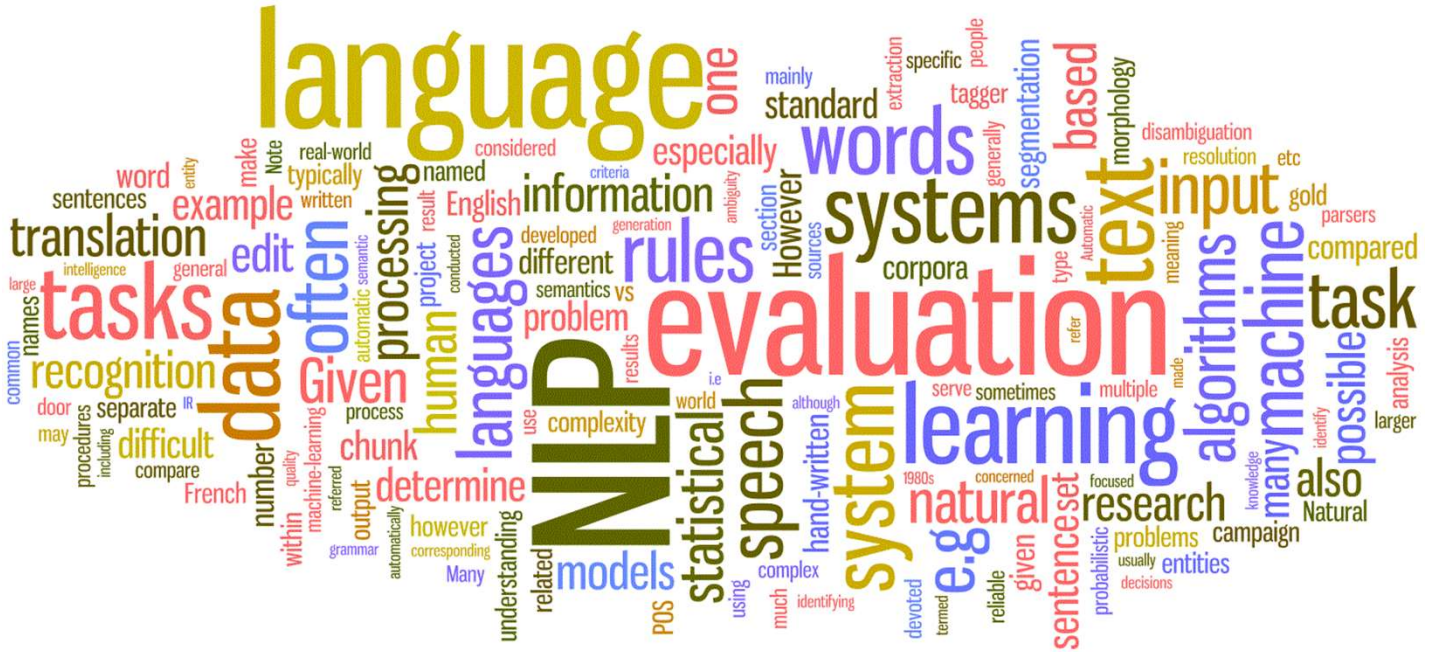


Lecture 6: Unknown Word Identification and Normalization

Xu Ruifeng

Harbin Institute of Technology, Shenzhen

Last Time

- Why Chinese Word Segmentation
- Difficulties in Chinese Word Segmentation
- Chinese Word Segmentation Algorithms
- Chinese Word Segmentation Evaluation

Today's class – Unknown Word Process

- Unknown Word Identification
  - Problem Statement
  - Unknown Word Identification
  - Chinese Person Name Identification
  - Chinese Organization Name Identification
  - Chinese Place Name Recognition
- Unknown Word Normalization : Abbreviation Expansion

Unknown Word

- Out Of Vocabulary (OOV) problem
  - Rapid growing new words 非典 给力 国考
  - OOV caused errors are about 5 times to ambiguity caused errors
- Five major types :
  - Abbreviation (acronym)
    - 国考=国家公务员考试
  - Proper name/Named Entity
    - Names of people PG One
    - Names of places 延坪岛
    - Names of organizations 上海合作组织
    - Certain key words are indicators for each different sub-category

- Derived words:
  - Have affix morphemes which are strong indicators
  - 审计 -> 审计人　审计员　审计局 审计处
- Compounds:
  - A word made up of other words
  - 光+敏感 -- 光敏感　　流体+力学 – 流体力学
- Numeric type compounds：
  - 五月三日　八点十分　第一

Unknown Word Detection and Recognition

- Decide whether an OOV string can form a word or not
  - Rule-based approach
    - Compositional Rule
    - Reject Rule
  - Statistics-based approach
    - In-word probability
    - Boundary probability
  - Learning-based approach

Typical Features and Techniques

- Frequency/Relative Frequency
  - Starting point to pick up new word candidates
  - low frequency new words are hardly identifiable
- In-word probability
  - Reflect the morphological property of Chinese
  - [Chen et al 2005; Wu et al 2000; Fu 2001] used in-word probability to combine adjacent single characters after initial segmentation
- Word Formation Pattern [Wu et al 2000]
  - Describe how likely a character appears in a certain position within a word

Chen A, Zhou Y, Zhang A, et al. Unigram language model for Chinese word segmentation.
Wu A,Jiang Z. Statistically-enhanced New Word identification in a rule-based Chinese system.
Fu G. Research on statistical methods of Chinese syntactic disambiguation.

- Mutual information [Church et al 1990]
  - Estimates the internal association strength among constituents of character n-grams
- Left/right entropy [Sornlertlamvanich et al 2000] and context dependency [Chien 1999]
  - Describe the dependency strength of current item (character sequence) on its context
  - The dependency strength of a current item on its context decreases as the probability of this item to be a Chinese word increases

Church kW, Hanks P. Word association norms, mutual information, and lexicography.
Chien LF. PAT-tree based adaptive keyphrase extraction for intelligent Chinese information retrieval.
Sornlertlamvanich V, Potipiti T, Charoenporn T. Automatic corpus-based Thai word extraction with the C4.5 learning algorithm.

- Independent word probability [Nie et al 1995]
  - A property of a single character or a string of characters
  - Likelihood for this character to appear as an independent word in texts

$$IWP(c) = \frac{N(Word(c))}{N(c)}$$

- Anti-word list [Li et al 2004; Nie et al 95]
  - A list of functional characters to exclude bad candidates
  - Single character prepositional, adverbial and conjunctive words etc.

Nie JY, Hannan ML, Jin WY. Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge.
Li H, Huang CN, Gao J, Fan X. The use of SVM for Chinese new word identification.

