Semi-supervised Bibliographic Element Segmentation with Latent Permutations

Tomonari Masada[1], Atsuhiro Takasu[2], Yuichiro Shibata[1], and Kiyoshi Oguri[1]

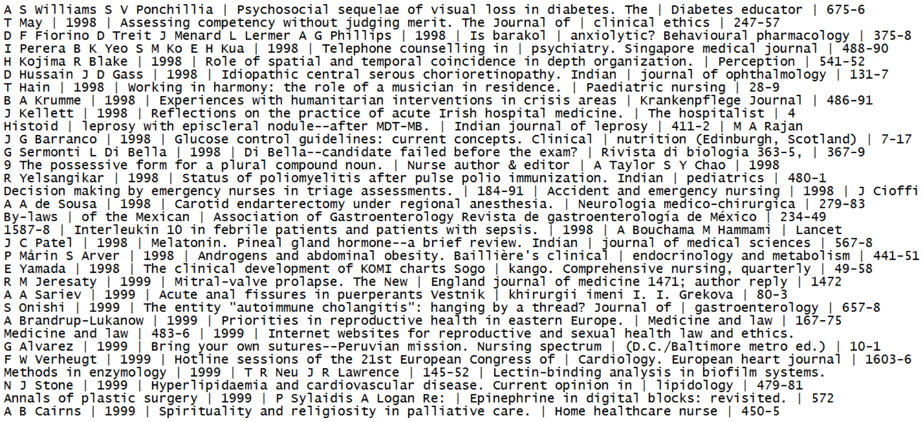
[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi, Nagasaki, Japan
{masada,shibata,oguri}@nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
takasu@nii.ac.jp

Abstract. This paper proposes a semi-supervised bibliographic element segmentation. Our input data is a large scale set of bibliographic references each given as an unsegmented sequence of word tokens. Our problem is to segment each reference into bibliographic elements, e.g. authors, title, journal, pages, etc. We solve this problem with an LDA-like topic model by assigning each word token to a topic so that the word tokens assigned to the same topic refer to the same bibliographic element. Topic assignments should satisfy contiguity constraint, i.e., the constraint that the word tokens assigned to the same topic should be contiguous. Therefore, we proposed a topic model in our preceding work [8] based on the topic model devised by Chen et al. [3]. Our model extends LDA and realizes unsupervised topic assignments satisfying contiguity constraint. The main contribution of this paper is the proposal of a *semi-supervised* learning for our proposed model. We assume that at most one third of word tokens are already labeled. In addition, we assume that a few percent of the labels may be incorrect. The experiment showed that our semi-supervised learning improved the unsupervised learning by a large margin and achieved an over 90% segmentation accuracy.

1 Introduction

Bibliographic element segmentation is an important problem when we build a large publication database from "raw" references, i.e., the references that are not segmented into bibliographic elements, e.g. author names, paper title, journal name, pages, publication year, etc. By scanning reference sections of printed articles or by crawling publication data from researchers' Web sites, we can obtain raw references as unlabeled word token sequences. Our problem is to segment each raw reference into bibliographic elements. Once bibliographic elements are identified, relevant tasks, e.g. reference alignment, reference deduplication, etc, will become easy. Figure 1 shows a segmentation example our method provides.

