S实 / B现 / E人 / B 均 / E国 / g内 / E生 / B产 / E总 / B值 / E 五 / B千 / M美 / M元 / E。 / S

B: the first character of a multi-character word

M: intermediate character in a multi-character word

E: the last character in a multi-character word

S: one character word

- [Peng et al 2004, Tseng et al 2005, Zhou et al 2005] Conditional Random Fields
  - CRF are undirected graphical models trained to maximize a conditional probability of the whole graph structure
  - CRF is a discriminative model which can capture many correlated features of the inputs
  - Suitable for sequence labeling
  - More accurate than the generative models
- Advantages of Character-based Approach
  - Simplicity and ease of application
  - Reduced costs and minimal overheads

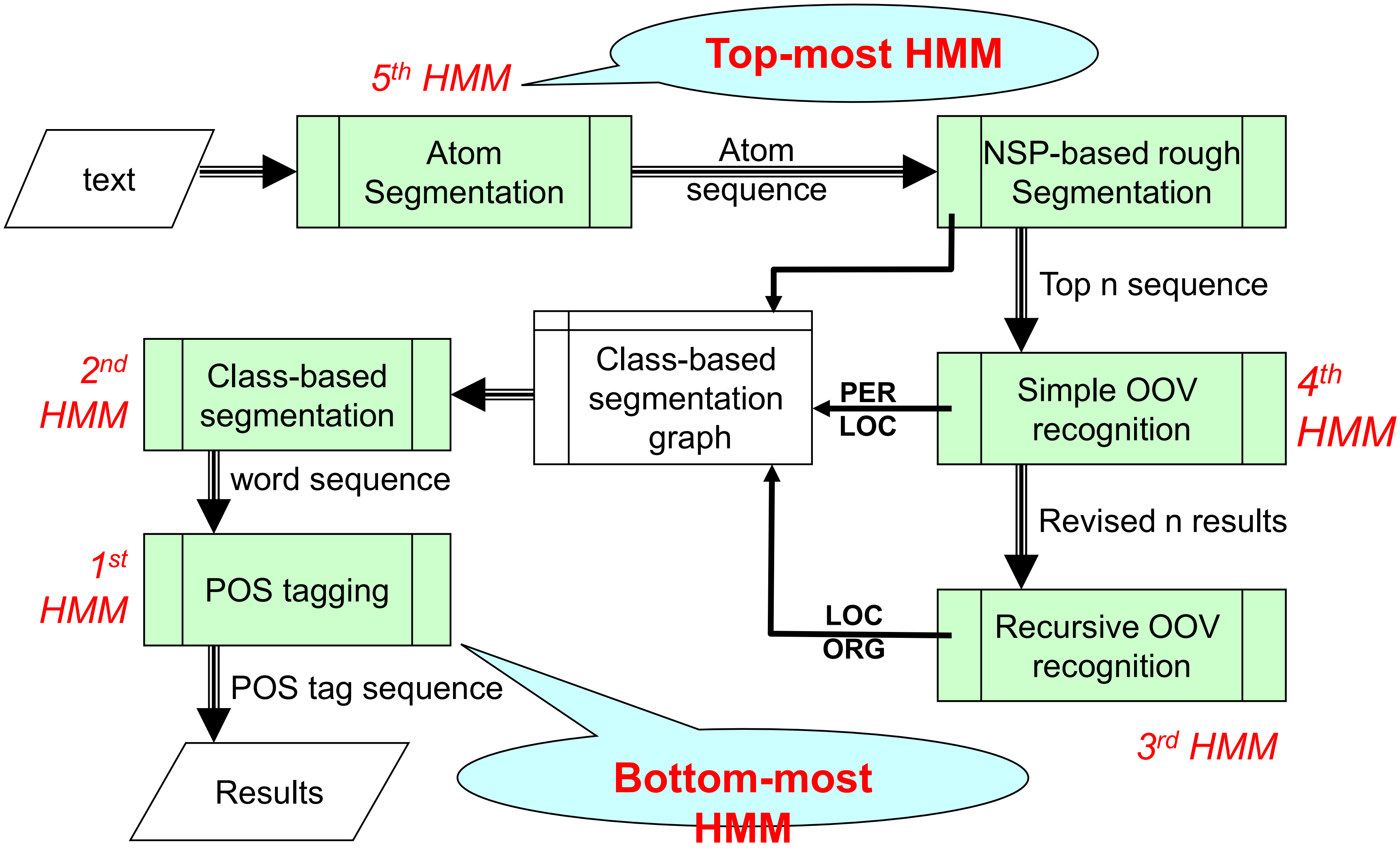
Peng F, Feng F，McCallum A. Chinese segmentation and new word detection using conditional random fields.
Tseng H, Chang P, Andrew G, Jurafsky D, Manning C. A conditional random fields word segmenter for SIGHAN bakeoff.
Zhou], Ni B, Chen]. A hybrid approach to Chinese word segmentation around CRFs.

Ref

ICTCLAS

- A Hierarchical Hidden Markov Model and Class-based Chinese word segmentation
- Developed by Huaping Zhang
- A widely used tool for it is open
- The performance is good

ICTCLAS: The Architecture



Ref

The Key Algorithms - 1

- Atom segmentation
  - To segment original text into atom sequence

    莱温斯基本来就很不爽

    ➔莱/温/斯/基/本/来/就/很/不/爽
- NSP-based rough segmentation
  - The top-N shortest path for segmentation

    ➔**(1)** 莱/温/斯/基本/来/就/很/不/爽

    ➔**(2)** 莱/温/斯/基/本来/就/很/不/爽

The Key Algorithms - 2

- OOV recognition
  - Locate boundary of a unknown word
  - Identify word class and
  - The association probability

    ➔莱温斯基/本来/就/很/不/爽

  - Role-based HMM

Ref

Role-based HMM - 1

- What is Role

角色	意义	例子
B	姓氏	张华平先生
C	双名的首字	张华平先生
D	双名的末字	张华平先生
E	单名	张浩说：“我是一个好人”
F	前缀	老刘、小李
G	后缀	王总、刘老、肖氏、吴妈、叶帅
K	人名的上文	又来到于洪洋的家。
L	人名的下文	新华社记者黄文摄
M	两个中国人名之间的成分	编剧邵钧林和稽道青说
U	人名的上文和姓成词	这里有关天培的壮烈
V	人名的末字和下文成词	龚学平等领导，邓颖超生前
X	姓与双名的首字成词	王国维、
Y	姓与单名成词	高峰、汪洋
Z	双名本身成词	张朝阳
A	以上之外其他的角色	