- Similarity between new words and dictionary words [Li et al 2004]
  - If two characters appear more times in the same word patterns, the analogy between them is more reliable
  - 下 (below) has the most identical word patterns with 上 (above), and there is a strong preference

$$ANA(a,x) = \frac{\sum_c (W(ac)W(xc) + W(ca)W(cx))}{\sum_c (W(ac) + W(ca) + W(xc) + W(cx))}$$

Li H, Huang CN, Gao J, Fan X. The use of SVM for Chinese new word identification.

- [He et al 2017] propose a unified model which can learn from out-of-domain corpora and in-domain unannotated texts.
  - The unified model contains two major functions.
  - Cross-domain learning function can learn out-of-domain information based on domain similarity.
  - Semi-Supervised learning function can learn in-domain unannotated information by self-training.

He H, Sun X. A Unified Model for Cross-Domain and Semi-Supervised Named Entity Recognition in Chinese Social Media.

- [Rui et al 2016] proposes a method to detect unknown words during natural reading of non-native language text
- using eye-tracking features
- They propose several eye gaze features and classifiers to detect unknown words.

Rui Hiraoka, Hiroki Tanaka, Sakriani Sakti. Personalized Unknown Word Detection in Non-native Language Reading using Eye Gaze.

- [Caglar et al 2016] propose a novel way to deal with the rare and unseen words
- the neural network models using attention.

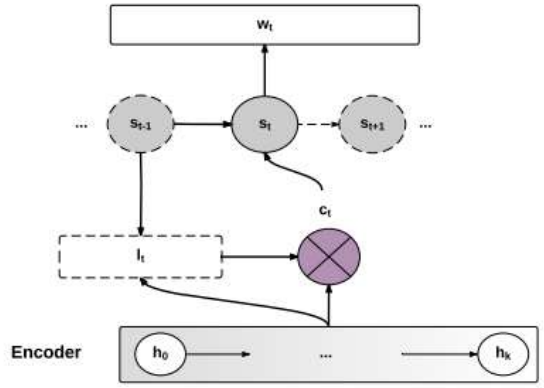


Figure 2: A depiction of neural machine translation architecture with attention. At each timestep, the model generates the attention weights $\mathbf{l}_t$. We use $\mathbf{l}_t$ the encoder's hidden state to obtain the context $\mathbf{c}_t$. The decoder uses $\mathbf{c}_t$ to predict a vector of probabilities for the words $\mathbf{w}_t$ by using softmax.

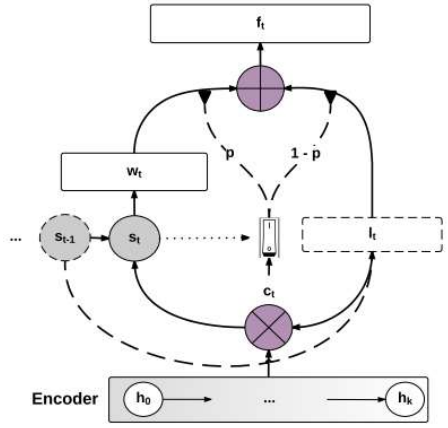


Figure 3: A simple depiction of the Pointer Softmax(PS) architecture. At each timestep as usual$\mathbf{l}_t$, $\mathbf{c}_t$ and the $\mathbf{w}_t$ for the words over the limited vocabulary(shortlist) is being generated. We have an additional switching variable z_t that decides whether to use $\mathbf{w}_t$ or copy the word from the input via $\mathbf{l}_t$. The final word prediction will be performed via pointer softmax $\mathbf{f}_t$ which can either copy the word from the source or predict the word from the shortlist vocabulary.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati. Pointing the Unknown Words.

Chinese Person Name

- A Chinese person name in Chinese is a family name + a given name
  - 李刚 朱五六 欧阳奋强 公孙杵臼 陈方安生
  - 25%+ out-of-vocabulary words
- Challenges:
  - Name formation is arbitrary
  - A name does not have “boundary tokens”
  - Different news agencies and people in different areas may translate the same foreign name to different Chinese names

Chinese Person Name Identification

- Driven by family name.
  - Less than 300 family names are commonly used
  - 千家姓 百家姓总汇 姓氏词典
- Detection of the name right boundaries

Types of Ordinary Words in Names	Examples
Two Character Name	高峰, 文静, 支流,幸运,关键
Three Character Name	黄灿灿
The First Two Characters	刘海亮, 黄金荣
The Last Two Characters	朱俊杰, 叶海燕

- Name identification using contextual information
  - Many studies constructed context word list
  - Title words: 总理/厂长
  - Verbs that are normally followed by person names,授予 接见
  - Patterns that contain both left and right context words flanking the names: 任命 … 为” and “记者 … 报道”

- Radical is the primary component of a Chinese character
  - Semantic radicals
    - Conveys the range or some aspect of the meaning of the whole character 玉 –璧 莹 瑜 琼 琪 玲 琅
    - Gender 蔷 薇 / 杏 桃
    - Contemporary characteristics 李援朝 王建国
    - Political Taboo 李世勣 →李勣 党怀英
  - Phonetic radicals
    - Often related to the pronunciation of the whole character

Transliteration of Foreign Names

- Foreign names are usually transliterated from the pronunciations in the original languages : Google, Facebook, Twitter
- Can be any lengths
- Features:
  - Chinese characters that appear particularly often in person name transliterations 尔 姆 斯
  - Boundary characters/string 夫、斯基
  - Structural feature 阿历克谢.马克西莫维奇.彼什科夫
  - Contextual information

