# The Grouped Author-Topic Model for Unsupervised Entity Resolution

Andrew M. Dai and Amos J. Storkey

Institute for Adaptive and Neural Computation, School of Informatics,
University of Edinburgh, U.K.
{a.dai,a.storkey}@ed.ac.uk

**Abstract.** This paper describes a generative approach for tackling the problem of identity resolution in a completely unsupervised context with no fixed assumption regarding the true number of identities. The problem of entity resolution involves associating different references to authors (in a paper's author list, for example) with real underlying identities. The references may be written in differing forms or may have errors, and identical references may refer to different real identities. The approach taken here uses a generative model of both the abstract of a document and its list of authors to resolve identities in a corpus of documents. In the model, authors and topics are associated with latent groups. For each document, an abstract and an author list are generated conditioned on a given group. Results are presented on real-world datasets, and outperform the best performing unsupervised methods.

**Keywords:** Bayesian nonparametrics, Dirichlet processes, nested Dirichlet processes, author disambiguation

## 1 Introduction

Entity resolution is a problem encountered widely in the literature and is referred to by a variety of names that vary depending on the domain area it is used in, including record linkage, deduplication and coreference resolution. The focus of the problem is essentially to discover duplicate entities in a dataset in the absence of unique identifiers. These entities may be things that are referenced in different ways in a document, duplicate records from merging customer databases or people being referenced within multiple documents in a single corpus. It is this latter task that we focus on. One common approach to tackling this problem includes the use of clustering, such as hierarchical agglomerative clustering and $k$-means clustering, where each cluster represents an entity. However, a problem with many of these existing approaches is that they require the number of clusters or a cut-off threshold to be set in advance.

Models where the number of clusters is unknown a priori, and which are flexible enough to incorporate a range of likelihood models are attractive for this problem. Additionally, since very little labelled data exists for entity resolution, unsupervised and generative approaches are useful. One class of models which satisfy these requirements are Bayesian nonparametric models, of which

the Dirichlet process (DP) [1] has been especially widely-used. The DP is a probability distribution on the space of probability measures. Since a sample from the DP is a discrete distribution, such a sample is a natural representation for clusters. Infinite mixture models that are based on the DP are not restricted to a finite number of latent classes and so offer extra modelling flexibility. A draw from a Dirichlet process (which we will denote by $G \sim \mathrm{DP}(\alpha, H)$), is dependent on two parameter terms $H$ and $\alpha$. $H$ is called the base measure and gives the expectation of $G$, and $\alpha$ is called the concentration. For a definition of the DP we refer the reader to Ferguson [1] or one of the many introductory texts on the subject. Structured variations of the DP include both the nested Dirichlet process (NDP) [2] and the hierarchical Dirichlet process (HDP) [3].

The model described in this paper is a hierarchal generative nonparametric model for document abstracts and author lists that differs from current approaches in a number of ways. It is the first approach (to our knowledge) to integrate both topic and co-author information for tackling the task of unsupervised identity resolution. Co-author information is captured through a concept of research groups that forms part of the generative model. Each group also has a number of topics on which they write. This integration of both topic and group information enables improved performance over methods that only consider individual information sources. Furthermore, unlike earlier methods we make no assumptions regarding the equivalence of authors with names that have the same transcription in the corpus. The approach here is compared to state of the art unsupervised models and is able to both separate identical references that refer to different identities as well as combine different references that refer to the same identity, while still performing better than the current state of the art.

The remainder of this paper is set out as follows. In Section 3, we develop our framework used to tackle this problem with a description of the generative story. In Section 4, we describe inference in this framework. We then describe results on real world datasets in Section 5 and conclude in Section 6 with a discussion.

## 2 Previous work

One way to attack the entity resolution problem is via an agglomerative approach where references are merged according to some criterion until a threshold is reached. Recently, approaches for entity resolution have aimed at avoiding the need to set a threshold at which to stop merging clusters or the number of author entities in advance. To avoid this problem, several approaches have been applied. Bhattacharya and Getoor [4] describe an entity-resolution approach (LDA-ER) based on latent Dirichlet allocation (LDA), that is able to infer the number of author entities in the data. However, the number of co-authorship groups need to be pre-specified and they require labelled data for setting the parameters.