Ref

Role-based HMM

- Example: Role in Chinese Organization Name

中文机构名称构成角色表		
角色	意义	例子
A	上文	参与亚太经合组织的活动
B	下文	中央电视台*报道*
X	连接词	北京电视台*和*天津电视台
C	特征词的一般性前缀	
F	特征词的译名性前缀	美国*摩托罗拉*公司
G	特征词的地名性前缀	交通银行*北京*分行
H	特征词的机构名前缀	*中共中央*顾问委员会
I	特征词的特殊性前缀	*中央*电视台
J	特征词的简称性前缀	
D	机构名的特征词	
Z	非机构名成份	

Ref

Role-based HMM - 2

Word sequence	$W = (w_1, w_2, ..., w_n)$
Role sequence	$R = (r_1, r_2, ..., r_n)$
Class sequence	$C = (c_1, c_2, ..., c_n)$

Optimal role sequence

$$R^{\#} = \arg\min \sum_{i=1}^{n} -\ln p(c_i \mid r_i) - \ln p(r_i \mid r_i)$$

Possibility of word is assigned such a class label

$$p(w_i \mid c_i) = \prod_{j=0}^{k-1} p(c_{p+j} \mid r_{p+j}) \times \prod_{j=1}^{k-1} p(r_{p+j} \mid r_{p+j-1})$$

Ref

The Key Algorithms - 3

- Recursive OOV recognition
  - Higher level Role-based HMM
  - Round 1

p(周恩来|PER)=p(周|C)p(恩|D)p(来|E)p(C|D)p(D|E)
p(邓颖超|PER)=p(邓|C)p(颖|D)p(超|E)p(C|D)p(D|E)

  - Round 2

p(周恩来和邓颖超纪念馆|ORG)
=p(周恩来|PER)p(和|B)p(邓颖超|PER)
p(纪念馆|F)p(PER|B)p(B|PER)p(PER|F)

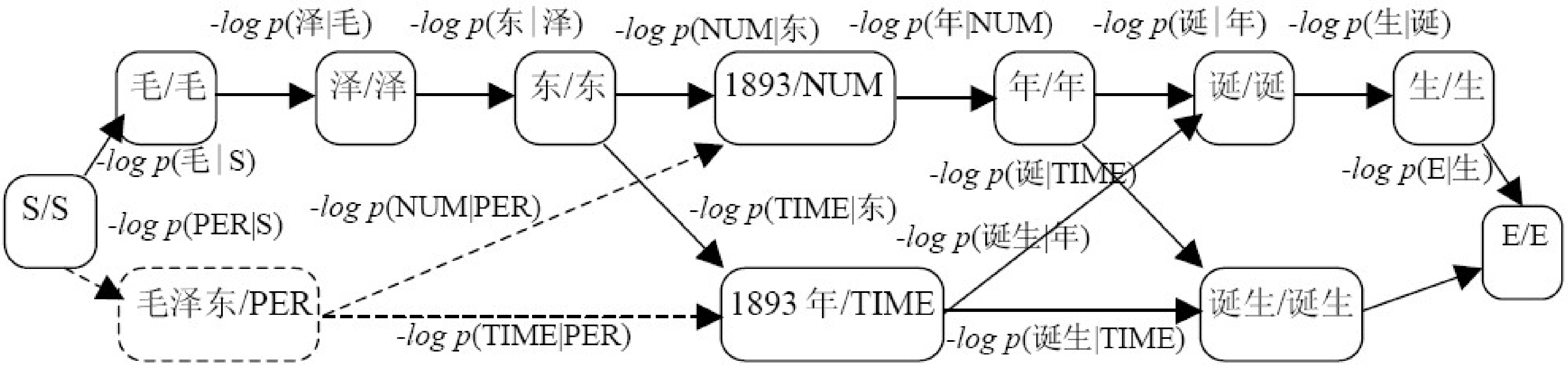
Ref

The Key Algorithms - 4

- Class-based HMM for word segmentation

Atom sequence	$A = (a_1, a_2, \ldots a_n)$
Word sequence	$W = (w_1, w_2, \ldots w_n)$
Class sequence	$C = (c_1, c_2, \ldots c_n)$

Optimal word sequence $W^{\#} = \arg\max_{W} p(W \mid A) = \arg\max_{W} p(W, A)\, p(A)$



Ref

The Key Algorithms - 5

- POS tagging
  - The top level HMM

Word sequence	$W=(w_1,w_2,\ldots,w_n)$
POS tag sequence	$T=(t_1,t_2,\ldots,t_n)$
Optimal tag sequence	$T^{\#}=\arg\max_{W} p(T\mid W)=\arg\max_{W} p(T,W)p(W)$

$$=\arg\max_{W} p(T,W)=\arg\max_{W} p(W,T\mid T)$$

$$\approx \arg\max_{w_1,w_2,\ldots w_n}\prod_{i=1}^{m} p(w_i\mid t_i)p(t_i\mid t_{i-1})$$

HMM

Ref

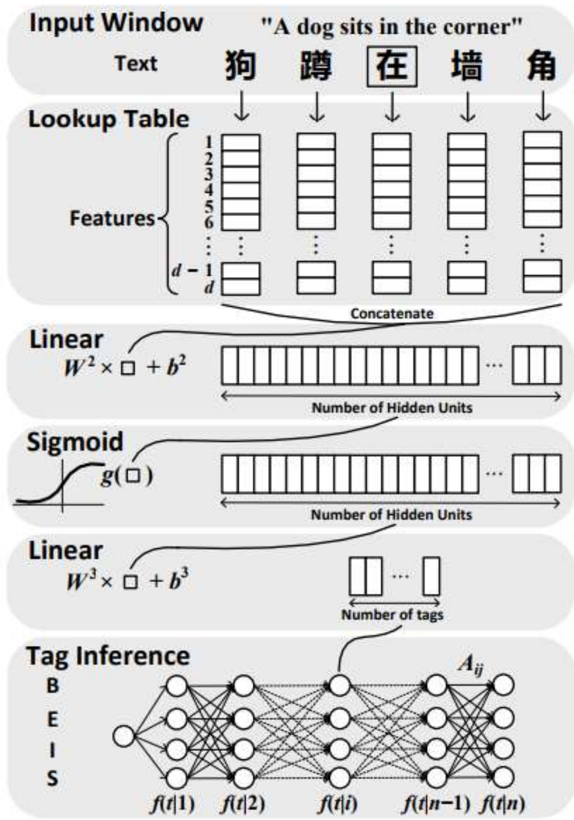
The Key Algorithms - 6

- The strategy to improve word segmentation and POS tagging at the same time.
  - Using a few word segmentation candidates rather than only one optimal considering only word
  - Applying p(w|c) to find the optimal result

Ref

Neural-based Approach

- Sequence labeling schemes approaches
  - Zheng et al. (2013) first adapted the sliding-window based sequence labeling (Collobert et al., 2011) with character embeddings as input.