Case Study: HMM Based Person Name Identification

- Transfer name identification problem into tagging problem:
  - Input text：

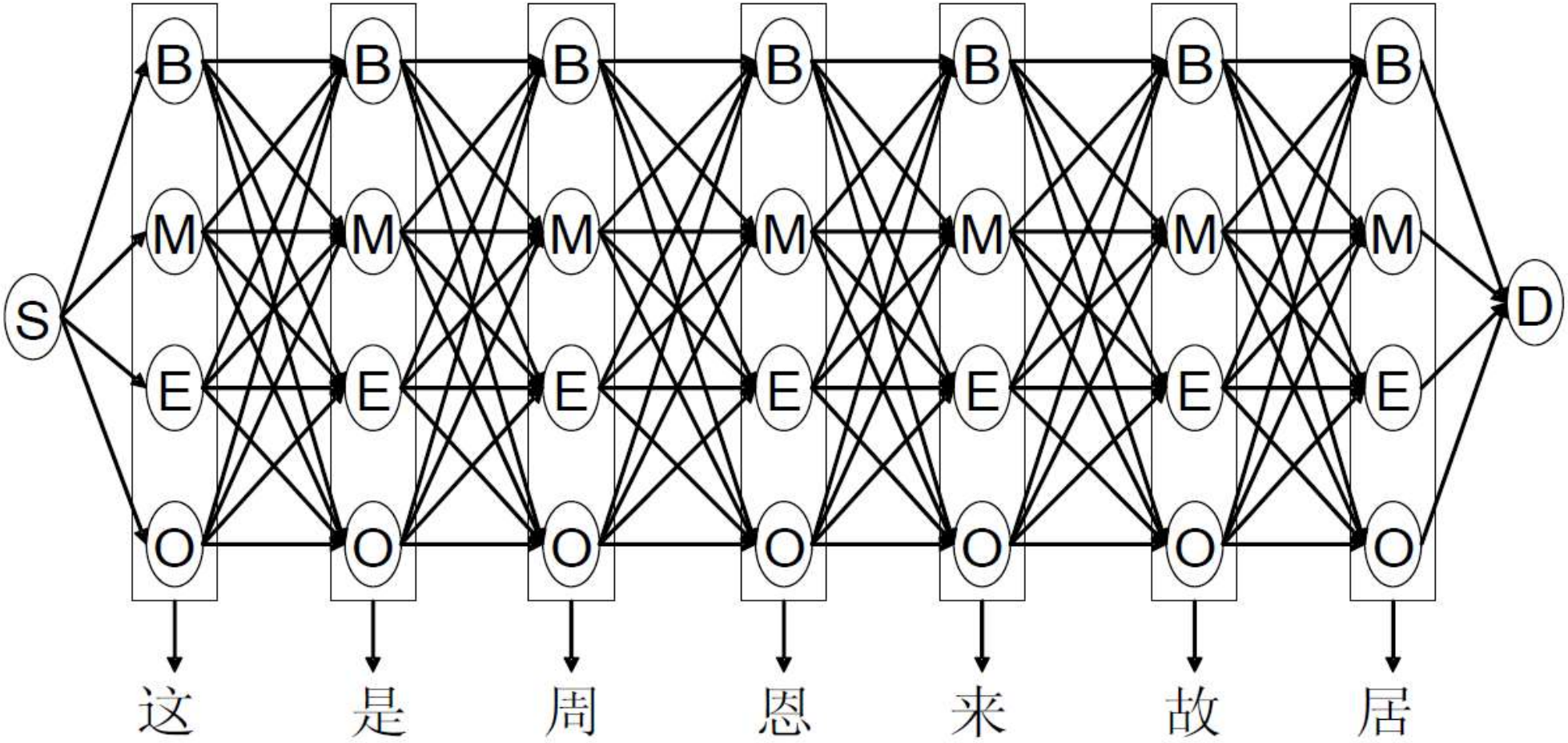
    这 是 周 恩 来 、 邓 颖 超 生 前 居 住 的 地 方
  - Tagged result：

    这 是 周 恩 来 、 邓 颖 超 生 前 居 住 的 地 方

    O O B M E O B M E O O O O O O O
  - O: not OOV, B: The first character of OOV
  - M: The middle character of OOV E: The last character of OOV
- "周恩来" and "邓颖超" which were tagged to "BME" are identified as OOV
- Use OOV tagged corpus as training corpus