In a preceding paper [8], we proposed a completely new approach for bibliographic element segmentation. Existing works solve this problem by modeling the appearance order of bibliographic elements as *state transition* [2, 7, 10, 11, 6, 5] with hidden Markov models or conditional random fields. In contrast, we



```
A S Williams S V Ponchillia | Psychosocial sequelae of visual loss in diabetes. The | Diabetes educator | 675-6
T May | 1998 | Assessing competency without judging merit. The Journal of | clinical ethics | 247-57
D F Fiorino D Treit J Menard L Lermer A G Phillips | 1998 | Is barakol | anxiolytic? Behavioural pharmacology | 375-8
I Perera B K Yeo S M Ko E H Kua | 1998 | Telephone counselling in | psychiatry. Singapore medical journal | 488-90
H Kojima R Blake | 1998 | Role of spatial and temporal coincidence in depth organization. | Perception | 541-52
D Hussain J D Gass | 1998 | Idiopathic central serous chorioretinopathy. Indian | journal of ophthalmology | 131-7
T Hain | 1998 | Working in harmony: the role of a musician in residence. | Paediatric nursing | 28-9
B A Krumme | 1998 | Experiences with humanitarian interventions in crisis areas | Krankenpflege Journal | 486-91
J Kellett | 1998 | Reflections on the practice of acute Irish hospital medicine. | The hospitalist | 4
Histoid | leprosy with episcleral nodule--after MDT-MB. | Indian journal of leprosy | 411-2 | M A Rajan
J G Barranco | 1998 | Glucose control guidelines: current concepts. Clinical | nutrition (Edinburgh, Scotland) | 7-17
G Sermonti L Di Bella | 1998 | Di Bella--candidate failed before the exam? | Rivista di biologia 363-5, | 367-9
9 The possessive form for a plural compound noun. | Nurse author & editor | A Taylor S Y Chao | 1998
R Yelsangikar | 1998 | Status of poliomyelitis after pulse polio immunization. Indian | pediatrics | 480-1
Decision making by emergency nurses in triage assessments. | 184-91 | Accident and emergency nursing | 1998 | J Cioffi
A A de Sousa | 1998 | Carotid endarterectomy under regional anesthesia. | Neurologia medico-chirurgica | 279-83
By-laws | of the Mexican | Association of Gastroenterology Revista de gastroenterología de México | 234-49
1587-8 | Interleukin 10 in febrile patients and patients with sepsis. | 1998 | A Bouchama M Hammami | Lancet
J C Patel | 1998 | Melatonin. Pineal gland hormone--a brief review. Indian | journal of medical sciences | 567-8
P Mårin S Arver | 1998 | Androgens and abdominal obesity. Baillière's clinical | endocrinology and metabolism | 441-51
E Yamada | 1998 | The clinical development of KOMI charts Sogo | kango. Comprehensive nursing, quarterly | 49-58
R M Jeresaty | 1999 | Mitral-valve prolapse. The New | England journal of medicine 1471; author reply | 1472
A A Sariev | 1999 | Acute anal fissures in puerperants Vestnik | khirurgii imeni I. I. Grekova | 80-3
S Onishi | 1999 | The entity "autoimmune cholangitis": hanging by a thread? Journal of | gastroenterology | 657-8
A Brandrup-Lukanow | 1999 | Priorities in reproductive health in eastern Europe. | Medicine and law | 167-75
Medicine and law | 483-6 | 1999 | Internet websites for reproductive and sexual health law and ethics.
G Alvarez | 1999 | Bring your own sutures--Peruvian mission. Nursing spectrum | (D.C./Baltimore metro ed.) | 10-1
F W Verheugt | 1999 | Hotline sessions of the 21st European Congress of | Cardiology. European heart journal | 1603-6
Methods in enzymology | 1999 | T R Neu J R Lawrence | 145-52 | Lectin-binding analysis in biofilm systems.
N J Stone | 1999 | Hyperlipidaemia and cardiovascular disease. Current opinion in | lipidology | 479-81
Annals of plastic surgery | 1999 | P Sylaidis A Logan Re: | Epinephrine in digital blocks: revisited. | 572
A B Cairns | 1999 | Spirituality and religiosity in palliative care. | Home healthcare nurse | 450-5
```

Fig. 1. A segmentation example of our method for M20 dataset (cf. Section 4), where the bibliographic elements to be segmented are `authors`, `year`, `title`, `journal`, and `pages`. Each line gives a different reference. The symbol "`|`" shows the segmentation.

proposed a Bayesian probabilistic model as an extension of latent Dirichlet allocation (LDA) [1] and modeled the appearance order as *topic permutation*.

We regard each raw reference as a document and assign each word token to a topic so that the word tokens assigned to the same topic refer to the same bibliographic element. However, the topic assignments obtained by LDA fails to solve our problem, because the assignments do not satisfy *contiguity constraint*, i.e., the constraint that the same bibliographic element should be referred to by contiguous word tokens. In the topic assignments given by LDA, the word tokens assigned to the same topic are in general not contiguous.