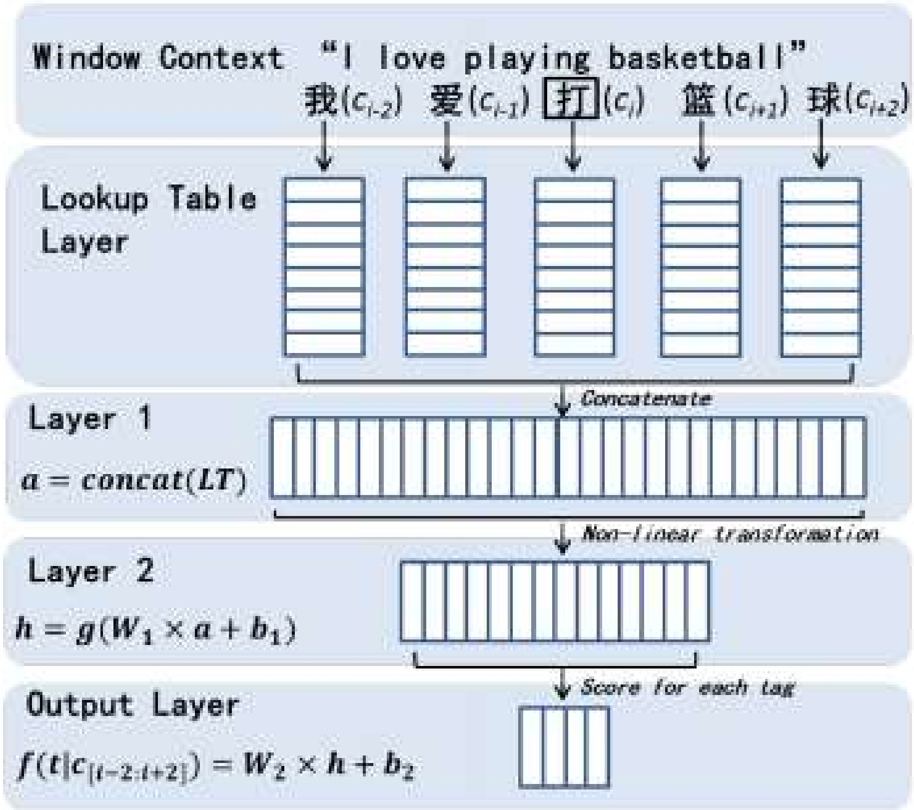


Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. Deep learning for Chinese word segmentation and POS tagging.
Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa.

Ref

Neural-based Approach

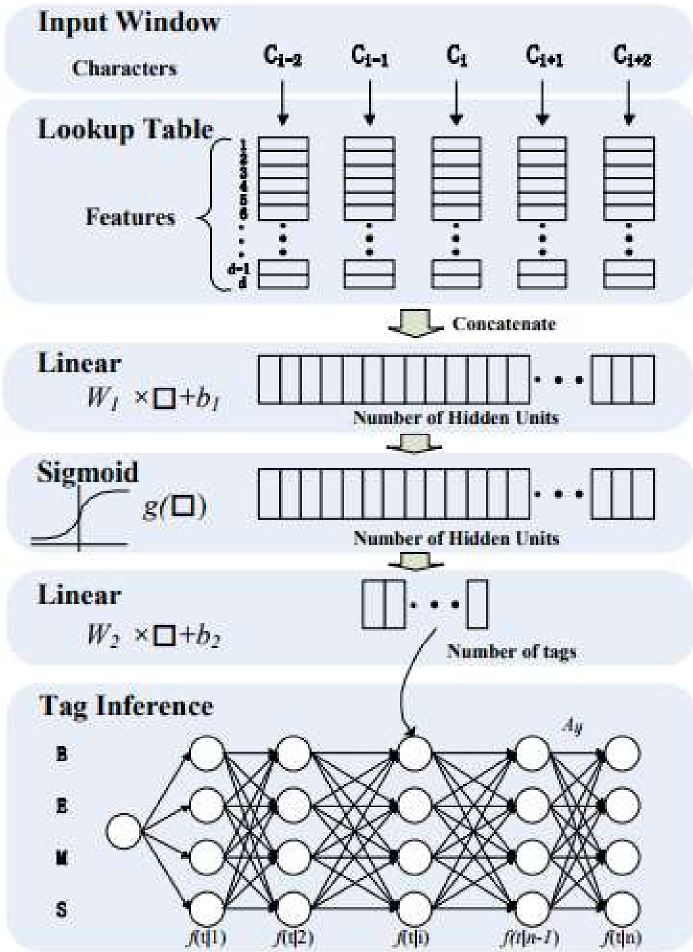
- Sequence labeling schemes approaches
  - Pei et al. (2014) introduced tag embedding.
  - Max-Margin Tensor Neural Network, MMTNN



Wenzhe Pei, Tao Ge, and Baobao Chang. Max-margin tensor neural network for Chinese word segmentation.

Neural-based Approach

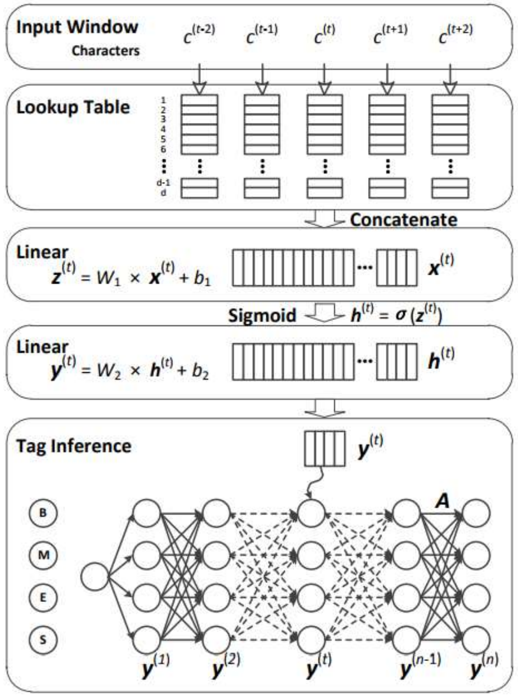
- Sequence labeling schemes approaches
  - Chen et al. (2015a) proposed to model n-gram features via a gated recursive neural network (GRNN)