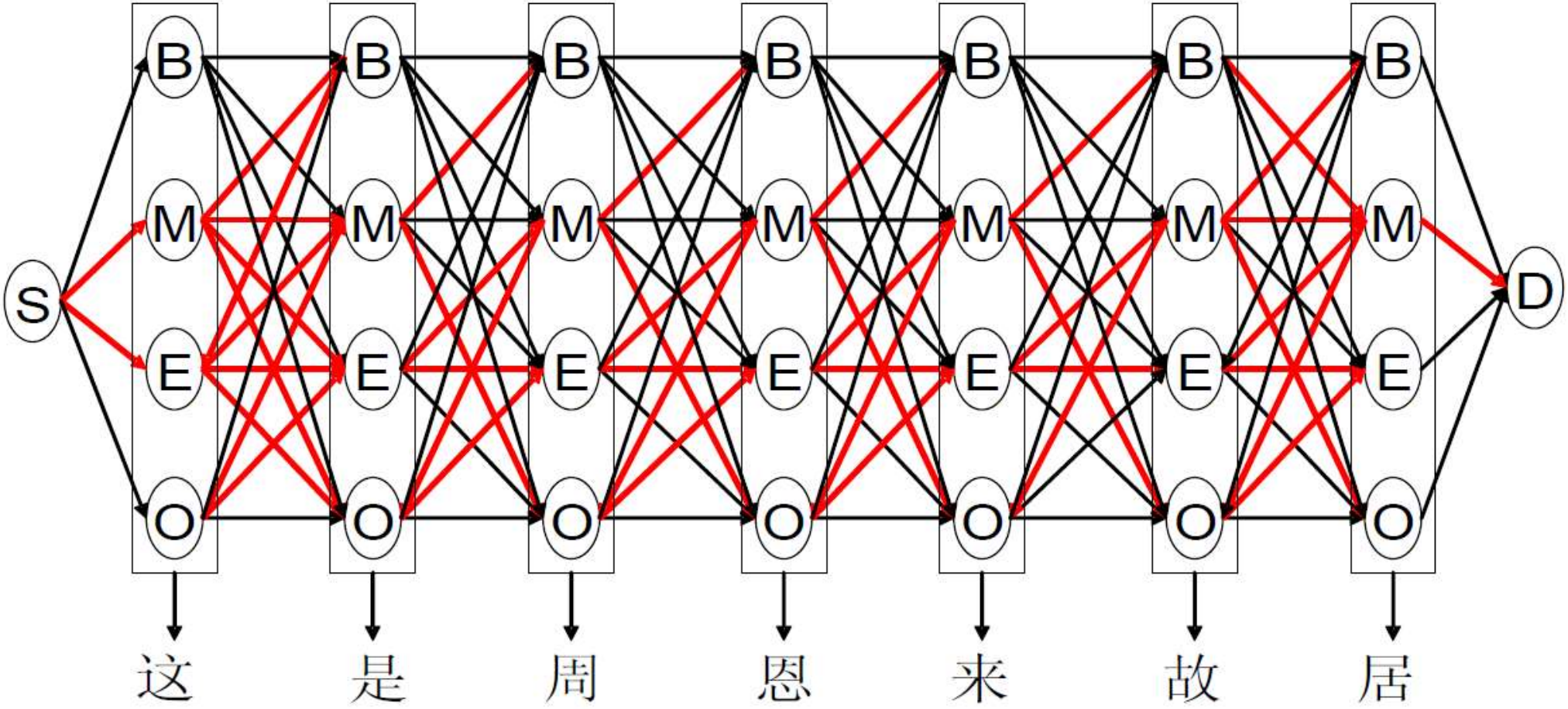
HMM Based Person Name Identification

Generate all of the edges



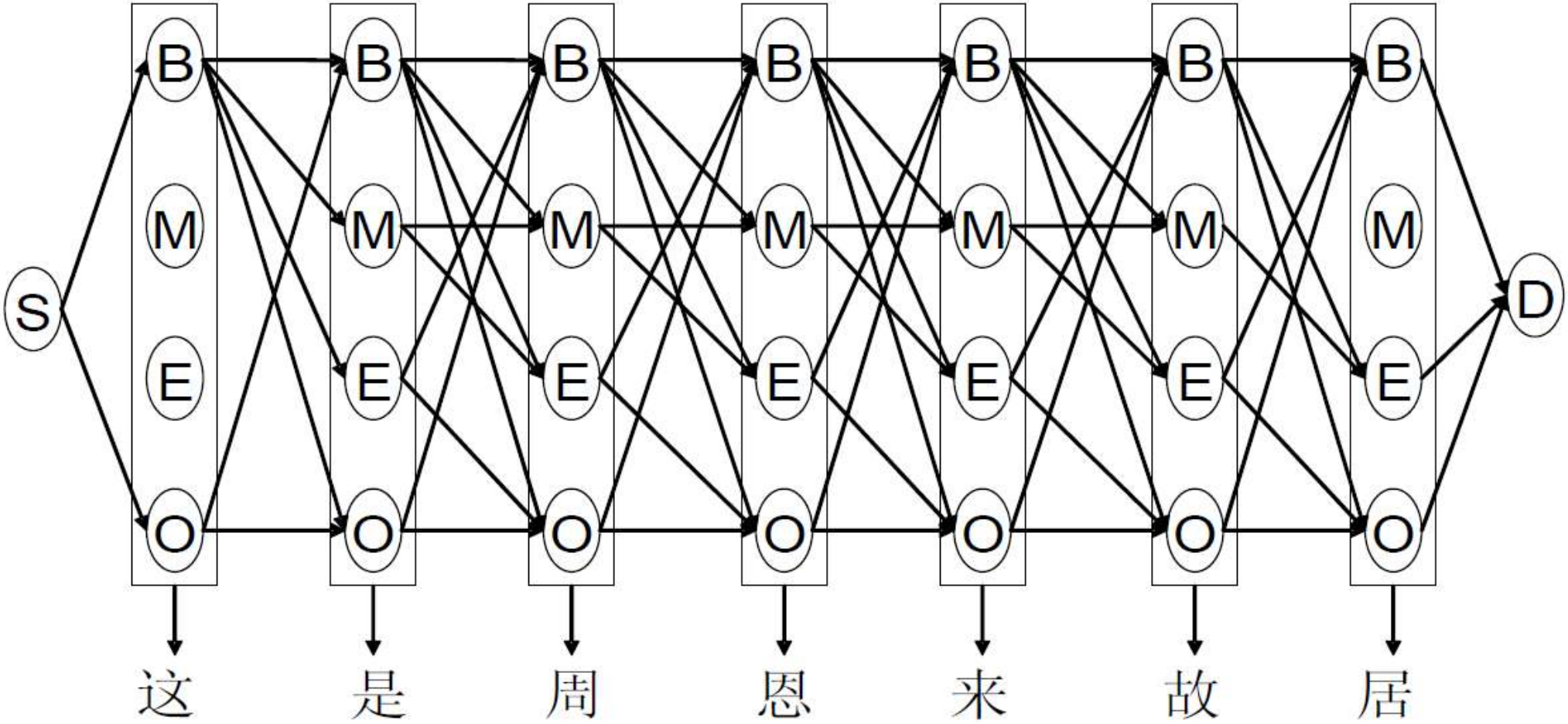
HMM Based Person Name Identification

- Unreachable edge:



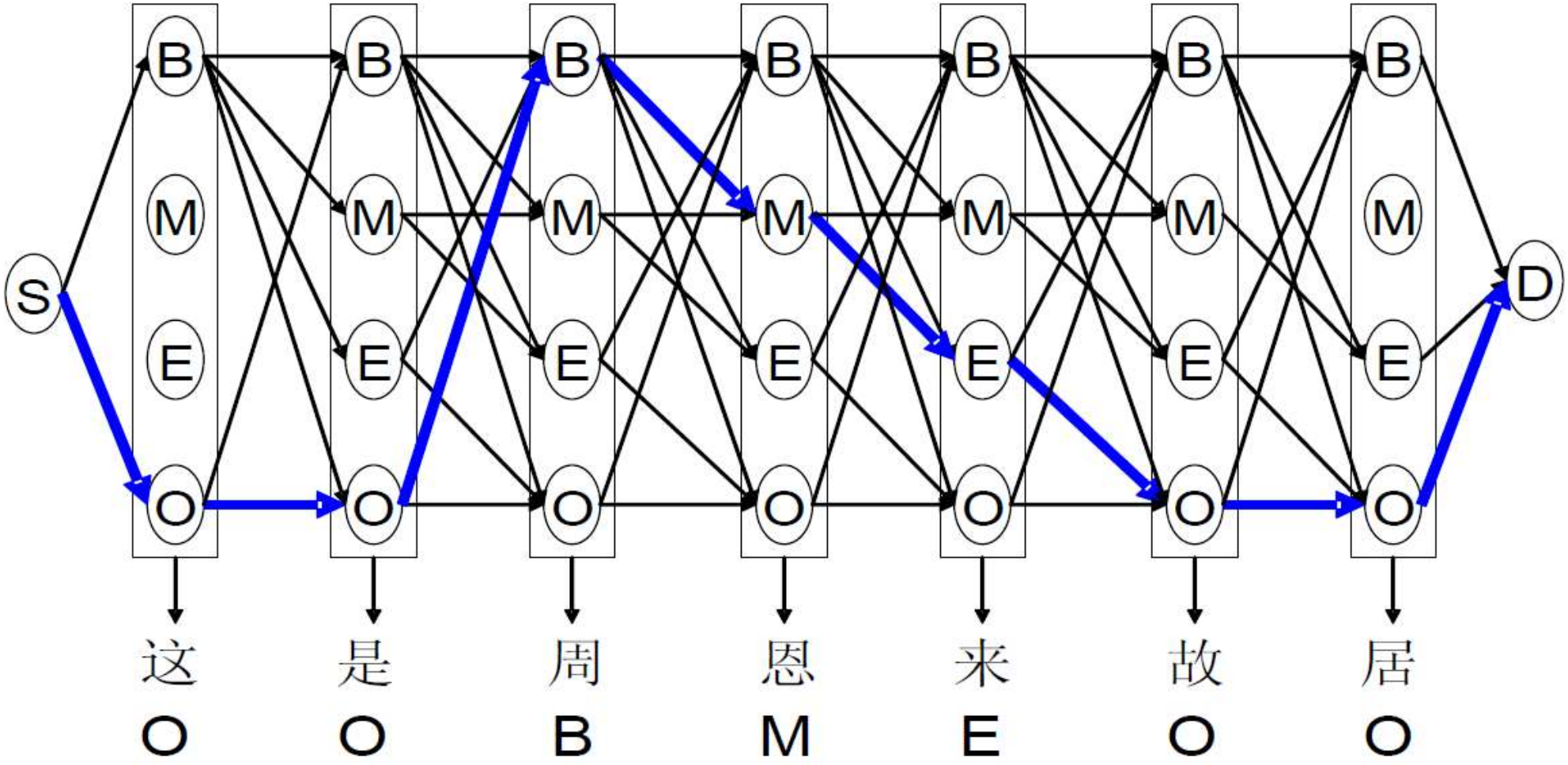
HMM Based Person Name Identification

- Remove unreachable edge:



HMM Based Person Name Identification

- Search for the optimal tagging path:



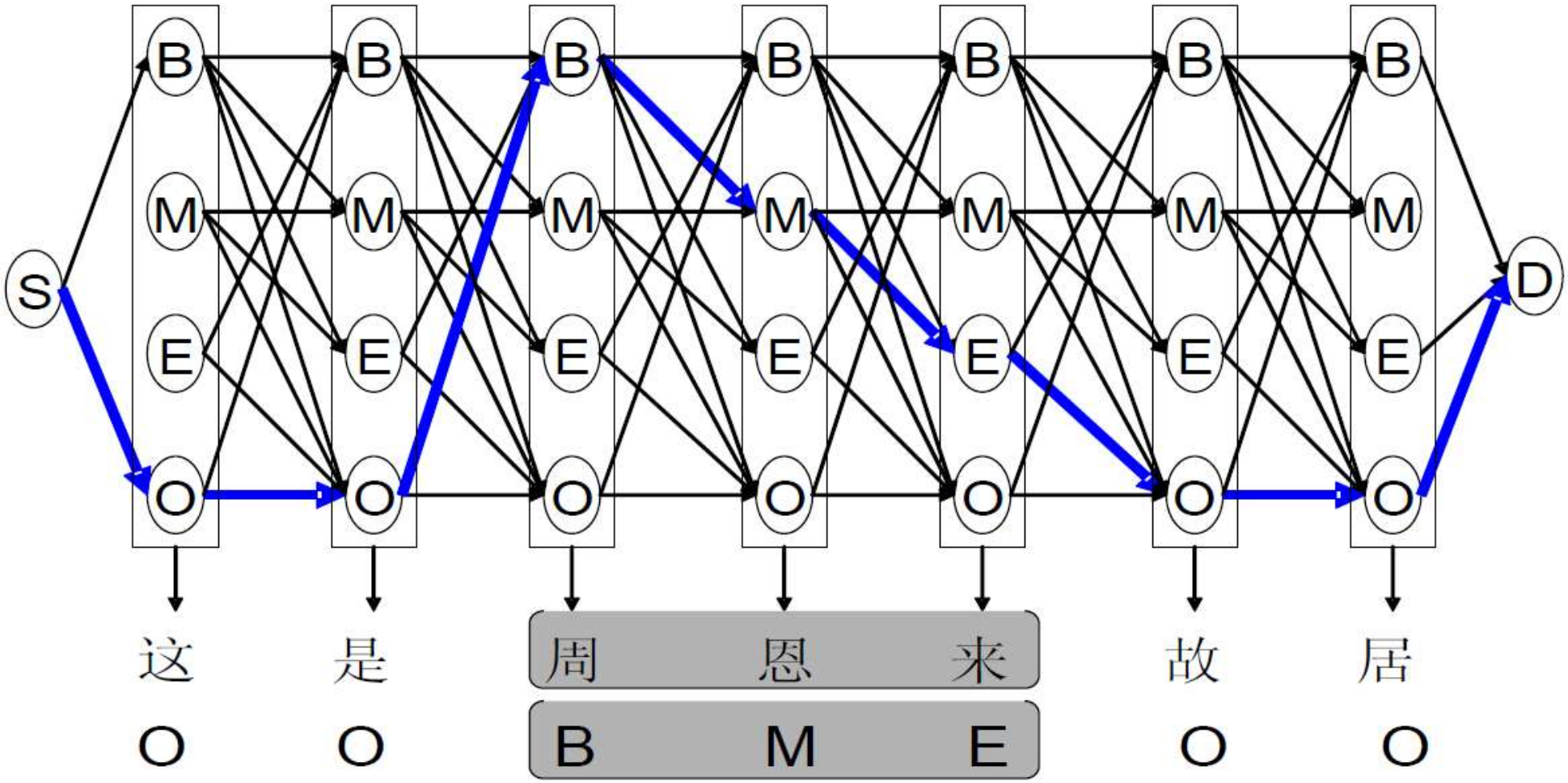
HMM Based Person Name Identification

- Match the following tagging segments in the optimal tagging path:
  - B
  - BE
  - BME
  - BMME
  - ……

人名标记序列

HMM Based Person Name Identification

- Find person name segment in the tagging sequence:



Chinese Organization Name

- Organization name composition are complex
  - The only clear clues are the organization types
- Have no predictable length
- Embody many other types of names
  - 福田区委 李嘉诚慈善基金会
- Words included in organization names are diverse
- Many organization names have abbreviations
  - 微软亚洲研究院

Internal Structure of Chinese Organization Name

- A Chinese organization name is normally composed of two major components
  - The first part is the proper name: Prefix word
    - Unlimited: 光明区委 三一重工集团
  - The second part is the organization type: organization name characteristic words
    - 公司 基金会 大学
  - The second part contains the major clue words guiding the identification of the right boundary
  - Determining the left boundaries becomes most crucial to organization name identification.

