Person Name Disambiguation based on Topic Model

Jiashen Sun, Tianmin Wang and Li Li
Center of Intelligence Science and Technology
Beijing University of Posts and Telecommunications
b.bigart911@gmail.com,
tianmin180@sina.com, wbg111@126.com

Xing Wu
School of Computer
Beijing University of Posts and Telecommunications
wuxing-6@163.com

Abstract

In this paper we describe our participation in the SIGHAN 2010 Task-3 (Person Name Disambiguation) and detail our approaches. Person Name Disambiguation is typically viewed as an unsupervised clustering problem where the aim is to partition a name's contexts into different clusters, each representing a real world people. The key point of Clustering is the similarity measure of context, which depends upon the features selection and representation. Two clustering algorithms, HAC and DBSCAN, are investigated in our system. The experiments show that the topic features learned by LDA outperforms token features and more robust.

1 Introduction

Most current web searches relate to person names. A study of the query log of the AllTheWeb and Altavista search sites gives an idea of the relevance of the people search task: 11-17% of the queries were composed of a person name with additional terms and 4% were identified simply as person names (Spink et al., 2004).

However, there is a high level of ambiguity where multiple individuals share the same name and thus the harvesting and the retrieval of relevant information becomes more difficult. This ambiguity has recently become an active research topic and, simultaneously, a relevant application domain for Web search services. Zoominfo.com, Spock.com and 123people.com are examples of sites which perform web people search, although with limited disambiguation capabilities (Artiles et al., 2009).

This issue directed current researchers towards the definition of a new task called Web People Search (WePS) or Personal Name Disambiguation (PND). The key assumption underlying the task is that the context surrounding an ambiguous person name is indicative of its ascription. The goal of the clustering task was to group web pages containing the target person's name, so that pages referring to the same individual are assigned to the same cluster. For this purpose a large dataset was collected and manually annotated.

Moreover, because of the ambiguity in word segmentation in Chinese, person name detection is necessary, which is subtask of Named Entity Recognition (NER). NER is one of difficulties of the study of natural language processing, of which the main task is to identify person names, place names, organization names, number, time words, money and other entities. The main difficulties of Chinese person name entity recognition are embodied in the following points: 1) the diversity of names form; 2) the Chinese character within names form words with each; 3) names and their context form words; 4) translation of foreign names require special considerations.

In this paper we describe our system and approach in the SIGHAN 2010 task-3 (Person Name Disambiguation). A novel Bayesian approach is adopt in our system, which formalizes the disambiguation problem in a generative model. For each ambiguous name we first draw a distribution over person, and then generate context words according to this distribution. It is thus assumed that different persons will correspond to distinct lexical

distributions. In this framework, Person Name Disambiguation postulates that the observed data (contexts) are explicitly intended to communicate a latent topic distribution corresponding to real world people.

The remainder of this paper is structured as follows. We first present an overview of related work (Section 2) and then describe our system which consists of NER and clustering in more details (Sections 3 and 4). Section 5 describes the resources and evaluation results in our experiments. We discuss our results and conclude our work in Section 6.

2 Related Work

The most commonly used feature is the bag of words in local or global context of the ambiguous name (Ikeda et al., 2009; Romano et al., 2009). Because the given corpus is often not large enough to learn the realistic probabilities or weights for those features, traditional algorithm such as vector-based techniques used in large-scale text will lead to data sparseness.

In recent years, more and more important studies have attempted to overcome the problem to get a better (semantic) similarity measures. A lot of features such as syntactic chunks, named entities, dependency parses, semantic role labels, etc., were employed. However, these features need many NLP preprocessing (Chen, 2009). Many studies show that they can achieve state-of-the-art performances only with lightweight features. Pedersen et al. (2005) present SenseClusters which represents the instances to be clustered using second order co–occurrence vectors. Kozareva (2008) focuses on the resolution of the web people search problem through the integration of domain information, which can represent relationship between contexts and is learned from WordNet. PoBOC clustering (Cleuziou et al., 2004) is used which builds a weighted graph with weights being the similarity among the objects.

Another way is to utilize universal data repositories as external knowledge sources (Rao et al., 2007; Kalmar and Blume, 2007; Pedersen and Kulkarni; 2007) in order to give more realistic frequency for a proper name or measure whether a bigram is a collocation.

Phan et al. (2008) presents a general framework for building classifiers that deal with short and sparse text and Web segments by making the most of hidden topics discovered from large-scale data collections. Samuel Brody et al. (2009) adopt a novel Bayesian approach and formalize the word sense induction problem in a generative model.