Therefore, in [8], we borrowed a remarkable idea proposed by Chen et al. [3] and modified LDA as follows: For each document, 1) draw as many topics as word tokens; 2) draw a topic permutation from a probability distribution defined by generalized Mallows model (GMM) [4] over all topic permutations; and 3) sort the drawn topics according to the drawn permutation (cf. Figure 2). The idea to use GMM in topic models was originally proposed in [3] to solve *document structure learning*. However, this problem is widely different from ours. Therefore, we modified their method and gave an unsupervised method for bibliographic element segmentation. In this paper, we give a fixed name *PaiFen*[1] to our segmentation method. The main contribution of this paper is to propose a *semi-supervised learning* for PaiFen and to improve the segmentation accuracy up to over 90%. We call the semi-supervised version of PaiFen *BanPaiFen*.[2]

BanPaiFen accepts an input set of references where some word tokens are already labeled. In our experiment, at most one third of word tokens are labeled. However, these supervised labels are allowed to be imperfect, because we assume

[1] "Pai" is the pronunciation of the Greek character π often used to denote permutations, and "Fen" is the pronunciation of the Chinese character meaning segmentation.

[2] "Ban" is the pronunciation of the Chinese character standing for "semi."

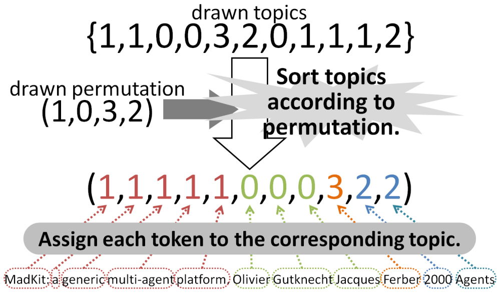

drawn topics
{1,1,0,0,3,2,0,1,1,1,2}
drawn permutation
(1,0,3,2)
Sort topics according to permutation.
(1,1,1,1,1,0,0,0,3,2,2)
Assign each token to the corresponding topic.
MadKit: a generic multi-agent platform. Olivier Gutknecht Jacques Ferber 2000 Agents

Fig. 2. How to make topic assignments satisfy contiguity constraint.

that the labels are given not by careful hand-labeling but by automated labeling. In our experiment, around 3% of the labels were incorrect. However, BanPaiFen could achieve an over 90% segmentation accuracy.

The rest of the paper is organized as follows. Section 2 describes the model PaiFen in a generative manner. Section 3 explains how we modify PaiFen for semi-supervised learning. Section 4 includes the settings and the results of our experiment. Section 5 concludes the paper with discussions and future work.

2 Topic Modeling for Segmentation

As is widely known, topic modeling trend in text mining was inaugurated by LDA [1]. By regarding references as documents and bibliographic elements as topics, we can interpret the topic assignments of LDA as bibliographic element labelings. However, the topic assignments by LDA do not satisfy *contiguity constraint*, i.e., the constraint that the word tokens assigned to the same topic should appear contiguously. This constraint is required for our problem, because each bibliographic element should be referred to by contiguous word tokens. Here we can borrow an intuition from the work of Chen et al. [3] and can use generalized Mallows model (GMM) for meeting contiguity constraint. Since GMM defines a probability distribution over topic permutations, we can model the topic appearance order in each document by a draw from this distribution.

However, the problem envisioned in [3] was *document structure learning*, a problem widely different from ours. Therefore, in [8], we proposed a modified version of their method. We call our method *PaiFen* in this paper. Chen et al. assumed *sparse topic distribution* in [3]. That is, they assumed that only a small number of topics appear in each document. This assumption is not appropriate for our problem, because all topics, i.e., all bibliographic elements, are basically expected to appear in every references. Therefore, while Chen et al. draw topic multinomial distributions from the Dirichlet prior to achieve topic sparseness, we replace all topic multinomials by the uniform distribution in PaiFen.