Features

- Features for organization name composition
  - Probability of word/character in/out organization name →organization word/character list
  - Famous organization names and their abbreviations
- Features for potential left boundary
  - Location Name 杭州商会
  - Organization Name 哈尔滨工业大学深圳研究生院
  - Special Word: NBA中国球迷会

Constraints For Organization Name Composition

- Some “stop words” such as 失败 and 其它 will never be selected to be part of the name
- Rule-based approaches[Chen et al 2000; Zhang et al 2008]
- Probability-based approaches [Yu et al 2003; Sun et al 2003; Wu et al 2003]
- Role-based tagging approach [Yu et al 2003]
- Web-based approach

Chen KJ, Chen CJ. Knowledge extraction for identification of Chinese organization names.
Yu H, Zhang H, Liu Q. Recognition of Chinese organization name based on role tagging.
Sun J, Zhou M., Gao J. A class-based language model approach to Chinese named entity identification.
Wu Y, Zhao J, Xu B. Chinese named entity recognition combining a statistical model with human knowledge.
Zhang Q Hu G, Yue L. Chinese organization entity recognition and association on web pages.

Chinese Place Name Identification

- Place name dictionary and dictionary look-up
  - The Chinese Place Name Set
    - published by Place Names Committee of China
    - 100,000 place names
  - 中华人民共和国地名词典
  - 中国古今地名大词典
  - 中文地名索引
- About 30% of the places names in a real text cannot be found [Zhang et al 2001]

- Combine corpus statistics and context rules
- The place suffix is a kind of useful features
- Corpus-based approaches:
  - Estimate the likelihood of a character being a part of a Chinese place name
  - Used as the beginning, middle and end character
  - Capture the capability of a character forming for Chinese place names
  - Provide a good means to locate the place name candidates

- Rule-based Candidates Confirmation
  - Example Rule: If two place name candidates are coordinated and one of them has been confirmed as a true place name, then the other should be confirmed as a true place name also
  - E.g. [罗湖区]、[福田区]和光明新区
- Rule-based Candidates Elimination
  - Example Rule: A place name candidate should be eliminated if its preceding word is a title of a person
  - E.g. 王四川 小沈阳

Abbreviation

- The first character of the each word in full name, 哈尔滨工业大学→ 哈工大
- A proper name in an organization full name, select the proper name, 耐克公司 → 耐克
- Location name plus the first character of the other words上海交通大学→ 上海交大
- First character of all the words excluding the organization name characteristic word. 中国交通银行总部→ 交行总部

Abbreviation in Chinese

Given a full-form $F = f_1 f_2 \cdots f_m$ and its corresponding abbreviation $S = s_1 s_2 \cdots s_n$, let $f_i (1 \le i \le m)$ denote a constituent word of the full-form and $s_j (1 \le j \le n)$ denote one component of the relevant short-form, then the above three types of Chinese abbreviations can be formally redefined as follows:

- Reduced Abbreviation: 空军政治部 – 空政

  If $n = m$ and s_i is the corresponding short-form of the constituent word f_i

- Eliminated abbreviation: 清华大学- 清华

  If $n < m$ and $\forall s_j \in F (1 \le j \le n)$

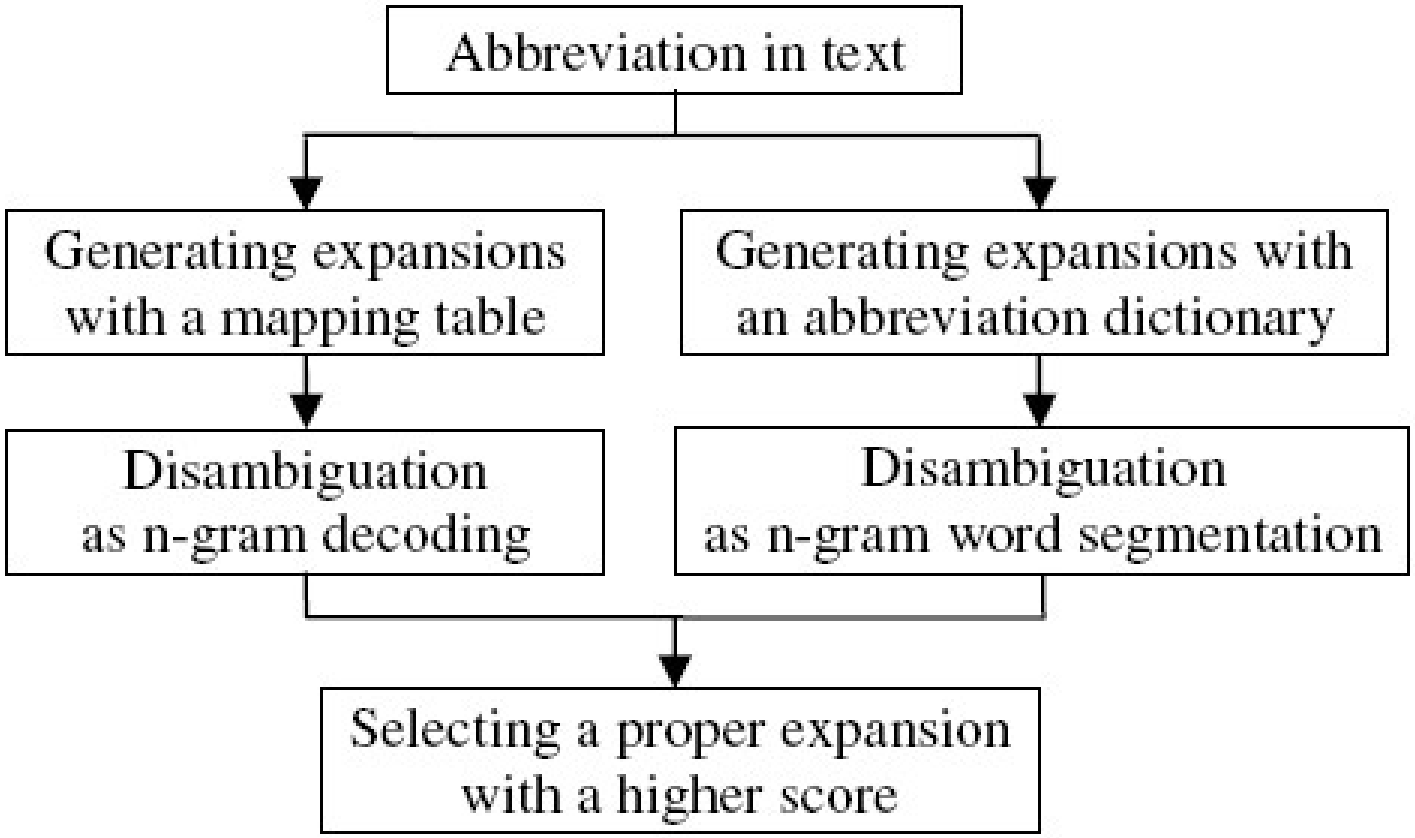
- Generalized abbreviation： 三从四德 八荣八耻

  If $n < m$ and $\exists s_j \notin F (1 \le j \le n)$

Abbreviation Statistics in PKU People's Daily Corpus

Corpus	No. Words	No. Sentences	Reduced Abbr.	Non-Reduced Abbr.	Total Abbr.	No. Sentences with abbr.
Jan.	1.12M	47,288	6,505 62.6%	3,883 37.4%	10,388	6,729 14.2%
Feb.	1.15M	48,095	6,655 59.7%	4,498 40.3%	11,153	7,137 14.8%