Often, models which use information from other attributes perform better than those that solely disambiguate based on names. The author-topic model, proposed by Rosen-Zvi et al. [5], associates latent topics with authors and identifies the topics that authors frequently write on. In this work, a latent topic is characterised by a distribution over words in the corpus. However, rather than entity resolution, their goal was to model the tendencies of authors to write on

certain topics or subject areas assuming the authors for each document are already known. Their model allocates words in the document to one of the known authors and does not use co-author information. However, this approach can require a large amount of data. An author must appear numerous times in the corpus for its topic distribution to be sufficiently tight for the purpose of disambiguation. Instead, in the model introduced in this paper, groups of authors are associated with topics rather than individual authors. This eliminates the difficult problem of associating authors with topics when data is limited.

## 3 Grouped Author-Topic Model

In this paper we aim to use as much of the commonly-shared information that is available for the purposes of entity resolution. This information is typically the words in the abstract, as well as the author list. This information is organised via the latent concept of a research group (which characterises which authors might be co-authors) along with topic information associated with each group (which helps disambiguate authors which could be members of a number of research groups). This leads to a model which we call the grouped author-topic model.

In the grouped author-topic model each real-world author identity will be represented by a latent *author entity*. Although a single entity, a real-world author may have a number of different names by which he or she is referred. These are known as *references* and different variants of the author's name occur due to variation in initialing, transcription errors, typographical errors, transliteration differences etc. These varying forms can be viewed as being generated by a name corruption process which, for each author, corrupts an underlying *canonical name* associated with that particular author. Any potential corruption model can be used in the context of the grouped author-topic model. We tested a generative bigram model, a trigram model and a previously-used pair hidden Markov model [4]. This last model uses domain knowledge that author names are often written with first or middle names initialled or middle name removed. We found that this corruption model performed the best.

To describe the model we need to introduce two concepts, that of *group* and that of *topic*. The idea of topic is common to other papers on topic modelling, where a topic is a mixture component defining a distribution of words. An individual abstract will only contain a small number of topics out of the total possible number. Intuitively, the idea of a group conceptualises authors who work/publish together and the associated topics they publish on. For each particular group, we define a Dirichlet process over author entities (to capture the authors that work together), and over topics (to capture the topics the group publishes on). This Dirichlet process is drawn hierarchically from a global author and topic DP. Hence author entities and topics can be shared between groups so that an author entity has non-zero probability of occurring in multiple groups, and similarly for the topics. In contrast to the author-topic model, the authors are not associated with topics directly. This model is depicted in Figure 1.

To complete the generative model we need to describe the process of generating the actual abstracts. Each abstract is associated with a group (again drawn from a DP). The group associated with the document determines which authors

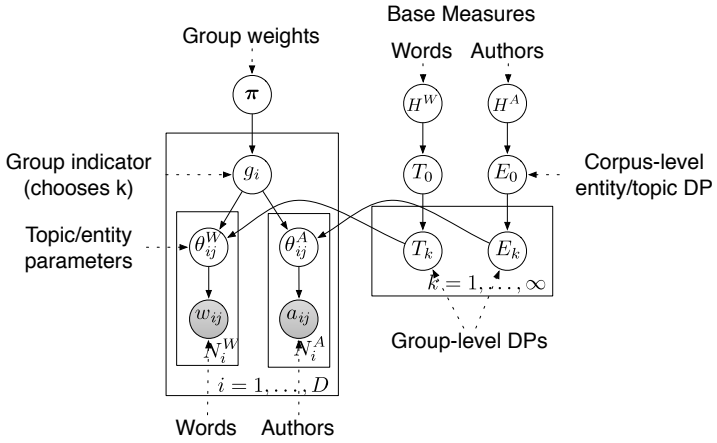


**Fig. 1.** Our generative model in plate notation. Filled in nodes are observed variables. The concentration parameters for the DPs have been omitted.

are potentially represented in a document and which topics are written about (i.e. those given significant probability by the associated group). Intuitively, this can be thought of as a document being authored by a single research group, which has a number of particular topics which they may choose to publish on, and which may be represented in the current document. The structure is loosely similar to the nested Dirichlet Process (NDP) of Rodriguez et al. [2]. However, due to the hierarchical structure in our framework, the clusters are shared between groups so that an author entity may be allocated to multiple groups. In contrast, in the standard NDP, clusters are not shared between groups.