Due to space limitation, we skip the mathematical details of PaiFen and refer the details to [8]. We only repeat the generative description of PaiFen as below:

1. Draw a word multinomial parameter $\phi_k = (\phi_{k1}, \ldots, \phi_{kW})$, $k = 1, \ldots, K$, from the symmetric Dirichlet prior $\mathrm{Di}(\beta)$ for each of the K latent topics.
2. Draw a parameter $\rho = (\rho_1, \ldots, \rho_{K-1})$ of generalized Mallows model $\mathrm{GMM}(\rho)$ from the conjugate prior, which is defined by Equation (2) in [3].
3. Let n_j be the number of word tokens contained in the jth reference d_j. Draw d_j as an ordered word token sequence $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn_j})$ as follows:
   (a) Draw n_j topics $\mathbf{t}_j = \{t_{j1}, \ldots, t_{jn_j}\}$ uniformly from the set of K topics.
   (b) Draw a permutation π_j of the K topics from $\mathrm{GMM}(\rho)$.
   (c) Sort the drawn topics $\mathbf{t}_j$ according to the drawn permutation π_j and obtain an ordered multiset of topics $\mathbf{z}_j = (z_{j1}, \ldots, z_{jn_j})$ (cf. Figure 2).
   (d) For x_{ji}, draw a word w from the multinomial $\mathrm{Multi}(\phi_{z_{ji}})$ and set $x_{ji} = w$.

3 Semi-supervised Inference

Basically, we adopt the MCMC inference described in [8]. However, we realize a semi-supervised learning by modifying the conditional posteriors in a *scheduled* manner. We need to generate MCMC samples for the following three types of variables: GMM parameters ρ, inversion counts $\mathbf{v}_j$, and topic draws $\mathbf{t}_j$. GMM parameters are sampled in the same manner as [8]. Therefore, we explain how inversion counts and topic draws are sampled in our semi-supervised inference.

We present necessary notations. *Inversion counts* are introduced to specify a topic permutation in a one-to-one manner. For each reference d_j, a set of $K-1$ inversion counts $\mathbf{v}_j = (v_{j1}, \ldots, v_{jK-1})$ are prepared as latent variables. Each v_{jk} is an integer satisfying $0 \leq v_{jk} \leq K-1$. Let $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK})$ be the topic permutation used for sorting the topic draws $\mathbf{t}_j$ for reference d_j. We denote the position of the kth topic in $\boldsymbol{\pi}_j$ by $\boldsymbol{\pi}_j^{-1}(k)$. $\boldsymbol{\pi}_j$ can be specified by $\mathbf{v}_j$ in a one-to-one manner via the following relationship: $v_{jk} \equiv |\{k' : k' > k \text{ and } \boldsymbol{\pi}_j^{-1}(k') < \boldsymbol{\pi}_j^{-1}(k)\}|$. That is, v_{jk} is the number of topics whose indices are larger than k and whose positions in $\boldsymbol{\pi}_j$ are earlier than the kth topic. For example, $\boldsymbol{\pi}_j = (4, 3, 1, 5, 2)$ can be uniquely specified by $\mathbf{v}_j = (2, 3, 1, 0)$.

In the unsupervised inference described in [8], we sample the inversion count v_{jk} for each j and k from the full conditional posterior $P(v_{jk}|\cdots)$. Further, we sample the latent topic t_{ji} for each j and i from the full conditional posterior $P(t_{ji}|\cdots)$. The mathematical details of these posteriors are referred to [8]. In BanPaiFen, these posteriors are modified by using the two types of penalties calculated for each reference d_j. The one is R_j^{mis}, the number of mismatches between the supervised label and the inferred topic. The other is R_j^{red}, the number of *redundant assignments* that are explained as follows. When more than one word tokens are assigned to the topics corresponding to the bibliographic elements that should be referred to by one word token, we call such assignments *redundant*. For example, when three word tokens are assigned to the topic corresponding to publication year, we set $R_j^{red} = 2$, because publication year should be referred to by one word token, and thus two among the three assignments are redundant. Then, $P(v_{jk}|\cdots)$ and $P(t_{ji}|\cdots)$ are modified by R_j^{mis} and R_j^{red} as

$$P(v_{jk}|\cdots)^{\exp(R_j^{mis} I/M)} \text{ and } P(t_{ji}|\cdots)^{\exp\{(R_j^{mis}+R_j^{red}) I/M\}}, \tag{1}$$

Table 1. Specifications shared by the three DBLP datasets, i.e., D0, D20, and D50.