Chinese Abbreviation Normalization: Framework



Expanding Reduced Abbreviations

- Generating Expansions with a Mapping Table
  - Assumption: Each character or word in a reduced abbreviation must match a word in its full-form
  - Characters or short-words vs. Full words

Short Words	Full Words	Real Cases
大	大学 大会 大型	北大 十五大 十大工程
联	联邦 联合会	足联 国联

- Disambiguation as N-Gram Decoding
  - Expansion decoding aims to search a combination of long-words as the best path from the lattice that maximizes the n-gram probability

$$\hat{F} = \arg\max_{F} P(LFR) = \arg\max_{F} P(Lf_1 f_2 \cdots f_n R)$$
$$\approx \arg\max_{F} P(f_1 \mid L) \times \prod_{i=1,n} P(f_i \mid f_{i-N+1,i-1}) \times P(R \mid f_{n-N+2,n})$$

$$\hat{F} = \arg\max_{F} P(Lf_1 f_2 \cdots f_n R) \approx \arg\max_{F} \prod_{i=1}^{n+1} P(f_i \mid f_{i-1})$$

Expanding Non-Reduced Abbreviations

- Generating Expansions with a Dictionary of Abbreviations
- Maps each non-reduced abbreviation to a set of full-forms from a abbreviation-expanded corpus

Short Words	Full Words
清华	清华大学
三通	通邮、通商、通航

- Disambiguation as N-Gram Word Segmentation
  - Select a proper expansion from a set of candidates if any
  - Using bigram word segmentation
  - The main idea is: each expansion candidate of a given abbreviation is segmented into a sequence of words using bigram LMs. The one whose segmentation has the maximum score will be identified as expansion output

Abbreviations Expansion Disambiguation

- Final disambiguation for expansion candidates
- Comparing their respective n-gram scores
- The one with a higher score is chosen as the resulting expansion

- Performance 85.5-88.0%
- Problem Left
  - 志愿: 北大 科大 清华
  - 上吊 自砂

Guohong Fu, Kang-Kwong Luke, Guodong Zhou, Ruifeng Xu, Automatic Expansion of Abbreviations in Chinese News Text, in Lecture Notes in Computer Science 4182, Springer, 2006, pp.530-536 (SCI IDS: BFG22)(EI Compendex 20064810275736)

Ref

Automatic Abbreviations for Titles

- Abbreviations are simply the combination of the first characters of all words in the title(example 1)
- Choosing to combine prefixes of certain words in the title (example 2-3)
- Using some external words (example 4)

Id	Abbreviation	Title Words
1	FAWN	A Fast Array of Wimpy Nodes
2	IPR	An Integrated Placement and Routing Algorithm
3	SeRQL	A Second Generation RDF Query Language
4	SIMPLIcity	Semantics-Sensitive Integrated Matching for Picture LIbraries

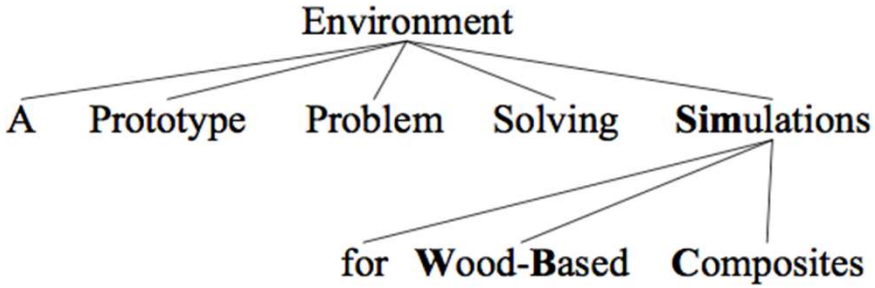
Jianbing Zhang, Yixin Sun, Shujian Huang*, Cam-Tu Nguyen, Xiaoliang Wang, Xinyu Dai, Jiajun Chen, Yang Yu. AGRA: An Analysis-Generation-Ranking Framework for Automatic Abbreviation from Paper Titles. IJCAI 2017

Description Analysis

- Lexical Analysis (Lex)

$$f_{\text{Lex-X}}(w_k) = \begin{cases} 1 & \text{the POS tag of } w_k \text{ is X,} \\ 0 & \text{otherwise.} \end{cases}$$

- Syntactic Analysis (Syn)



- Semantic Analysis (Sem)

$$f_{\text{Syn-i}}(w_k) = \begin{cases} 1 & w_k \text{ is at the } i^{th} \text{ layer of the tree,} \\ 0 & \text{otherwise.} \end{cases}$$

Ref

Candidate Generation

- Scoring the Candidates

$$f_{\text{Char-i}}(p_k, w_k) = \begin{cases} 1 & p_k \text{ uses the } i^{th} \text{ character of } w_k, \\ 0 & \text{otherwise.} \end{cases}$$

- Searching for Candidates
  - Append: Append the next character of current word to the current prefix
  - Jump: Move to the next word in the title
  - Stop: Keep the current abbreviation, with no further extension

Ref

Abbreviation Ranking

- Abbreviations that are, or similar to, existing words are usually good candidates,

- Denote the language model score of abbreviation:

$$s(a) = s_{\text{ABBR}}(a) + w_{\text{LM}} \times f_{\text{LM}}(a) + w_{\text{len}} \times f_{\text{len}}(a)$$

- Performance 50.0-55.0%
- Problem Left
  - Do not consider sophisticated language phenomena

Network Informal Language (NIL)

- Internet used information language
- Observation

Morphological form	No. of unique terms	No. of occur.	Percentage of occur.
Chinese character/ word/phrase	446	24438	55.59
Letter	197	15514	35.29
Number	25	2886	6.56
Mixed form	147	1029	2.34
Other	46	96	0.22
Total	861	43963	100

Xia Y, Wong K F, Gao W. NIL is not nothing: Recognition of Chinese network informal language expressions

- NIL widely distributed in many documents
- Brings more ambiguities in Chinese Word Process
- Anomalous entry 酱紫 细巴细
- Anomalous sense 玉米 钢丝 神马