The generative process for a whole corpus is as follows, where $\gamma$ and $\alpha$ denote concentration parameters for the global and lower level DPs respectively, the superscripts $W$ and $A$ denote the parameters or distributions for the topics and the author entities respectively. $H$ denotes the base measure, $\pi_k$ denotes the weight from the stick-breaking construction for each group $k$, and GEM represents the distribution from the stick-breaking construction [3]. These stick breaking weights determine the group DP over entities and topics $E_k, T_k$ respectively. $\theta$ denotes the parameters for the likelihood models for the authors and topics and finally $f(a|\theta^A)$ is the probability the name $a$ is corrupted from the canonical name $\theta^A$ by the name corruption model.

1. Draw (from their prior distributions) the concentration parameters for the global DPs, $\gamma^W, \gamma^A, \gamma^G$ for the topics, authors and groups respectively. Likewise, draw the concentration parameters for the lower-level DPs, $\alpha^W, \alpha^A$ from their priors.
2. Draw a global distribution over topics $T_0 \sim \mathrm{DP}(\gamma^W, H^W)$ and author entities $E_0 \sim \mathrm{DP}(\gamma^A, H^A)$. Draw a distribution over groups $\boldsymbol{\pi} \sim \mathrm{GEM}(\gamma^G)$.
3. For each group $k$, draw a distribution over topics $T_k \sim \mathrm{DP}(\alpha^W, T_0)$ and author entities $E_k \sim \mathrm{DP}(\alpha^A, E_0)$.

4. Now for each document $i = 1, \ldots, D$:
   (a) Draw a group to generate the document $g_i|\boldsymbol{\pi} \sim \boldsymbol{\pi}$.
   (b) For each word $w_{ij}, j = 1, \ldots, N_i^W$:
      i. Draw a topic $\theta_{ij}^W|g_i, T_{g_i} \sim T_{g_i}$. Draw a word $w|\theta_{ij}^W \sim \text{Mult}(w|\theta_{ij}^W)$.
   (c) For each author reference $a_{ij}, j = 1, \ldots, N_i^A$:
      i. Draw an author entity $\theta_{ij}^A|g_i, E_{g_i} \sim E_{g_i}$. Draw a (possibly corrupted) author's name from the corruption model $a|\theta_{ij}^A \sim f(a|\theta_{ij}^A)$.

In the grouped author-topic model, $H^W$ is a symmetric Dirichlet($\eta$) prior distribution over topic parameters, where a topic is parameterised by the probabilities of each word appearing in a corpus. Since this is conjugate to the likelihood (a multinomial distribution), during inference $\theta^W$ can be integrated out.

## 4 Inference

Since calculating the exact posterior under DP models is intractable, we use approximate algorithms. Due to the ease of implementing and verifying a Markov chain Monte Carlo approach, we use collapsed Gibbs sampling based on the Polya urn scheme for inference. Collapsed Gibbs sampling is described in Teh et. al [3] and involves Gibbs sampling while integrating out over conjugate distributions and random measures. The group allocations can be sampled given the word and author allocations and vice versa. As noted earlier, we integrate out the parameters for each topic, which are the multinomial distributions over the words. Since the base measure for the author names is not conjugate, we use Algorithm 8 described by Neal [6] for the author name parameters.

The true names in the corpus are considered latent variables in the grouped author-topic model. However, for practical purposes, to avoid the search over all possible canonical names, we make the computationally simplifying assumption that the true name can be sufficiently well represented by one of the references in the corpus. Every unique author name that appears in the corpus is therefore given a uniform prior probability of being the canonical name for an entity, $H^A = \text{Multinomial}(1/A_N)$ where $A_N$ is the number of unique names observed. This is equivalent to using an empirical prior for the space of canonical names.

## 5 Experiments

We tested the grouped author-topic model on the author lists and abstracts from several standard publicly available citation databases. We chose the real-world CiteSeer and Rexa databases as their ground truth is publicly available. The CiteSeer dataset, created by Giles [7] with ground truth compiled by Culotta and McCallum, consists of citations to four areas in machine learning. After removing duplicate documents in the CiteSeer dataset, it contains 1,695 references to 1,158 authors across 862 documents. The Rexa dataset [8] contains 9,366 author references in total with 1,972 of those labelled, by Culotta, to 105 author identities across 2,697 documents. Compared to the Rexa dataset, the CiteSeer dataset contains many more singleton author entities, authors that only appear once in the corpus. We applied a standard stoplist and stemming.