#references = 944,755, #different words = 685,799, #word tokens = 17,408,876					
#labels	`authors`	`title`	`journal`	`year`	total
	2,851,470	1,761,034	126,420	950,721	5,689,645
(/ #word tokens)	(16.4%)	(10.1%)	(0.73%)	(5.5%)	(32.7%)
#incorrect labels	`authors`	`title`	`journal`	`year`	total
	7,549	160,325	1,218	5,966	175,058
(/ total #labels)	(0.13%)	(2.8%)	(0.02%)	(0.10%)	(3.1%)

respectively. In Eq. (1), I is the number of iterations, and M is the parameter that controls how fast the effect of the supervised labels becomes strong as iterations proceed. The modified probabilities are used after being normalized. As Eq. (1) shows, when we sample inversion counts, we only consider R_j^{mis}. This is because R_j^{red} does not change by any update of $\mathbf{v}_j$. While we tested various ways to modify the posteriors, Eq. (1) gave the best result.

We can explain what Eq. (1) means as follows. Suppose that we terminate the inference after 300 iterations. Further, suppose that we choose M as 100. Then, for earlier iterations, $R_j^{mis} I/M$ and $(R_j^{mis} + R_j^{red})I/M$ are nearly equal to zero, because $I \ll M$ holds. Therefore, the penalty scores only have a small effect. This means that we conduct an almost unsupervised inference. However, as iterations proceed, $R_j^{mis} I/M$ and $(R_j^{mis} + R_j^{red})I/M$ become larger, because I becomes larger. Consequently, the posterior probabilities become closer to zero for the sample values making penalties positive. The inference then comes to sharply avoid such sample values. For example, if some sample value for t_{ji} increases the number of the word tokens referring to publication year from one to two, the inference after hundreds of iterations sharply avoids this sample value.

We can find a previous work [9] that utilizes supervised labels by introducing response variables into topic models. However, we use the labels to directly modify the "shape" of the posterior distributions. We think that our method is more intuitive and efficient, though not mathematically elegant.

4 Evaluation Experiment

4.1 Dataset Composition

We composed the datasets by using DBLP database[3] and MEDLINE/PUBMED database[4]. With respect to DBLP database, we composed the three datasets D0, D20, and D50 based on the file `dblp.xml` dated February 8, 2010 as follows:

- We collected the references having publication years ranging from 2000 to 2009 and extracted the bibliographic elements `authors`, `title`, `booktitle`,

[3] `http://dblp.uni-trier.de/xml/`

[4] MEDLINE®/PUBMED®, a database of the U.S. National Library of Medicine.

Table 2. Specifications shared by the MEDLINE datasets, i.e., M0, M20, and M50.

#references = 3,001,207, #different words = 2,168,061, #word tokens = 87,085,708						
#labels	`authors`	`year`	`title`	`journal`	`pages`	total
	4,983,698	2,770,489	483,878	294,330	2,710,264	11,242,659
(/ #word tokens)	(5.7%)	(3.2%)	(0.56%)	(0.34%)	(3.1%)	(12.9%)
#incorrect labels	`authors`	`year`	`title`	`journal`	`pages`	total
	60,536	13,951	179,958	4,570	8,198	267,213
(/ total #labels)	(0.54%)	(0.12%)	(1.6%)	(0.04%)	(0.07%)	(2.4%)

`journal`, and `year`. We identified `booktitle` with `journal`, because these two elements could be regarded as playing the same role for our problem.

- We defined the *canonical order* of bibliographic elements as `authors` < `title` < `journal` < `year` and sorted the bibliographic elements in this order for all references. The canonical order determines the correspondence between the bibliographic elements and the latent topics. That is, any inferred segmentation is expected to assign word tokens to topic 1 (resp. 2, 3, and 4) when those tokens refer to `authors` (resp. `title`, `journal`, and `year`).
- We randomly selected $Q\%$ of the references and randomly shuffled the order of the bibliographic elements. We did not change the word token order in each bibliographic element. By forgetting bibliographic elements information, we obtained each reference as a “raw” word token sequence. D0, D20, and D50 are the datasets obtained when $Q = 0, 20, \text{and } 50$, respectively.

We call D0, D20, and D50 *DBLP datasets*. Table 1 summarizes the specifications. Note that these specifications are shared by D0, D20, and D50, because only the order of the bibliographic elements is different among these datasets.

