Unsupervised Name Ambiguity Resolution Using A Generative Model

Zornitsa Kozareva and Sujith Ravi
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{kozareva,sravi}@isi.edu

Abstract

Resolving ambiguity associated with names found on the Web, Wikipedia or medical texts is a very challenging task, which has been of great interest to the research community. We propose a novel approach to disambiguating names using Latent Dirichlet Allocation, where the learned topics represent the underlying senses of the ambiguous name. We conduct a detailed evaluation on multiple data sets containing ambiguous person, location and organization names and for multiple languages such as English, Spanish, Romanian and Bulgarian. We conduct comparative studies with existing approaches and show a substantial improvement of 15 to 35% in task accuracy.

1 Introduction

Recently, ambiguity resolution for names found on the Web (Artiles et al., 2007), Wikipedia articles (Bunescu and Pasca, 2006), news texts (Pedersen et al., 2005) and medical literature (Ginter et al., 2004) has become an active area of research. Like words, names are ambiguous and can refer to multiple entities. For example, a Web search for *Jerry Hobbs* on Google returns a mixture of documents associated with two different entities in the top 10 search results. One refers to a computational linguist at University of Southern California and the other refers to a fugitive and murderer. Disambiguating the names and identifying the correct entity is very important especially for Web search applications since 11-17% of the Web search queries are composed of person name and a term (Artiles et al., 2009a).

In the past, there has been a substantial body of work in the area of name disambiguation under a variety of different names and using diverse set of approaches. Some refer to the task as *cross-document coreference resolution* (Bagga and Baldwin, 1998), *name discrimination* (Pedersen et al., 2005) or *Web People Search* (WebPS) (Artiles et al., 2007). The majority of the approaches focus on person name ambiguity (Chen and Martin, 2007; Artiles et al., 2010), some have also explored organization and location name disambiguation (Pedersen et al., 2006).

The intuition behind most approaches follows the distributional hypothesis (Harris, 1954) according to which ambiguous names sharing the same contexts tend to refer to the same individual. To model these characteristics, Bunescu and Pasca (2006) and Cucerzan (2007) incorporate information from Wikipedia articles, Artiles et al. (2007) use Web page content, Mann and Yarowsky (2003) extract biographic facts. The approaches used in the WebPS tasks mainly rely on bag-of-words representations (Artiles et al., 2007; Chen and Martin, 2007; Artiles et al., 2009b). Most methods suffer from a common drawback—they rely on surface features such as word co-occurrences, which are insufficient to capture hidden information pertaining to the entities (senses) associated with the documents.

We take a novel approach for tackling the problem of name ambiguity using an unsupervised topic modeling framework. To our knowledge, no one has yet explored the disambiguation of names using Latent Dirichlet Allocation (LDA) nor has shown LDA's behavior on multiple data sources and settings. Our motivation for using an unsupervised

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 105–112,
Edinburgh, Scotland, UK, July 27–31, 2011.

topic modeling framework for name disambiguation is based on the advantages generative models offer in contrast to the existing ones. For instance, topic models such as Latent Dirichlet allocation (LDA) method (Blei et al., 2003) have been widely used in the literature for other applications to uncover hidden (or latent) groupings underlying a set of observations. Topic models are capable of handling ambiguity and distinguishing between uses of words with multiple meanings depending on context. Thereby, they provide a natural fit for our name disambiguation task, where latent topics correspond to the entities (name senses) representing the documents for an ambiguous name. Identifying these latent topics helps us identify the particular sense of a given ambiguous name that is used in the context of a particular document and hence resolve name ambiguity. In addition, this approach offers several advantages—(1) entities (senses) can be learnt automatically from a collection of documents in an unsupervised manner, (2) efficient methods already exist for performing inference in this model so we can easily scale to Web data, and (3) unlike typical approaches, we can easily apply our learnt model to resolve name ambiguity for unseen documents.

The main contributions of this paper are:

- We propose a novel model for name disambiguation using Latent Dirichlet Allocation.
- Unlike previous approaches, which are designed for specific tasks, corpora and languages, we conduct a detailed evaluation taking into consideration the multiple properties of the data and names.
- Our experimental study shows that LDA can be used as a general name disambiguation framework, which can be successfully applied on any *corpora* (i.e. Web, news, Wikipedia), *languages* (i.e. English, Spanish, Romanian and Bulgarian) and *types of ambiguous names* (i.e. people, organizations, locations).
- We conduct a comparative study with existing state-of-the-art clustering approaches and show substantial improvements of 15 to 35% in task accuracy.