We compare the grouped author-topic model with other similar approaches. The *words with authors* model can be seen as a non-parametric version of the

**Table 1.** $B^3$ results on Rexa and CiteSeer datasets. Means and standard deviations are across 10 parallel chains, each with a 1,000 iteration burn-in. *Grouped A-T* is the grouped author-topic model, *group per word* relaxes the model allowing abstracts to be allocated to multiple groups, *words with authors* is the model similar to the author-topic model where words are allocated to entities without groups and *without abstracts* is a simple HDP model that ignores abstracts and does not use groups.

| | Rexa | | | CiteSeer | | |
|---|---|---|---|---|---|---|
| Model | Recall | Precision | F1 | Recall | Precision | F1 |
| Grouped A-T | 95.6 | 99.7 | **97.6 (± 0.3)** | 98.7 | 99.5 | **99.2 (± 0.1)** |
| Group per word | 95.2 | 99.5 | 97.3 (± 0.3) | 99.3 | 85.7 | 92.0 (± 0.9) |
| Words with authors | 93.6 | 97.3 | 95.4 (± 1.0) | 95.1 | 39.3 | 55.6 (± 0.4) |
| Without abstracts | 93.0 | 99.3 | 96.0 (± 0.3) | 97.2 | 97.4 | 97.3 (± 0.2) |
| LDA-ER | 92.6 | 99.4 | 95.9 (± 1.2) | 97.0 | 100 | 98.4 (± 0.1) |
| Baseline distance | 57.4 | 99.6 | 72.8 | 78.5 | 100 | 88.0 |

author-topic model [5] adapted for author disambiguation. We implemented the LDA-ER model [4], which uses the concept of groups to perform disambiguation but does not use any abstract or title information. We also evaluate against a baseline distance measure that assigns identical names to the same identity. $\eta$ was set to 0.01 in common with the author-topic model and for the entities we placed an uninformative Gamma$(1, 0.01)$ prior on the global concentration parameter and a Gamma$(1, 0.1)$ prior on the lower-level concentration parameter and updated by sampling from their posterior. These priors and similar priors on concentration parameters were chosen to give a uniform prior on the number of clusters following the algorithm in Dorazio [9]. Changing the priors by an order of magnitude did not significantly influence the results. We calculated the standard $B^3$ score [10] used for coreference and the results are shown in Table 1.

The sampler converged in terms of the log likelihood of each chain and between chains after 200 iterations. It took 40 minutes to sample 1,000 iterations running on a single core of an Intel Xeon server for the CiteSeer dataset. We burned-in for 1,000 iterations and sampled for a further 1,000, evaluating on the posterior author entity assignments. For each round of sampling the entity and topic allocations, we perform 10 iterations of group sampling to improve mixing of groups. An example of an inferred group from the Rexa dataset spread across 20 documents is: *N. Cristianini, Taylor J. Shawe, J. Kandola, J. Platt, H. Lodhi, P. L. Montgomery* with the topics: *spectral, clustering, classification, semantic, kernel, method, extension.* Our results show that the grouped author-topic model performs better than other unsupervised approaches including LDA-ER. Even though LDA-ER performs well in the CiteSeer dataset, their approach assumes that identical author references always refer to the same author identity. As can be seen in the baseline, there is little ambiguity in the CiteSeer dataset. Applying this assumption to the grouped author-topic model can be done by requiring identical references to be assigned to the same entity. However, this would result

**Table 2.** Macro-averaged $B^3$ disambiguation results on the WePS 2 dataset.

| Model | Recall | Precision | F1 |
|---|---|---|---|
| Unsupervised grouped Author-Topic | 50 | 82 | **56** |
| Supervised bag of words | 48 | 95 | 59 |
| Baseline (each document in individual cluster) | 24 | 100 | 34 |

in a model that would no longer be able to handle ambiguous names, the handling of which was an advantage over LDA-ER. Our results also show that our grouped author-topic model succeeds in integrating abstract and co-author information as compared to the models which do not. The model with words directly assigned to authors likely performs poorer due to the posterior overweighting author entities with many assigned words.