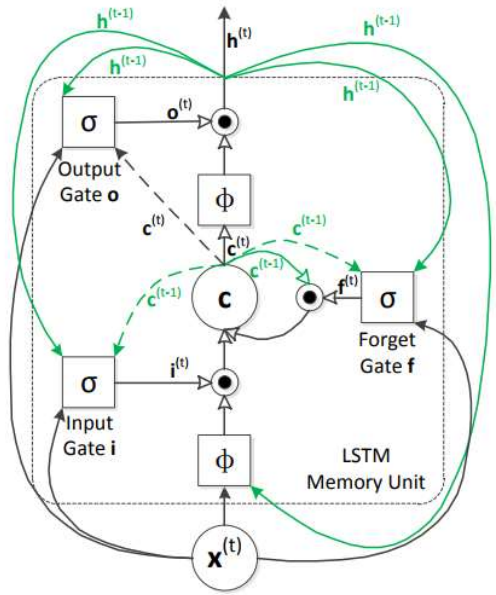
Xinchi Chen, Xipeng Qiu, Chenxi Zhu, and Xuanjing Huang. Gated recursive neural network for Chinese word segmentation.

Neural-based Approach

- Sequence labeling schemes approaches
  - Chen et al. (2015b) used a Long shortterm memory network (LSTM) (Hochreiter and Schmidhuber, 1997) to capture long-distance context.



General architecture of neural model for Chinese word segmentation.



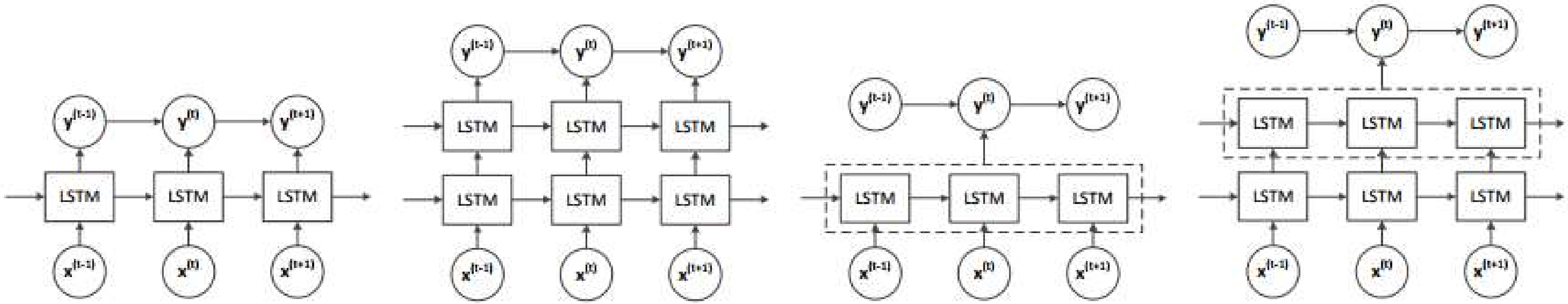
LSTM Memory Unit.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. Long short-term memory neural networks for Chinese word segmentation.

Ref

Neural-based Approach

- Sequence labeling schemes approaches
  - Chen et al. (2015b) used a Long-short term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) to capture long-distance context.



(a) LSTM-1 (b) LSTM-2 (c) LSTM-3 (d) LSTM-4

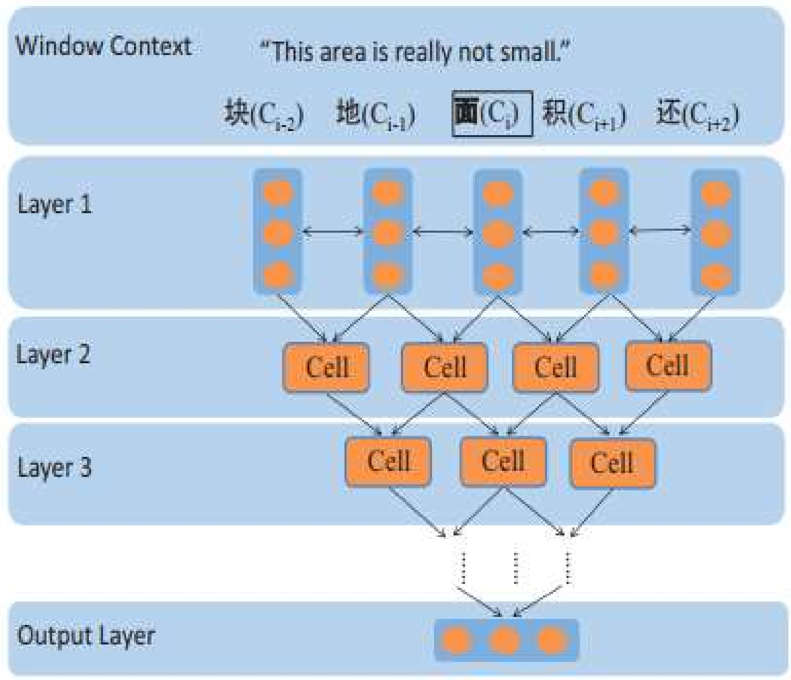
Chen proposed LSTM architectures for Chinese word segmentation.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. Long short-term memory neural networks for Chinese word segmentation.

Ref

Neural-based Approach

- Sequence labeling schemes approaches
  - Xu and Sun (2016) integrated both GRNN and LSTM for deeper feature extraction.
  - Called dependency-based gated recursive neural network(DGRNN).



Architecture of DGRNN for Chinese Word Segmentation. Cell is the basic unit of GRNN.

Jingjing Xu and Xu Sun. Dependency-based gated recursive neural network for Chinese word segmentation.