Previous work using the WePS1 (Artiles et al., 2007) or WePS2 data set (Artiles et al., 2009) shows that standard document clustering methods can deliver excellent performance if similarity measure is enough good to represent relationship of context.

The study in Chinese PND is still in its infancy. Person Name detection is often necessary in Chinese. At present, the main technology of person name recognition is used statistical models, and the hybrid approach. Liu et al. (2000) designed a Chinese person name recognition system based on statistical methods, using samples of names from the text corpus and the real amount of statistical data to improve the system performance, while the shortcoming is that samples of name database are too small, resulting in low recall. Li et al. (2006) use the combination of the boundary templates and local statistics to recognize Chinese person name, the recognition process is to use the boundary with the frequency of template to identify potential names, and to recognize the results spread to the entire article in order to recall missing names caused by sparse data.

3 Person Name Recognition

In this section, we focus on Conditional Random Fields (CRFs) algorithm to establish the appropriate language model. Given of the input text, we may detect the potential person names in the text fragments, and then take various features into account to recognize of Chinese person names.

Conditional Random Fields as a sequence learning method has been successfully applied in many NLP tasks. More details of the its principle can be referred in (Lafferty, McCallum, and Pereira, 2001; Wallach, 2004). We here will focus on how to apply CRFs in our person name recognition task.

3.1 CRFs-based name recognition

CRFs is used to get potential names as the first stage name recognition outcome. To avoid the

interference that caused by word segmentation errors, we use single Chinese character information rather than word as discriminative features for CRFs learning model.

We use BIEO label strategy to transfer the name recognition as a sequence learning task. The label set includes: B-Nr (Begin, the initial character of name), I-Nr (In, the middle character of name), E-Nr(End, the end character of name) and O (Other, other characters that aren't name).

3.2 Rule-based Correction

After labeling the potential names by CRFs model, we apply a set of rules to boost recognition result, which has been proved to be the key to improve Chinese name recognition.

The error of the potential names outcome by CRFs model is mainly divided into the following categories: the initial character of name is not recognized, the middle character of name is not recognized, the end character of name is not recognized, and their combinations of those three errors. The other two extreme errors, including non-name recognition for the anchor name, and the name is not recognized as potential names.

In the stage of rule-based correction, we first conduct word segmentation for the text. The segmentation process is also realized with the method of CRFs, without using dictionaries and other external knowledge. The detailed description is beyond this paper, which can be accessible in the paper (Lafferty, McCallum, and Pereira, 2001). The only thing we should note is that part of the error in such segmentation result obtained in this way can be corrected through the introduction of an external dictionary.

For each potential name, and we examine it from the following two aspects:

1) It is reasonable to use the word in a person name, including checking the surname and the character used in names;

2) The left and right borders are correct. Check the left and right sides of the cutting unit can be added to the names, including the words used before names, the words used behind names and the surname and character used in names.

4 Clustering

4.1 Features

The clustering features we used can be divided into two types, one is token features, including word (after stop-word removal), uni-character and bi-character, the other is topic features, which is topic-based distribution of global or window context learned by LDA (Latent Dirichlet Allocation) model.

4.1.1 Token-based Features

Simple token-based features are used in almost every disambiguation system. Here, we extract three kinds of tokens: words, uni-char and bi-char occurring in a given document.

Then, each token in each feature vector is weighed by using a tf-idf weighting and entropy weighting schemes defined as follows.

tf-idf weighting:

$$a_{ik} = f_{ik} \bullet \log(\frac{N}{n_i})$$

entropy weighting:

$$a_{ik} = \log(f_{ik} + 1.0) \bullet \left(1 + \frac{1}{\log(N)} \sum_{j=1}^{N} \left[\frac{f_{ij}}{n_i} \log(\frac{f_{ij}}{n_i}) \right] \right)$$

where f_{ik} is the frequency of term i in document k, N is the number of document in corpus, n_i is the frequency of term i in corpus. So,

$$\frac{1}{\log(N)} \sum_{j=1}^{N} \left[\frac{f_{ij}}{n_i} \log(\frac{f_{ij}}{n_i}) \right]$$

is the average uncertainty or entropy of term i. Entropy weighting is based on information theoretic ideas and is the most sophisticated weighting scheme.

4.1.2 Features Selection

In this Section, we give a brief introduction on two effective unsupervised feature selection methods, DF and global tf-idf.

DF (Document frequency) is the number of documents in which a term occurs in a dataset. It is the simplest criterion for term selection and easily scales to a large dataset with linear computation complexity. It is a simple but effective feature selection method for text categorization (Yang & Pedersen, 1997).