The rest of the paper is organized as follows. In Section 2 we describe related work. Section 3 describes the Latent Dirichlet Allocation model used to disambiguate the names. Section 4 describes the experiments we have conducted on multiple data sets and languages. Finally, we conclude in Section 5.

2 Related Work

Ambiguous names have been disambiguated with varying success from structured texts (Pedersen et al., 2006), semi-structured texts such as Wikipedia articles (Bunescu and Pasca, 2006; Cucerzan, 2007) or unstructured texts such as those found on the Web (Pedersen and Kulkarni, 2007; Artiles et al., 2009b). Most approaches (Artiles et al., 2009b; Chen et al., 2009; Lan et al., 2009) focus on person name disambiguation, while others (Pedersen et al., 2006) also explore ambiguity in organization and location names. In the medical domain, Hatzivassiloglou et al. (2001) and Ginter et al. (2004) tackle the problem of gene and protein name disambiguation.

Due to the high interest in this task, researchers have explored a wide range of approaches and features. Among the most common and efficient ones are those based on clustering and bag-of-words representation (Pedersen et al., 2005; Artiles et al., 2009b). Mann and Yarowsky (2003) extract biographic facts such as date or place of birth, occupation, relatives among others to help resolve ambiguous names of people. Others (Bunescu and Pasca, 2006; Cucerzan, 2007; Nguyen and Cao, 2008) work on Wikipedia articles, using infobox and link information. Pedersen et al. (2006) rely on second order co-occurrence vectors. A few others (Matthias, 2005; Wan et al., 2005; Popescu and Magnini, 2007) identify names of people, locations and organizations and use them as a source of evidence to measure the similarity between documents containing the ambiguous names. The most similar work to ours is that of Song et al. (2007) who use a topic-based modeling approach for name disambiguation. However, their method explicitly tries to model the distribution of latent topics with regard to person names and words appearing within documents whereas in our method, the latent topics represent the underlying entities (name senses) for an ambiguous name.

Unlike the previous approaches which were specifically designed and evaluated on the WebPS

task or a corpus such as Wikipedia or the Web, in this paper we show a novel unsupervised topic modeling approach for name disambiguation for any corpora (i.e. Web, news, Wikipedia), languages (i.e. English, Spanish, Romanian and Bulgarian) and semantic categories (i.e. people, location and organization). The obtained results show substantial improvements over the existing approaches.

3 Name Disambiguation with LDA

Recently, topic modeling methods have found widespread applications in NLP for various tasks such as summarization (Daumé III and Marcu, 2006), inferring concept-attribute attachments (Reisinger and Pasca, 2009), selectional preferences (Ritter et al., 2010) and cross-document co-reference resolution (Haghighi and Klein, 2010).

Topic models such as LDA are generative models for documents and represent hidden or latent topics (where a topic is a probability distribution over words) underlying the semantic structure of documents. An important use for methods such as LDA is to infer the set of topics associated with a given document (or a collection of documents). Next, we present a novel approach for the task of name disambiguation using unsupervised topic models.

3.1 Method Description

Given a document corpus D associated with a certain ambiguous name, our task is to group the documents into K sets such that each document set corresponds to one particular entity (sense) for the ambiguous name. We first formulate the name disambiguation problem as a topic modeling task and then apply the standard LDA method to infer hidden topics (senses). Our generative story is as follows:

for each name sense s_k where $k \in \{1, ..., K\}$ **do**
  Generate β_{s_k} according to $Dir(\eta)$
end for
for each document i in the corpus D **do**
  Choose $\theta_i \sim Dir(\alpha)$
  for each word $w_{i,j}$ where $j \in \{1, ..., N_i\}$ **do**
    Choose a sense $z_{i,j} \sim Multinomial(\theta_i)$
    Choose a word $w_{i,j} \sim Multinomial(\beta_{z_{i,j}})$
  end for
end for

3.2 Inference

We perform inference on this model using collapsed Gibbs sampling, where each of the hidden sense variables $z_{i,j}$ are sampled conditioned on an assignment for all other variables, while integrating over all possible parameter settings (Griffiths and Steyvers, 2002). We use the MALLET (McCallum, 2002) implementation of LDA for our experiments. We ran LDA with different parameter settings on a held out data set and found that the following configuration resulted in the best performance. We set the hyperparameter η to the default value of 0.01. For the name discrimination task, we have to choose from a smaller set of name senses and each document is representative of a single sense, so we use a sparse prior (α=0.1). On the other hand, the Web People Search data is more noisy and also involves a large number of senses, so we use a higher prior (α=50).

