Domain Adaptation for CRF-based Chinese Word Segmentation using Free Annotations

Yijia Liu[†‡], Yue Zhang[†], Wanxiang Che[‡], Ting Liu[‡], Fan Wu[†]

[†] Singapore University of Technology and Design

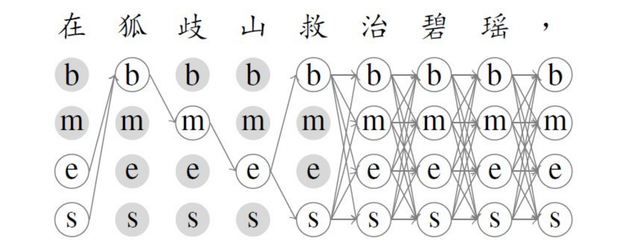
[‡] Research Center for Social Computing and Information Retrieval

Harbin Institute of Technology, China

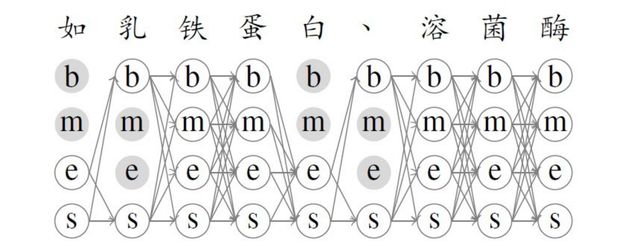
Introduction

- Statistical Chinese word segmentation gains high accuracies on newswire
- Performance drops when testing domain switch from newswire to blog, computer forums and Internet literature
- There are *free* data which contain limited but useful segmentation information over the internet.
  - Lexicon
  - Wikipedia
- Contributions
  - adopting freely available data
  - different sources of free data are represented as partial segmentation.
  - a variant of CRF is used to model partially annotated data.
  - https://github.com/ExpResults/partial-crfsuite

Example of Partial Data



(a) "在 (at) 狐岐山 (Huqi Mountain) 救治 (save) 碧瑶 (Biyao)", where "狐岐山" matches a lexicon word.



(b) "如 (e.g.) 乳铁蛋白 (lysozyme) 、 溶菌酶 (lactoferrin)", where "乳铁蛋白" is a hyperlink.

Obtaining Partial Data

- Free lexicon
  - Forward maximum matching scheme is used to find subsequence that matches lexicon in the unlabeled data.
  - If a lexicon entry is matched in sentence, the subsequence is tagged with the corresponding tags, and its surrounding characters are also constrained to a small set of tags.
  - (a) in left figure illustrates this method.
- Free natural annotation
  - *Natural annotation* refers to word boundaries that can be inferred from URLs, fonts or colors on web pages, also result in partially annotated sentences.
  - Problems:
    - incompatibility of segmentation standards between the annotated training data and Wikipedia.

People's Daily	看到(saw) 海南(Hainan) 旅游业(tourist industry) 充满(full) 希望(hope) saw tourist industry in Hainan is full of hope
Wikipedia	主要(mainly) 是(is) 旅游(tourist) 业(industry) 和(and) 软件(software) 产业(industry) mainly is tourist industry and software industry

    - intrinsic ambiguity of segmentation

Literature	《说文解字(Shuo Wen Jie Zi, a book) 段(segmented) 注(annotated)》 the segmented and annotated version of Shuo Wen Jie Zi
Computer	每条(each) 记录(record) 被(is) 分隔(splitted) 为(into) 字段(fields) each record is splitted into several fields

    - selection on natural annotated data is needed.
      - Any URL-tagged entry in a Wikipedia sentence matches the target domain data, the sentence is selected for training.

Experimental Results

- Free lexicon
  - CTB -> Zhuxian (A Chinese Internet novel)
    - use the lexicon from Zhang et al. (2014), 159 entries for entity names.
  - People's Daily Corpus -> Medicine and Computer (*sighan 2010 bakeoff data*)
    - use wiki page titles
  - The model with lexicon feature (Sun and Xu, 2010) are also used as comparison.