We introduce a new feature selection method called "global tf-idf" that takes the term weight into account. Because DF assumes that each term is of same importance in different documents, it is easily biased by those common terms which have high document frequency but uniform distribution over different classes. Global tf-idf is proposed to deal with this problem:

$$g_i = \sum_{k=1}^{N} tfidf_{ik}$$

4.1.3 Latent Dirichlet Allocation (LDA)

Our work is related to Latent Dirichlet Allocation (LDA, Blei et al. 2003), a probabilistic generative model of text generation. LDA models each document using a mixture over K topics, which are in turn characterized as distributions over words. The main motivation is that the task, fail to achieve high accuracy due to the data sparseness.

LDA is a generative graphical model as shown in Figure 1. It can be used to model and discover underlying topic structures of any kind of discrete data in which text is a typical example. LDA was developed based on an assumption of document generation process depicted in both Figure 1 and Table 1.

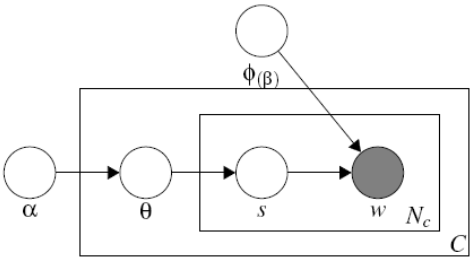

Figure 1 Generation Process for LDA

4.1.4 LDA Estimation with Gibbs Sampling

Estimating parameters for LDA by directly and exactly maximizing the likelihood of the whole data collection is intractable. The solution to this is to use approximate estimation methods like Gibbs Sampling (Griffiths and Steyvers, 2004).

Here, we only show the most important formula that is used for topic sampling for words. After finishing Gibbs Sampling, two matrices Φ and Θ are computed as follows.

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{v=1}^{V} n_k^{(v)} + \beta_v}$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{j=1}^{K} n_m^{(j)} + \alpha_j}$$

where Θ is the latent topic distribution corresponding to real world people.

Table 1: Generation process for LDA

for all topics $k \in [1, K]$ **do**
  sample mixture components $\vec{\varphi}_k \sim Dir(\vec{\beta})$
end for
for all documents $m \in [1, M]$ **do**
  sample mixture proportion $\vec{\vartheta}_m \sim Dir(\vec{\alpha})$
  sample document length $N_m \sim Poiss(\xi)$
  for all words $n \in [1, N_m]$ **do**
    sample topic index $z_{m,n} \sim Mult(\vec{\vartheta}_m)$
    sample term for word $w_{m,n} \sim Mult(\vec{\varphi}_{z_{m,n}})$
  end for
end for

Parameters and variables:
- M: the total number of documents
- K: the number of (hidden/latent) topics
- V: vocabulary size
- $\vec{\alpha}$, $\vec{\beta}$: Dirichlet parameters
- $\vec{\vartheta}_m$: topic distribution for document m
- $\Theta = \{\vec{\vartheta}_m\}_{m=1}^{M}$: a $M \times K$ matrix
- $\vec{\varphi}_k$: word distribution for topic k
- $\Phi = \{\vec{\varphi}_k\}_{k=1}^{K}$: a $K \times V$ matrix
- N_m: the length of document m
- $z_{m,n}$: topic index of nth word in document m
- $w_{m,n}$: a particular word for word placeholder [m, n]

4.1.5 Topic-based Features

Through the observation for the given corpus, many key information, like occupation, affiliation, mentor, location, and so on, in many cases, around the target name. So, both local and global context are choose to doing topic analysis. Finally, the latent topic distributions are topic-based representation of context.

4.2 Clustering

Our system trusts the result of Person Name detection absolutely, so contexts need to do clustering only if they refer to persons with the same name. We experimented with two different classical clustering methods: HAC and DBSCAN.

4.2.1 HAC

At the heart of hierarchical clustering lies the definition of similarity between clusters, which based on similarity between individual

documents. In my system, a linear combination of similarity based on both local and global context is employed:

$$sim = \alpha \cdot sim_{global} + (1-\alpha) sim_{local}$$

where, the general similarity between two features-vector of documents di and dj is defined as the cosine similarity:

$$sim(d_i, d_j) = \frac{d_i \bullet d_j}{|d_i||d_j|}$$

We will now refine this algorithm for the different similarity measures of single-link, complete-link, group-average and centroid clustering when clustering two smaller clusters together. In our approach we used an overall similarity stopping threshold.

4.2.2 DBSCAN

In this section, we present the algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) (Ester et al., 1996) (Table 2) which is designed to discover the clusters and the noise in a spatial database.