Neural-based Approach

- Sequence labeling schemes approaches
  - Traditional Model ***VS*** Neural Sequence Model

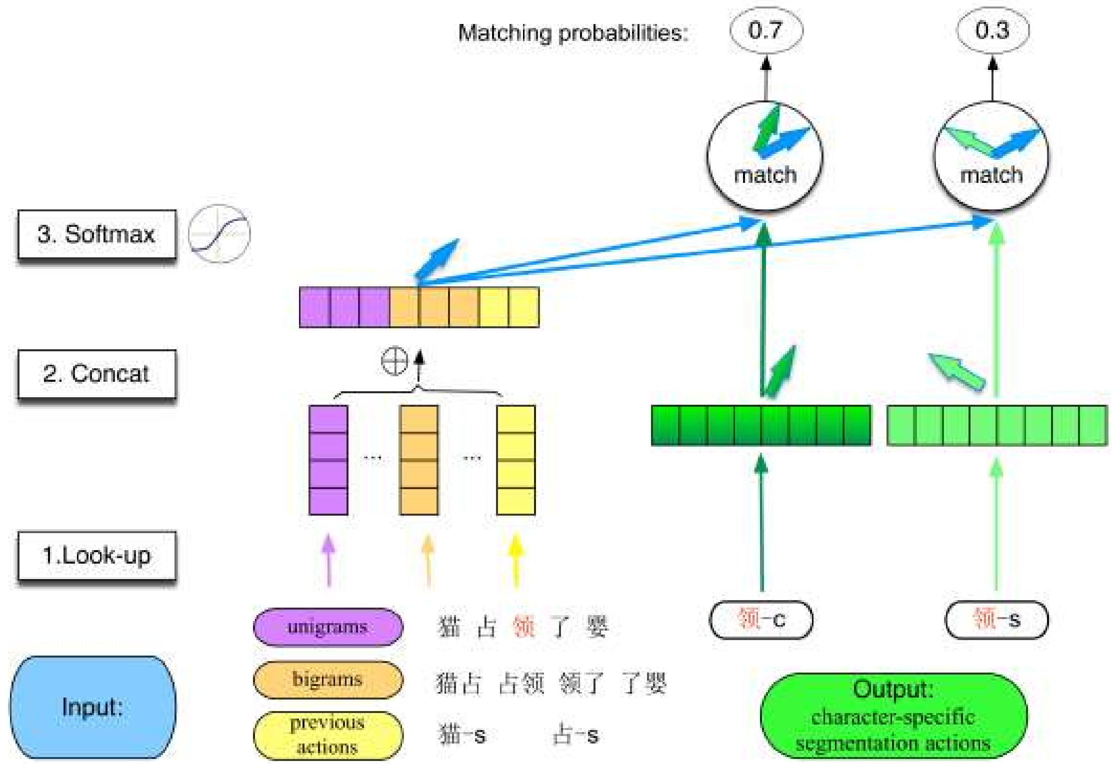
结构化建模演进（每行展示结构化建模对应的两种模型）

结构分解	传统模型	神经模型
分类模型	Xue (2003)	
Markov模型	Ng and Low (2004); Low et al. (2005)	Zheng et al. (2013) Pei et al. (2014)
标准串学习建模	CRF: Peng et al. (2004) semi-CRF: Andrew (2006); Sun et al. (2009)	LSTM: Chen et al. (2015b) Liu et al. (2016)
全局模型	Zhang and Clark (2007)	Cai and Zhao (2016) Cai et al. (2017)

Ref

Neural-based Approach

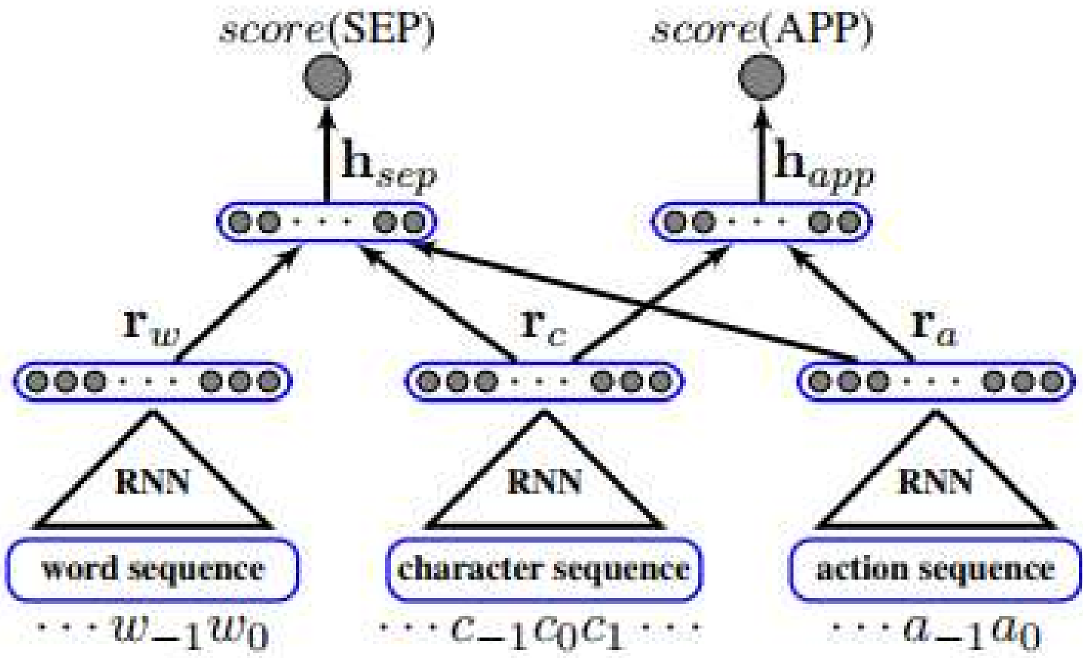
- Other schemes approaches
  - Ma & Hinrichs (2015) proposed a character-based embedding matching approach.

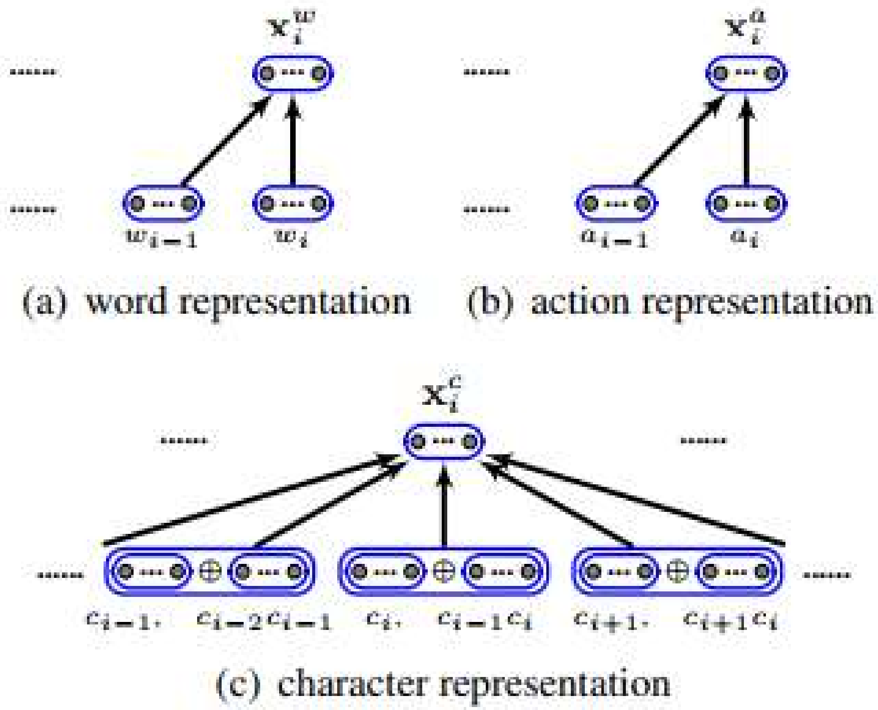


Jianqiang Ma and Erhard Hinrichs. Accurate linear-time Chinese word segmentation via embedding matching.