With respect to MEDLINE/PUBMED database, the 100 files whose names ranged from `medline09n0400.xml` to `medline09n0499.xml` were used to compose the three datasets M0, M20, and M50 by applying the same procedure with D0, D20, and D50, respectively. We call the three datasets *MEDLINE datasets*, whose specifications are shown in Table 2. For MEDLINE datasets, we extracted the five bibliographic elements: `authors`, `year`, `title`, `journal`, and `pages`, and regarded this order as the canonical order. Further, we eliminated the parentheses `[` and `]` at the head and the tail of every title, because they are artifacts. Except this, we applied no preprocessing like stemming, punctuation removal, and stop word elimination to MEDLINE datasets and also to DBLP datasets.

4.2 Automated Labeling

We applied the following automated procedures to obtain supervised labels:

1. We extracted the references having publication years earlier than 2000 from DBLP database. When a word appeared as a part of some bibliographic element and never appeared as a part of the other bibliographic elements

in the extracted references, we labeled all tokens of that word by the corresponding bibliographic element. The same procedure was applied also for MEDLINE datasets by using the references contained in the 100 files from `medline09n0000.xml` to `medline09n0099.xml` of MEDLINE database. This automated labeling was applied only for `authors`, `title`, and `journal`.

2. We labeled all tokens of the words giving an integer in the interval $[1900, 2012]$ as `year`. Further, we labeled all tokens of the words matching the regular expression `[1-9][0-9]*\-[1-9][0-9]*` as `pages`.

Since the first procedure gave many incorrect labels, we removed the labels of `authors` and `journal` from all tokens of the words included in the SCOWL word list[5]. Consequently, the tokens of the words included in the SCOWL list cannot have any labels other than `title`. The label statistics are shown in Table 1 for DBLP datasets and in Table 2 for MEDLINE datasets. For DBLP datasets, 32.7% of the word tokens are labeled, but 3.1% of them are incorrectly labeled. For MEDLINE datasets, only 12.9% of the word tokens are labeled, and 2.4% of them are incorrect.

4.3 Implementation

We refer the details of the inference implementation to [8]. The number of iterations were 300 for every experiment setting. Therefore, I in Eq. (1) is at most 300. We ran the inference 10 times for every experiment setting by starting from a random initialization. The segmentation quality was measured by the F-score defined in [3], which can be roughly viewed as the proportion of the correctly segmented word tokens. Based on the 10 runs, we calculated the mean and standard deviation of the corresponding 10 F-scores. However, we do not report the standard deviation here, because it was always negligibly small.

4.4 Evaluation Results for DBLP Datasets

Figure 3 summarizes the evaluation results for DBLP datasets. We set $M = 10000$ in Eq. (1), because this setting gave fairly good results. The wall clock time of 300 iterations was 4,400 seconds on a Fedora 14 PC equipped with Intel Core i7 970 at 3.20GHz. Each bar in Figure 3 represents the mean F-score.

In Figure 3, the charts in the top left, top right, and bottom panels present the mean F-scores for D0, D20, and D50, respectively. The bottom black bar in each chart gives the baseline F-score, i.e., the F-score achieved by PaiFen. The other bars give the F-scores achieved by BanPaiFen under various settings. The tag of each bar shows the labels used in semi-supervised inference. A, T, J, and Y mean `authors`, `title`, `journal`, and `year`, respectively. For example, the tag ATY means that we used the supervised labels of `authors`, `title`, and `year`.

The best F-scores in each chart are underlined. We can observe that the settings TJY and ATJY gave the best F-scores for all D0, D20, and D50. Interestingly, we could achieve better F-scores for D50 than for D0 and D20. In D50,