Table 2 Algorithm of DBSCAN

Arbitrary select a point p
Retrieve all points density-reachable from p wrt Eps and MinPts.
If p is a core point, a cluster is formed.
If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
Continue the process until all of the points

5 Experiments and Results Analysis

We run all experiments on SIGHAN 2010 training and test corpus.

5.1 Preprocessing and Person Name Recognition

Firstly, a word segmentation tool based on CRF is used in each document. Then, person name recognition is processing. The training data for word segmentation and PNR is People's Daily in January, 1998 and the whole 2000, respectively.

5.2 Feature Space

Our experiments used five types of feature (uni-char, bi-char, word and topic in local and global), two feature weighting methods (tf-idf and entropy) and two feature selection methods (DF and global tf-idf).

5.3 Model Selection in LDA

Our model is conditioned on the Dirichlet hyperparameters α and β, the number of topic K and iterations. The value for the α was set to 0.2, which was optimized in tuning experiment used training datasets. The β was set to 0.1, which is often considered optimal in LDA-related models (Griffiths and Steyvers, 2004). The K was set to 200. The Gibbs sampler was run for 1,000 iterations.

5.4 Clustering Results and Analysis

Since the parameter setting for the clustering system is very important, we focus only on the B-cubed scoring (Artiles et al., 2009), and acquire an overall optimal fixed stop-threshold from the training data, and then use it in test data. In this section, we report our results evaluated by the clustering scoring provided by SIGNAN 2010 evaluation, which includes both the B-cubed scoring and the purity-based scoring.

Table 3 and 4 demonstrate the performance (F scores) of our system in different features representation and clustering for the training data of the SIGNAN 2010. In Table 3, the numbers in parentheses are MinPts and Eps respectively, and stop-threshold in Table 4. As shown in Table 3, DBSCAN isn't suitable for this task, and the results are very sensitive to parameters. So we didn't submit DBSCAN-based results.

Table 4 shows that the best averaged F-scores for PND are based on topic model, which meet our initial assumptions, and result based on merging local and global information is a bit better than both local and global information independently. Also, the results based on topic model are the most robust because the F-score of variation is slightly with stop-threshold changing. Conversely, the results based on token are not like this. As the performance of segmentation is not very satisfactory, results based on word are worst, even worse than uni-char-based. In

addition, it is found that global tf-idf is better than DF, which is the simplest unsupervised feature selection method. Entropy weighting is more effective than tf-idf weighting.

Table 5 shows that the evaluation results in test data on SIGHAN 2010, and the last two lines are results in diagnosis test. We are in fifth place. The evaluation results (F-score) of Person Name Recognition in training data is 0.965.

Features	FS	Weighting	B-Cubed			P-IP		
			precision	recall	F	P	IP	F
word (0.19)	DF	tf-idf	79.05	79.68	76.49	83.25	85.84	82.72
word (0.2)			80.99	75.72	75.54	84.67	83.08	82.2
word (0.3)		entropy	78.8	80.71	77.42	83.13	86.62	83.58
word (0.25)	global tf-idf	tf-idf	80.79	83.1	80.53	84.88	88.32	85.79
word (0.23)			79.45	84.49	79.66	83.76	89.25	85.08
uni-char (0.43)	DF	tf-idf	76.47	85.46	78.77	81.7	90.05	84.45
uni-char (0.5)			82.34	75.97	77	86.11	83.54	83.78
uni-char (0.48)			80.42	79.44	78.01	84.53	86.17	84.26
bi-char (0.35)			88.3	67.75	75.34	89.96	77.38	82.44
bi-char (0.315)			81.84	81.58	80.54	85.72	87.17	85.8
local topic (0.6)			78.76	86.8	80.63	83.27	91.16	85.88
global topic (0.4)			77.92	88.72	81.04	82.67	92.64	86.26
global topic (0.7)			80.54	88.43	83.55	84.76	92.55	88.02
merged topic (0.63)			81.39	87.82	83.88	85.42	91.94	88.21

Table 3 Performance of HAC

	B-Cubed			P-IP		
MinPts and Eps	precision	recall	F	P	IP	F
2 0.9	64.15	95.84	74.19	71.95	97.36	80.97
2 0.4	71.34	62.25	63.95	76.56	71.94	72.59
3 0.9	64.15	95.88	74.2	71.95	97.37	80.97
6 0.95	64.12	96.55	74.44	71.92	97.79	81.12

Table 4 Performance of DBSCAN

B-Cubed			P-IP		
precision	recall	F	P	IP	F
80.33	94.52	85.79	85.1	96.46	89.77
80.56	92.56	85.29	85.34	95.19	89.5
80.43	95.41	86.18	85.07	97.06	89.96
80.82	93.41	85.77	85.62	95.76	89.91

Table 5 Evaluation Results in test data

6 Discussion and Future Work

In this paper, we present implementation of our systems for SIGHAN-2010 PND bekeoff,.The experiments show that the topic features learned by LDA outperform token features and exhibit good robustness.