Neural-based Approach

- Other labeling schemes approaches
  - Zhang et al. (2016) proposed a transition-based framework.





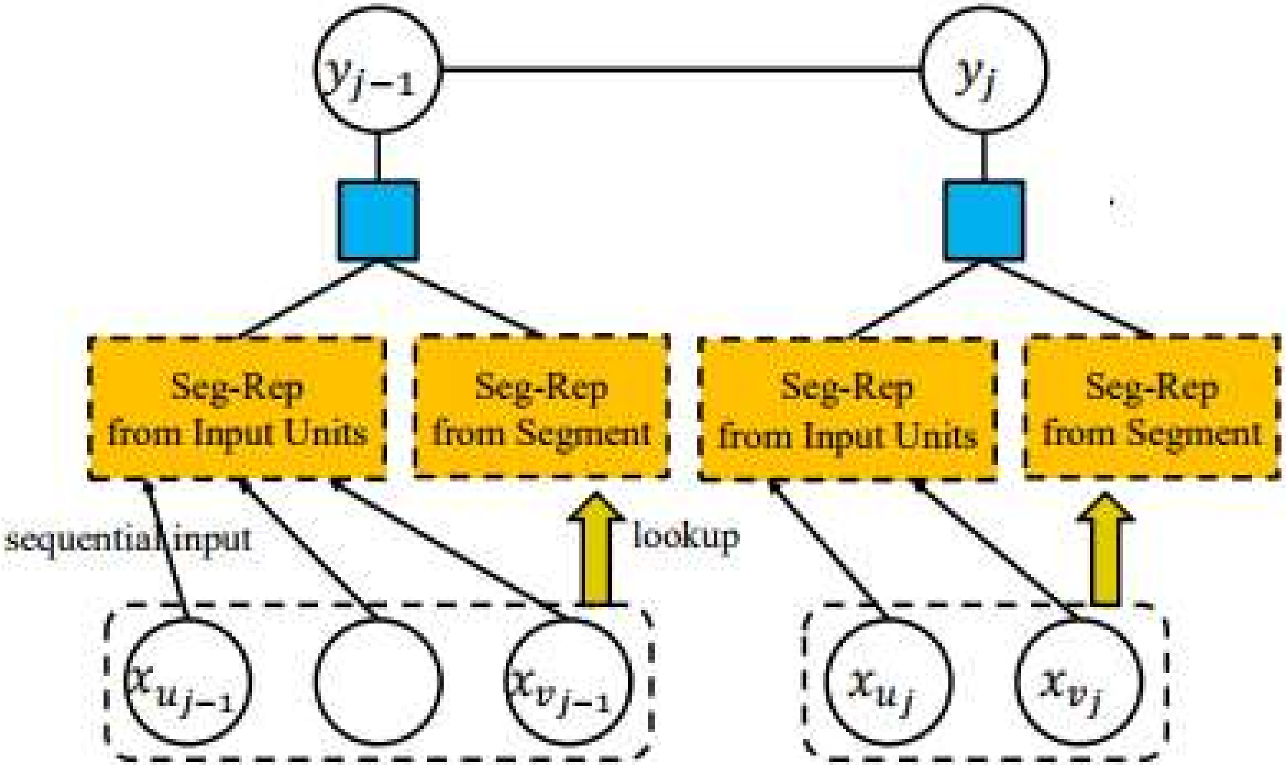
Meishan Zhang, Yue Zhang, and Guohong Fu. Transition-based neural word segmentation.

Scorer for the neural transition-based Chinese word segmentation model.

Input representations of LSTMS

Neural-based Approach

- Other labeling schemes approaches
  - Liu et al. (2016) used a zero-order semi-CRF based model.



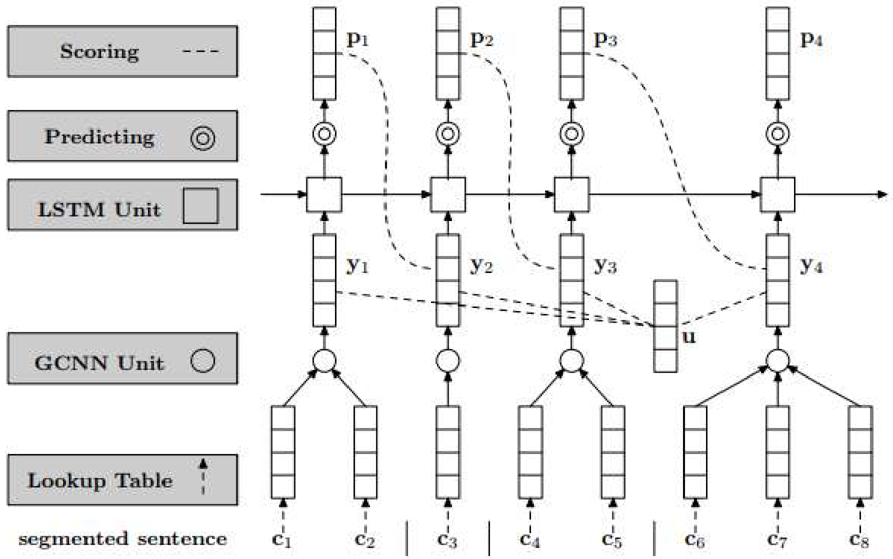
Yijia Liu, Wanxiang Che, Jiang Guo, Bing Qin, and Ting Liu. Exploring segment representations for neural segmentation models.

neural semi-CRF model with segment representation from input composition and segment embeddings.

Ref

Neural-based Approach

- Other labeling schemes approaches
  - Cai and Zhao (2016) proposed to score candidate segmented outputs directly
  - Employing a gated combination neural network over characters for word representation generation and an LSTM scoring model for segmentation result evaluation.