Anomalous type	No. of unique terms	No. of occur.	Percentage of occur.
Anomalous entry	650	20585	46.82
Anomalous sense	211	23378	53.18
Total	861	43963	100

Senses in NIL

No. of senses	No. of unique terms	No. of occur.	Percentage of occur.	Max. occur.	Avg. occur.
1	613	14514	33.01	5405	24
2	211	22598	51.40	8735	107
3	21	3054	6.95	502	145
4	13	2285	5.20	909	176
5	1	24	0.05	24	24
6	1	413	0.94	413	413
7	1	1075	2.45	1075	1075
Total	861	43963	100	–	–

Creation and Relinquish Rates

- Creation rate

Set	Feb-05	Apr-05	Jun-05	Aug-05	Oct-05	Dec-05	Feb-06
Dec-04	0.0231	0.0531	0.0912	0.1410	0.1640	0.1728	0.1880
Feb-05	–	0.0307	0.0697	0.1207	0.1442	0.1532	0.1688
Apr-05	–	–	0.0402	0.0928	0.1171	0.1264	0.1425
Jun-06	–	–	–	0.0548	0.0801	0.0898	0.1065
Aug-05	–	–	–	–	0.0268	0.0370	0.0547
Oct-05	–	–	–	–	–	0.0105	0.0287
Dec-05	–	–	–	–	–	–	0.0184

- Relinquish rate

Set	Feb-05	Apr-05	Jun-05	Aug-05	Oct-05	Dec-05	Feb-06
Dec-04	0.0259	0.0560	0.0920	0.1265	0.1578	0.1782	0.1878
Feb-05	–	0.0309	0.0679	0.1033	0.1354	0.1563	0.1662
Apr-05	–	–	0.0381	0.0747	0.1078	0.1294	0.1396
Jun-06	–	–	–	0.0380	0.0725	0.0949	0.1055
Aug-05	–	–	–	–	0.0358	0.0592	0.0702
Oct-05	–	–	–	–	–	0.0242	0.0356
Dec-05	–	–	–	–	–	–	0.0117

- Within 12 months,17.28% NILs are created and 17.82% NILs relinquished

Phonetic behavior

Phonetic behaviour	No. of unique terms	No. of occur.	Percentage of occur.
Created using phonetic clue	802	42767	97.28
Created using no phonetic clue	59	1196	2.72
Total	861	43963	100

NIL Normalization - I

- Character-based Source Channel Model
  - Given an input chat text string, T
  - Source Channel Model aims to find the most probable translation character string

$$C = \arg\max_{C} p(C|T) = \arg\max_{C} p(T|C)p(C)$$

- Problems:
  - Data sparseness problem in NIL corpus is serious
  - Training effectiveness is poor again due to the dynamic nature of the chat language

NIL Normalization - II

- Phonetic mapping
  - Phonetic mapping connects two characters via phonetic transcription, i.e. Chinese pinyin
- The phonetic mapping probability is calculated by combining phonetic similarity and character frequency in the standard language

$$\Pr_{pm}(c|\bar{c}) = \frac{(fr_{slc}(\bar{c}) \times ps(c,\bar{c}))}{\sum_i (fr_{slc}(c_i) \times ps(c,c_i))}$$

$$ps(c,c) = Sim(py(c), py(c)) = Sim(initial(py(c)), initial(py(c))) \times Sim(final(py(c)), final(py(c))).$$

- Phonetic mapping probability
  - Chinese to Chinese

(1) 偶 $\xrightarrow{(wo,ou,0.685)}$ 我: 偶 (even; *ou3*) replaces我 (me, *wo3*) with $p = 0.685$.

(2) 介 $\xrightarrow{(zhe,jie,0.56)}$ 这: 介 (interrupt; *jie4*) replaces 这 (this; *zhe4*) with $p = 0.560$.

(3) 素 $\xrightarrow{(shi,su,0.491)}$ 是: 素 (white, *su4*) replaces 是 (is, *shi4*) with $p = 0.491$.

(4) 银 $\xrightarrow{(ren,yin,0.457)}$ 人: 银 (silver, *yin2*) replaces 人 (human, *ren2*) with $p = 0.457$.

(5) 米 $\xrightarrow{(mei,mi,0.452)}$ 没: 米 (rice, mi3) replaces 没 (have not, *mei2*) with $p = 0.452$.

– Letter to Chinese

(6) J $\xrightarrow{(jie,ji,0.671)}$ 姐: *J* replaces 姐 (older sister; *jie3*) with $p = 0.671$.

(7) M $\xrightarrow{(mei,mi,0.593)}$ 妹: *M* replaces 妹 (younger sister; *mei4*) with $p = 0.593$.

(8) S $\xrightarrow{(si,si,0.587)}$ 死: *S* replaces 死 (die; *si3*) with $p = 0.587$.

(9) T $\xrightarrow{(ti,ti,0.465)}$ 踢: *T* replaces 踢 (kick; *ti1*) with $p = 0.465$.

(10) K $\xrightarrow{(kuai,ki,0.447)}$ 快: *K* replaces 快 (quick; *kuai4*) with $p = 0.447$.

– Number to Chinese

(11) 9 $\xrightarrow{(jiu,jiu,0.541)}$ 酒: 9 replaces 酒 (wine; *jiu3*) with $p = 0.541$.

(12) 8 $\xrightarrow{(bu,ba,0.519)}$ 不: 8 replaces 不 (no; *bu4*) with $p = 0.519$.

(13) 7 $\xrightarrow{(chi,qi,0.454)}$ 吃: 7 replaces 吃 (eat; *chi1*) with $p = 0.454$.

(14) 4 $\xrightarrow{(si,si,0.449)}$ 死: 4 replaces 死 (die; *si3*) with $p = 0.449$.

(15) 5 $\xrightarrow{(wu,wu,0.297)}$ 呜: 5 replaces 呜 (crying sound; *wu1*) with $p = 0.297$.

- Extend the source channel model by inserting a phonetic mapping model

$$\hat{C} \approx \arg\max_{M.C} p(T, M|C)p(C) = \arg\max_{M.C} p(T|M, C)p(T|M, C)p(M|C)p(C)$$

- Three components
  - chat term normalization observation model p(T|M,C) $p(T|M, C) = \prod_i p(t_i|m_i, c_i)$
  - phonetic mapping model p(M|C)

$$p(M|C) = \prod_i p(m_i|c_i) = \prod_i \Pr^*_{pm}(t_i|c_i)$$

  - language model p(C)
- 78%-88% normalization accuracy

The Next Lecture

- Lecture 7

  Word Meaning