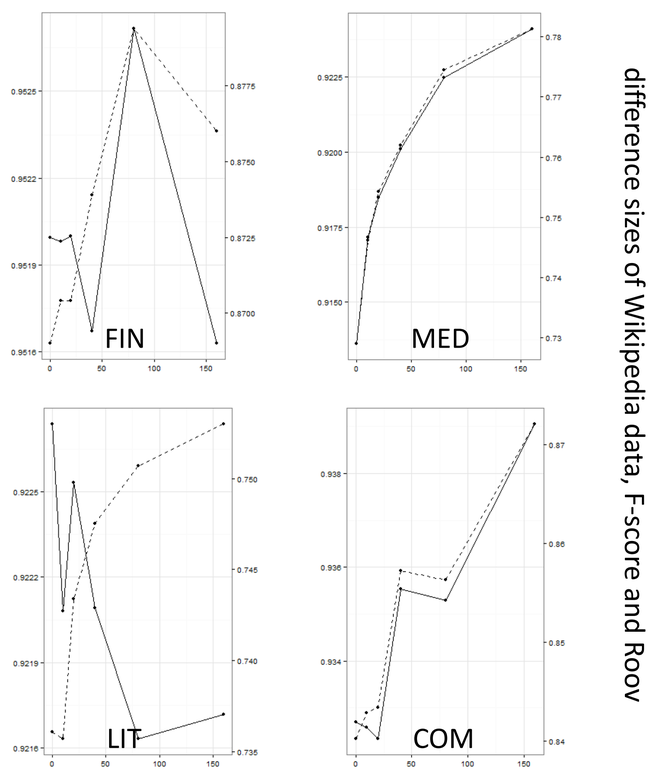
System	ZX		Medicine		Computer	
	F	Roov	F	Roov	F	Roov
Baseline	87.50	73.65	91.36	72.96	93.16	84.02
Baseline + Lexicon Feature	90.36	80.69	91.60	74.39	93.14	84.27
Baseline + PA(lex)	**90.63**	**84.88**	**91.68**	**74.99**	**93.47**	**85.63**
Zhang et al. (2014)	88.34					

  - Our method give better result than the lexicon feature method and Zhang et al.
    - combining a lexicon with unannotated sentences is a better option than using the lexicon directly
- Free natural annotation
  - People's Daily Corpus -> Finance, Medicine, Literature, Computer
    - using Wikipedia page as free natural annotation, perform selection against randomly selected data

Method	Finance		Medicine		Literature		Computer		Avg-F
	F	Roov	F	Roov	F	Roov	F	Roov	
Baseline	95.20	86.90	91.36	72.90	92.27	73.61	93.16	83.48	93.00
Baseline+PA(random)	95.16	87.60	92.41	78.13	92.17	75.30	93.91	83.48	93.41
Baseline+PA(selected)	**95.54**	**88.53**	92.47	**78.28**	92.49	**76.84**	**93.93**	**87.53**	**93.61**
Jiang et. al (2013)	93.16		**93.34**		**93.53**		91.19		92.80

    - The model incorporating selected data achieves better performance compared to the model with randomly sampled data
  - Combining lexicion and natural annotation

	Medicine	Computer
	F	F
Baseline	91.36	93.16
Baseline + PA(lex)	91.68	93.47
Baseline + PA(natural)	92.47	93.93
Baseline + PA(lex+ natural)	**92.63**	**94..07**



Analysis

1. Natural annotation on Wikipedia data contributes to the recognition of OOV words on domain adaptation;
2. target domains with more OOV words benefit more from Wikipedia data.
3. along with the positive effect on OOV recognition, Wikipedia data can also introduce noise, and hence data selection can be useful

Modelling the partially annotated data

- CRF-based method by modeling the marginal probability over partially annotated data (Tsuboi et al. 2008).
  - Fully and partially annotated data are modeled together.
- For each character's possible labels:
  - $\boldsymbol{L} = (L_1, L_2, \ldots, L_T)$
- $\boldsymbol{Y_L}$ be the set of all possible label sequences.
- Probability:
  - $p(Y_L|x) = \frac{1}{Z}\sum_{y \in Y_L} \exp \sum_{t=1}^{T} \sum_k \lambda_k f_k(y_t, y_{t-1}, x)$
- Likelihood:
  - let $Z_{Y_l} = \sum_{y \in Y_L} \exp \sum_{t=1}^{T} \sum_k \lambda_k f_k(y_t, y_{t-1}, x)$
  - $\mathcal{L}_{Y_l} = \sum_{n=1}^{N} (\log Z_Y - \log Z)$
- A modification on forward-backward algorithm is used to calculate the likelihood and its gradient.
- L-BFGS is used to learning parameter from data.

Acknowledge

- National Key Basic Research Program of China via grant 2014CB340503
- National Natural Science Foundation of China (NSFC) via grant 61133012 and 61370164
- The Singapore Ministry of Education (MOE) AcRF Tier 2 grant T2MOE201301
- SRG ISTD 2012 038 from Singapore University of Technology and Design