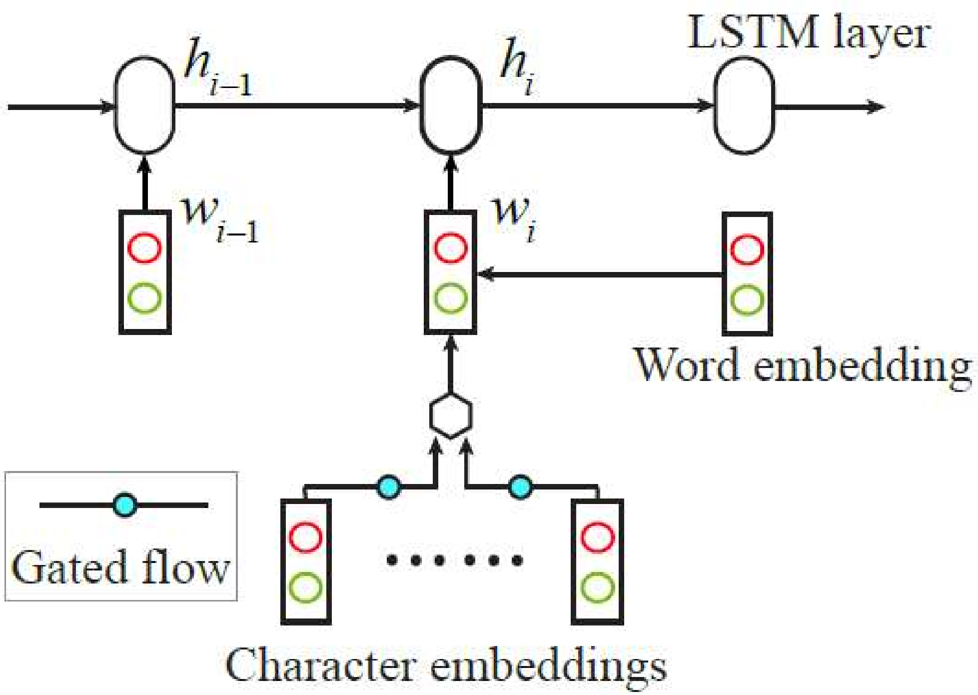
Architecture of proposed neural network scoring model

Deng Cai and Hai Zhao. Neural word segmentation learning for Chinese.

Ref

Neural-based Approach

- Other labeling schemes approaches
  - Cai and Zhao (2017) presented a fast and accurate word segmentor using neural networks.



Neural network scoring for word candidate

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. Fast and accurate neural word segmentation for Chinese.

Recent Works

- Is Word Segmentation necessary for deep learning of Chinese representations?
  – Meng et al.(2019) found that in the Chinese environment, **using character-based models (e.g., BERT) is more suitable than word-based models**. As the latter often suffers from data sparsity and out-of-vocabulary problems.
  – The experiments on four end-to-end NLP benchmark tasks: Language Modeling, Machine Translation, Sentence Matching/Phrasing and Text Classification.

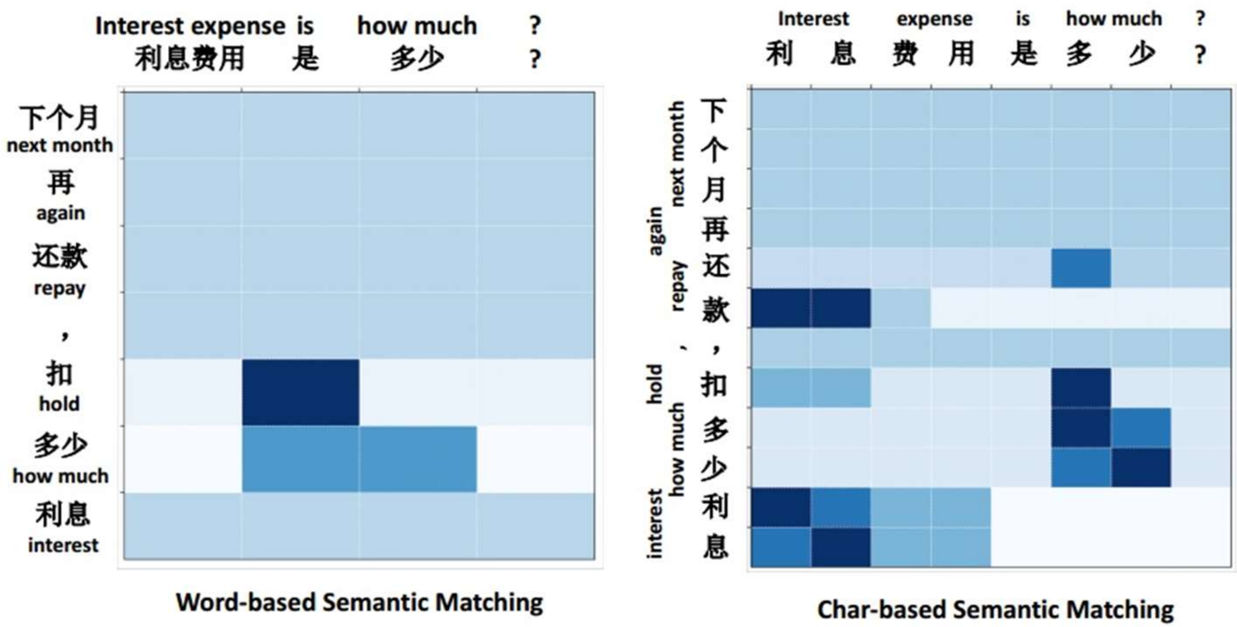
Yuxian Meng, Xiaoya Li, Xiaofei Sun, Qinghong Han, Arianna Yuan and Jiwei Li. Is Word Segmentation Necessary for Deep Learning of Chinese Representation. ACL 2019.

Recent Works

- Results on text classification datasets

Dataset	description	char valid	word valid	char test	word test
chinanews	1260K/140K/112K	91.81	91.82	91.80	**91.85** (+0.05)
dianping	1800K/200K/500K	78.80	78.47	**78.76** (+0.36)	78.40
ifeng	720K/80K/50K	86.04	84.89	**85.95** (+1.09)	84.86
jd_binary	3600K/400K/360K	92.07	91.82	**92.05** (+0.16)	91.89
jd_full	2700K/300K/250K	54.29	53.60	**54.18** (+0.81)	53.37

- Visualization



Yuxian Meng, Xiaoya Li, Xiaofei Sun, Qinghong Han, Arianna Yuan and Jiwei Li. Is Word Segmentation Necessary for Deep Learning of Chinese Representation. ACL 2019.