[5] http://wordlist.sourceforge.net/

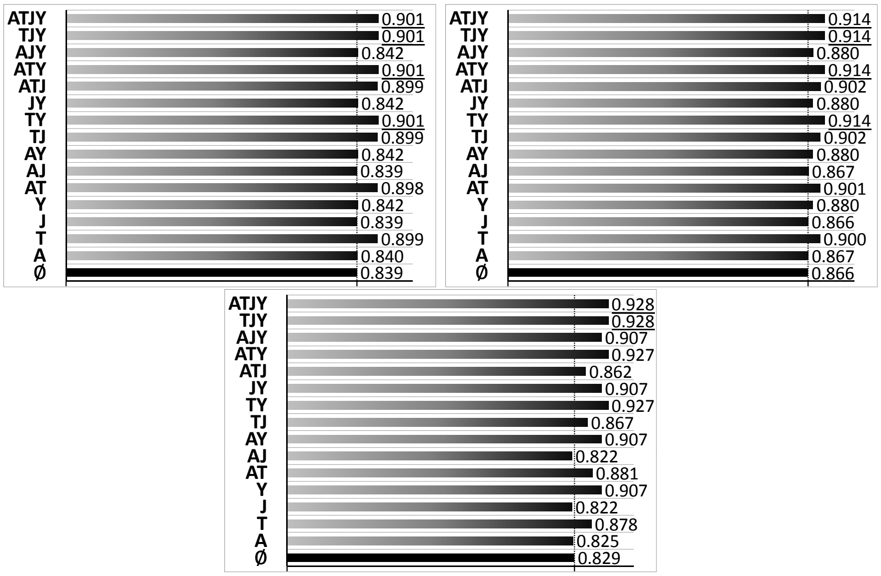


Fig. 3. Evaluation results for DBLP datasets. The charts in the top left, top right, and bottom panels present the F-scores for D0, D20, and D50, respectively. Each F-score is the mean calculated over 10 runs of the inference. The best F-scores are underlined.

50% of the references contain the bibliographic elements in a random order. It seems that the existence of a wide variety of orderings of bibliographic elements helped the inference to reach better local optima at least for DBLP datasets. Obviously, Figure 3 shows that BanPaiFen improved PaiFen by a large margin.

4.5 Evaluation Results for MEDLINE Datasets

Figure 4 presents the evaluation for MEDLINE datasets. We set $M = 100$ in Eq. (1), because this setting gave better results than $M = 10000$. 42,000 seconds were required for 300 iterations on a Fedora 14 PC equipped with Intel Core i7 970 at 3.20GHz. The tag characters A, T, J, Y, and P attached to each bar mean `authors`, `title`, `journal`, `year`, and `pages`, respectively. We only show the results for the baseline case tagged as $\emptyset$ and the following eight cases: A, T, J, Y, P, YP, ATJ, and ATJYP, because other cases gave no better results.

Figure 4 shows that the combined use of the supervised labels of `year` and `pages` achieved the best F-score for all of M0, M20, and M50. Especially, we can observe that the supervised labels of `year` played an important role. With no `year` labels, we could not improve PaiFen. This may be because `year` is placed between `authors` and `title` in the canonical order of bibliographic elements for MEDLINE datasets, and thus the correct segmentation of `year` drastically contributed to the correct segmentation of both `authors` and `title`.

Further, we can observe that the segmentation quality was worse for M50 than for M0 and M20, though we obtained better F-scores for D50 than D0 and

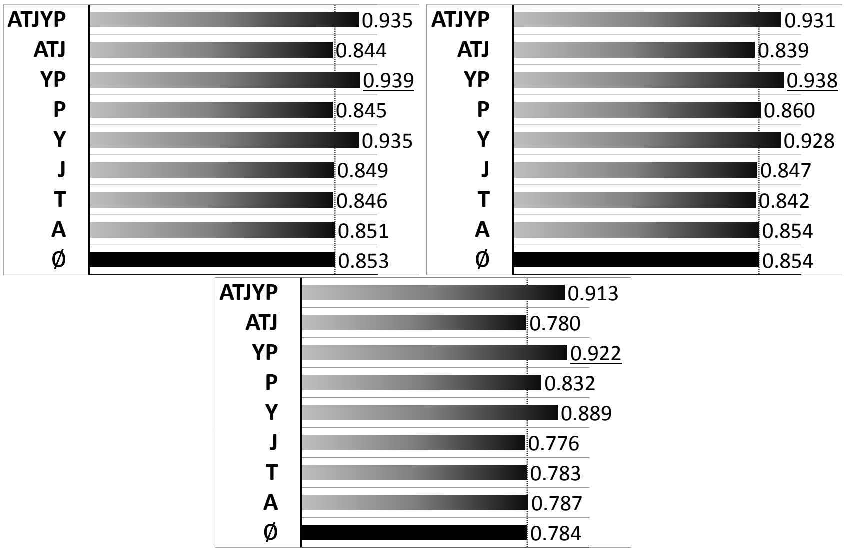


Fig. 4. F-scores for MEDLINE datasets. The charts in the top left, top right, and bottom panels present the F-scores for M0, M20, and M50, respectively. Each F-score is the mean calculated over 10 runs of the inference. The best scores are underlined.