Finally, we show results on a task that LDA-ER cannot tackle due to its assumption that authors with identical names always refer to the same entity. We ran experiments on the dataset from the WePS 2 [11] people clustering task. The goal of the task is to disambiguate person names in web search results. 30 randomly chosen names were searched for on an Internet search engine. The top 150 search results were retrieved and each document was hand annotated to match with a real identity. The dataset is highly ambiguous with an average of 18 different people per name. We extracted the words from each webpage, removed stopwords and ran the result through the Stanford named entity recogniser [12]. We used the extracted named entities in place of the author references in our model and used the Jaro-Winkler distance metric as the name corruption model. This flexible model was chosen to allow matching of name, location and organization entities written in different forms. We used the non-entity words as the observed words for each document. We then performed experiments with priors on the concentration parameters that were scaled logarithmically in proportion to the given real-world frequency of that name. Since identities are at the document level, we evaluate our model using the posterior group assignments. The results in Table 2 show that our unsupervised model almost matches the performance of the supervised bag of words approach. Our model performs well compared to other teams [11] despite the majority of the other teams being reliant on supervised approaches with additional features based on extracted attributes of the person, cutoff distances or additional queries on a search engine.

## 6 Discussion

Our grouped author-topic model models the authorship of a document through a hierarchical model that combines a topic model with a multiple authorship model. This allows information that comes from a document having multiple authors and the topic specific content in a document to be leveraged to usefully disambiguate the authors that are represented in the corpus. We have evaluated the model against real world data and shown that it performs well in the task of identity resolution against other unsupervised state of the art approaches. The

model shows significant improvement over ignoring groups or abstracts in the citation database examples and shows that it can perform well at disambiguating a set of documents where the names are identical.

Our model is versatile in that it can disambiguate identical name references that refer to different entities as well as combine differing references to the same entity. The model is fully automated in that it does not require pre-specification of numbers of entities, research groups, topics etc. This is a result of the model taking a Bayesian non-parametric approach to the problem and allowing broad uninformative priors to be set on the number of entities, etc. while allowing more informative priors over the number of entities to be chosen if needed. Although the base measure for the entities is non-conjugate, using an auxiliary variable Gibbs sampler still resulted in good performance. The name corruption model could be changed to a bigram model or a discriminative name model to simplify inference or to use the model in other settings. For example, the appropriate likelihood and base measure may allow the modelling of co-entity relationships to be used for word sense disambiguation.

## References

1. Ferguson, T.S.: A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics **1**(2) (1973) 209–230
2. Rodriguez, A., Dunson, D.B., Gelfand, A.E.: The nested Dirichlet process. Journal of the American Statistical Association **103**(483) (2008) 1131–1154
3. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet Processes. Journal of the American Statistical Association **101**(476) (2006) 1566–1581
4. Bhattacharya, I., Getoor, L.: A Latent Dirichlet Model for Unsupervised Entity Resolution. In: The SIAM International Conference on Data Mining (SIAM-SDM), Bethesda, MD, USA (2006)
5. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: UAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence, Arlington, Virginia, United States, AUAI Press (2004) 487–494
6. Neal, R.M.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics **9**(2) (2000) 249–265
7. Giles, C.L., Bollacker, K.D., Lawrence, S.: CiteSeer: An Automatic Citation Indexing System. In: Digital Libraries 98 - The Third ACM Conference on Digital Libraries. (1998) 89–98
8. Peng, F., Mccallum, A.: Information extraction from research papers using conditional random fields. Information Processing & Management **42**(4) (2006) 963–979
9. Dorazio, R.M.: On selecting a prior for the precision parameter of Dirichlet process mixture models. Journal of Statistical Planning and Inference **139**(9) (September 2009) 3384–3390
10. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the Vector Space Model. In: Proceedings of the 17th international conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1998) 79–85
11. Artiles, J., Gonzalo, J., Sekine, S.: WePS 2 Evaluation Campaign : Overview of the Web People Search Clustering Task. In: Evaluation. (2009)
12. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics ACL 05 **43** (2005) 363–370