However, in our system, only given data is exploited. We are going to collect a very large external data as universal dataset to train topic model, and then do clustering on both a small set of training data and a rich set of hidden topics discovered from universal dataset. The universal dataset can be snippets returned by search engine or Wikipedia queried by target name and some keywords, and so on.

We built our PDN system on the result of person name recognition. However, it is not appropriate to toally trust the result of Person Name detection. So an algorithm that can correct NER mistakes should be investigated in future work..

Moreover, Cluster Ensemble system can ensure the result to be more robust and accurate accordingly, which is another direction of future work..

Acknowledgments

This research has been partially supported by the National Science Foundation of China (NO. NSFC90920006). We also thank Xiaojie Wang, Caixia Yuan and Huixing Jiang for useful discussion of this work.

References

Spink, B. Jansen, and J. Pedersen. 2004. Searching for people on web search engines. *Journal of Documentation*, 60:266 -278.

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2009. Weps 2 evaluation campaign: overview of the web people search clustering task. In *WePS 2 Evaluation Workshop. WWW Conference*.

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.ACL.

M. Ikeda, S. Ono, I. Sato, M. Yoshida, and H. Nakagawa. 2009. Person name disambiguation on the web by twostage clustering. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference.*

L. Romano, K. Buza, C. Giuliano, and L. Schmidt-Thieme. 2009. Person name disambiguation on the web by twostage clustering. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference.*

Y. Chen, S. Y. M. Lee, and C.-R. Huang. 2009. Polyuhk: A robust information extraction system for web personal names. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference.*

Z. Kozareva, R. Moraliyski, and G. Dias. 2008. Web people search with domain ranking. In *TSD '08: Proceedings of the 11th international conference on Text, Speech and Dialogue,* 133-140, Berlin, Heidelberg.

Pedersen, Ted, Amruta Purandare, and Anagha Kulkarni. 2005. Name Discrimination by Clustering Similar Contexts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico.*

G. Cleuziou, L. Martin, and C. Vrain. 2004. Poboc: an overlapping clustering algorithm. application to rule-based classification and textual data, 440-444.

Kalmar, Paul and Matthias Blume. 2007. FICO: Web Person Disambiguation Via Weighted Similarity of Entity Contexts. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).*ACL.

Rao, Delip, Nikesh Garera and David Yarowsky. 2007. JHU1 : An Unsupervised Approach to Person Name Disambiguation using Web Snippets. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).*ACL.

Pedersen, Ted and Anagha Kulkarni. 2007. Unsupervised Discrimination of Person Names in Web Contexts. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico.*

Phan, X., Nguyen, L. and Horiguchi. 2008. Learning to Classify Short and Sparse Test & Web with Hidden Topics from large-scale

Data collection. In *Proceedings of 17th International World Wide Web Conference. (Beijing, China, April 21-25, 2008)*. ACM Press, New York, NY, 91-100.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 103-111.

Sun et al. 1995. Identifying Chinese Names in Unrestricted Texts (in Chinese). In *Journal of Chinese Information Processing*, 9(2):16-27.

Liu et al. 2000. Statistical Chinese Person Names Identification (in Chinese). In *Journal of Chinese Information Processing*, 14(3):16-24.

Huang et al. 2001. Identification of Chinese Names Based on Statistics (in Chinese). In *Journal of Chinese Information Processing*, 15(2):31-37.

Li et al. 2006. Chinese Name Recognition Based on Boundary Templates and Local Frequency (in Chinese). In *Journal of Chinese Information Processing*, 20(5):44-50.

Mao et al. 2007. Recognizing Chinese Person Names Based on Hybrid Models (in Chinese). In *Journal of Chinese Information Processing*, 21(2):22-27.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML-01*, 282-289, 2001.

Wallach, Hanna. 2004. Conditional random fields: An introduction. Technical report, University of Pennsylvania, Department of Computer and Information Science.

Yang, Y. and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, US, 1997)*, 412–420.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press. 226-231.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. In *J. Machine Learn. Res. 3*, 993-1022.

T. Griffiths and M. Steyvers. 2004. Finding scientific topics. In *The National Academy of Sciences*, 101:5228-5235.