D20 in case of DBLP datasets. While there are only $4! = 24$ permutations of bibliographic elements in total for DBLP datasets, we have $5! = 120$ permutations for MEDLINE datasets. Therefore, the inference of the correct permutation for each reference is far more difficult in case of MEDLINE datasets. This may be the reason we obtained worse F-scores for M50 than M0 and M20. However, BanPaiFen improved PaiFen by a large margin for all of M0, M20, and M50.

5 Conclusion

We proposed a semi-supervised bibliographic element segmentation called *BanPaiFen* by modifying our unsupervised method PaiFen [8]. We modified posterior probabilities in PaiFen so that the inference can avoid the sample values leading to mismatches between supervised labels and inferred topics or leading to redundant topic assignments, where the word “redundant” means that we assign more than one word tokens to the topic corresponding to the bibliographic element that should be referred to by one word token, e.g. publication year. Our experiment showed that BanPaiFen could improve PaiFen by a large margin.

In a more realistic situation, OCR errors may be included in the references obtained from the scanned articles, and crawled Web pages may include typos. Therefore, it is an important future work to incorporate an error correction at the word token level into our model, as a preceding work did for hidden Markov models [10]. One promising research direction is to propose a modeling of each word token as a character string for calculating the penalties used in Eq. (1).

We know that existing successful databases regard bibliographic element order as *state transition* and use hidden Markov models or conditional random fields. However, such approaches achieve their superiority not only with such data modelings but also with practical tunings. While our approach also requires additional practical tunings, we think that our approach can be an alternative to existing approaches as a new style of segmentation based on *permutation*, not on transition, because we have already achieved an over 90% accuracy without getting into any details about how each word token is composed as a string.

Acknowledgement

This work was supported in part by Japan Society for the Promotion of Science Grant-in-Aid for Young Scientists (B) 60413928.

References

1. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, pp. 993–1022 (2003)
2. Connan, J., Omlin, C. W.: Bibliography Extraction with Hidden Markov Models. Technical Report US-CS-TR-00-6, University of Stellenbosch (2000)
3. Chen, H., Branavan, S. R. K., Barzilay, R., Karger, D. R.: Global Models of Document Structure Using Latent Permutations. In: Proc. of North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) 2009 conference, pp. 371–379 (2009)
4. Fligner, M. A., Verducci, J. S.: Distance Based Ranking Models. Journals of the Royal Statistical Society B, Vol. 48, No. 3, pp. 359–369 (1986)
5. Hetzner, E.: A Simple Method for Citation Metadata Extraction Using Hidden Markov Models. In: Proc. of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 280–284 (2008)
6. Kramer, M., Kaprykowsky, H., Keysers, D., Breuel, T. M.: Bibliographic Meta-Data Extraction Using Probabilistic Finite State Transducers. In: Proc. of the 9th International Conference on Document Analysis and Recognition, pp. 609–613 (2007)
7. Lafferty, J. D., McCallum, A., Pereira, F. C. N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proc. of the Eighteenth International Conference on Machine Learning, pp. 282–289 (2001)
8. Masada, T., Shibata, Y., Oguri, K.: Unsupervised Segmentation of Bibliographic Elements with Latent Permutations. International Journal of Organizational and Collective Intelligence, Vol. 2, Issue 2, pp. 49–62 (2011)
9. Sharifi, M.: Semi-supervised Extraction of Entity Attributes Using Topic Models. Master's Thesis, Carnegie Mellon University (2009)
10. Takasu, A.: Bibliographic Attribute Extraction from Erroneous References Based on a Statistical Model. In: Proc. of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 49–60 (2003)
11. Yin, P., Zhang, M., Deng, Z.-H., Yang, D.-Q.: Metadata Extraction from Bibliographies Using Bigram HMM. In: Proc. of the 7th International Conference on Asian Digital Libraries, pp. 1–14 (2004)