Recent Works

- Neural Chinese Word Segmentation with Dictionary Knowledge
  – Liu et al.(2019) find that many neural network based methods require a large number of labeled sentences and usually **cannot utilize the useful information in Chinese dictionary.**
  – They proposed two methods to exploit the dictionary information for Chinese Word Segmentation.
  – The experiments on on two benchmark Chinese word segmentation datasets.

Junxin Liu, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, Xing Xie. Neural Chinese word segmentation with dictionary. Neurocomputing, 2019

Recent Works

- Results on CWS datasets

	1%			10%			100%		
	P	R	F	P	R	F	P	R	F
Chen et al. [3]	75.50	75.80	75.64	87.71	86.22	86.96	94.24	93.35	93.80
LSTM-CRF	75.88	74.86	75.36	85.52	84.81	85.16	94.26	93.29	93.78
CNN-CRF	75.59	74.43	75.00	89.72	89.14	89.43	95.03	94.53	94.78
Zhang et al. [17]	75.75	75.95	75.85	89.52	89.01	89.27	95.71	95.41	95.56
Ours_Pseudo	80.58	77.97	79.25	90.49	89.59	90.04	95.36	94.71	95.03
Ours_Multi	78.47	77.31	77.88	89.91	89.27	89.59	95.10	94.50	94.80

	5%			25%			100%		
	P	R	F	P	R	F	P	R	F
Chen et al. [3]	82.31	82.60	82.44	88.00	89.90	88.94	90.79	92.92	91.84
LSTM-CRF	81.08	80.88	80.98	86.76	88.40	87.57	91.39	92.58	91.98
CNN-CRF	82.44	84.50	83.46	89.95	91.57	90.75	92.22	93.84	93.02
Zhang et al. [17]	83.38	84.98	84.17	89.93	91.41	90.66	92.60	93.89	93.24
Ours_Pseudo	87.37	86.56	86.97	90.97	92.04	91.50	92.77	94.09	93.43
Ours_Multi	84.59	86.22	85.40	90.43	91.68	91.05	92.35	93.93	93.13

- Case Study

	Example 1	Example 2
Original	5名男子和被害人有恩怨	警方一口气带回了5 0多人
CNN-CRF	5/名/男子/和/被/害/人/有/恩怨	警方/一/口/气/带回/了/5 0多/人
+Internal dictionary	5/名/男子/和/被/害/人/有/恩怨	警方/一口气/带回/了/5 0多/人
+External dictionary	5/名/男子/和/被害人/有/恩怨	警方/一口气/带回/了/5 0多/人

Junxin Liu, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, Xing Xie. Neural Chinese word segmentation with dictionary. Neurocomputing, 2019

Benchmark - Standards

- Peking University Standard
  - Contemporary Chinese Corpus
  - Characterized by a large set of specific rules
  - Each rule explains how a particular kind of character string should be segmented
  - Grammatical Knowledge-Base of Contemporary Chinese：73,000 entries
  - Used as a reference lexicon

- Academia Sinica Standard
  - Sinica Corpus developed by Academia Sinica
  - Assumes the use of a reference lexicon for segmentation
  - No large set of specific rules
  - Two segmentation principles and four specific segmentation guidelines
  - A segmentation unit is defined as the smallest string of characters with an independent meaning and a fixed grammatical category

- University of Pennsylvania Standard
  - Similar to the Peking University Standard in the sense
  - Also consists of a large set of specific rules.
  - Each rule specifies how a particular kind of character string should be handled
  - Do not assume the use of any reference lexicon for segmentation.
  - Needs to determine the wordhood status for each character string without lexicion
  - The rules in the standard attempt to cover all possible scenarios, and the set become inevitably large

Benchmark - Bakeoffs

- To compare the accuracy of various method
- The "Bakeoff evaluation" first held at the 2nd SIGHAN Workshop at ACL in 2003
- Four corpora
  - Academia Sinica, Hong Kong CityU, Upenn, PKU
- Open Track
  - Allowed to use any other resources such as dictionaries or more training data
- Close Track
  - Only training data is allowed

Bakeoff Evaluation

- SigHan Bakeoff are evaluated in five measurements
  - recall, precision and F-measure for overall segmentation
  - recall for unknown words and known words

$$Recall = \frac{number\ of\ correctly\ segmented\ words}{total\ number\ of\ words\ in\ gold\ data}$$

$$Precision = \frac{number\ of\ correctly\ segmented\ words}{total\ number\ of\ words\ segmented}$$

$$F\text{-}measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

$$Recall(OOV) = \frac{number\ of\ correctly\ segmented\ unknown\ words}{total\ number\ of\ unknown\ words\ in\ gold\ data}$$

$$Recall(IV) = \frac{number\ of\ correctly\ segmented\ known\ words}{total\ number\ of\ known\ words\ in\ gold\ data}$$

4^{th} Bakeoff Evaluation

- New Tasks
  - Chinese Word Segmentation
  - Chinese Named Entity Recognition
  - Chinese POS-tagging
- Seven Corpora
  - Academia Sinica, Hong Kong CityU, Microsoft Research Asia, PKU, Shanxi University, National Chinese Corpus, Chinese Tree Bank

Agreement on Words

- Wu Dekai 1995 nk-blind
  - N person segmentation
  - K person agree $0<k<n$
  - N=8 all-agree rate 30% at-least-one agree rate 90%
- Sproat 1996
  - Based on one person's output
  - Agree rate 0.76

Agreement on Words

- Human Agreement on Words

	M1	M2	M3	T1	T2	T3
M1		0.77	0.69	0.71	0.69	0.70
M2			0.72	0.73	0.71	0.70
M3				0.89	0.87	0.80
T1					0.88	0.82
T2						0.78

- Corpora Agreement

测试语料库	分 词 系 统			
	As	CTB	CityU	MSRA
AS2006	1.0	0.959 3	0.925 6	0.858 3
CTB2006	0.942 0	1.0	0.910 4	0.877 4
CityU2006	0.932 1	0.934 6	1.0	0.848 8
MSRA2006	0.857 0	0.886 6	0.848 3	1.0

Open Resources

- NLPIR ICTCLAS
  - Beijing Institute of Technology
  - http://ictclas.nlpir.org/

- JIEBA Word Segmentation System
  - https://github.com/fxsjy/jieba

- HIT LTP Platform
  - www.**ltp**-cloud.com/

The Next Lecture

- Lecture 6

  Unknown Word Identification and Normalization

Questions？