For the name discrimination task (Section 4.1), we are given a set of senses to choose from and hence we can use this value to fix the number of topics (senses) K in LDA. However, it is possible that the number of senses may be unknown to us apriori. For example, it is difficult to identify all the senses associated with names of people on the Web. In such scenarios, we set the value of K to a fixed value. For experiments on Web People Search, we set $K = 40$, which is roughly the average number of senses associated with people names on the Web. An alternative strategy is to automatically choose the number of senses based on the model that leads to the highest posterior probability (Griffiths and Steyvers, 2004). It is easy to incorporate this technique into our model, but we leave this for future work.

3.3 Interpreting Name Senses From Topics

As a result of training, our model outputs the topic (sense) distributions for each document in the corpus. Although the LDA model can assign multiple senses to a document, the name disambiguation task specifies that each document should be assigned only to a single name sense. Hence, for each document i we assign it the most probable sense from its sense distribution. This allows us to cluster all the documents in D into K sets.

To evaluate our results against the gold standard

data, we further need to find a mapping between our document clusters and the true name sense labels. For each cluster k, we identify the true sense labels (using the gold data) for every document which was assigned to sense k in our output, and pick the majority sense label $label_{k_{maj.}}$ as being representative of the entire cluster (i.e., all documents in cluster k will be labeled as belonging to sense $label_{k_{maj.}}$). Finally, we evaluate our labeling against the gold data.

4 Experimental Evaluation

Our objective is to study LDA's performance on multiple datasets, name categories and languages. For this purpose, we evaluate our approach on two tasks: *name discrimination* and *Web People Search*, which are described in the next subsections. We use freely available data from (Pedersen et al., 2006) and (Artiles et al., 2009b), which enable us to compare performance against existing methods.

4.1 Name Discrimination

Pedersen et al. (2006) create ambiguous data by conflating together tuples of non-ambiguous well known names. The goal is to cluster the contexts containing the conflated names such that the original and correct names are re-discovered. This task is known as *name discrimination*.

An advantage of the name conflation process is that data can be easily created for any type of names and languages. In our study, we use the whole data set developed by Pedersen et al. (2006) for the English, Spanish, Romanian and Bulgarian languages.

Table 1 shows the conflated names and the semantic category they belong to (i.e. person, organization or location) together with the distribution of the instances for each underlying entity in the name. In total there are eight person, eight location and three organization conflated name pairs which represent a diverse set of names of politicians, countries, cities, political parties and software companies. For four conflated name pairs the data is balanced. For example, there are 3800 examples in total for the conflated name *Bill Clinton – Tony Blair* of which 1900 are for the underlying entity *Bill Clinton* and 1900 for *Tony Blair*. For the rest of the cases the data is imbalanced. For example, there are 3344 examples for the conflated name *Yaser Arafat – Bill Clinton* of which 1004 belong to *Yaser Arafat* and 2340 to *Bill Clinton*. The balanced and imbalanced data also lets us study whether LDA's performance if affected by the different sense distributions.

Next, we show in Table 2 the overall results from the disambiguation process. For each name, we first show the baseline score which is calculated as the percentage of instances belonging to the most frequent underlying entity over all instances of that conflated name pair. For example, for the *Bill Clinton – Tony Blair* conflated name pair, the baseline is 50% since both underlying entities have the same number of examples. This baseline is equivalent to a clustering method that would assign all of the contexts to exactly one cluster.

The second column corresponds to the results achieved by the second order co-occurrence clustering approach of (Pedersen et al., 2006). This approach is considered as state-of-the-art in name discrimination after numerous features like unigram, bigram, co-occurrence and multiple clustering algorithms were tested. We denote this approach in Table 2 as *Pedersen* and use it as a comparison. Note that in this experiment (Pedersen et al., 2006) predefine the exact number of clusters, therefore we also use the exact number of senses for the LDA topics. The third column shows the results obtained by our LDA approach. The final two columns represent the difference between our LDA approach and the baseline denoted as Δ_B, as well as the difference between our LDA approach and those of *Pedersen* denoted as Δ_P. We have highlighted in bold the improvements of LDA over these methods.

The obtained results show that for all experiments independent of whether the name sense data was balanced or imbalanced, LDA has a positive increase over the baseline. For some conflated tuples like the Spanish *NATO–ETZIN*, the improvement over the baseline is 47%. For seventeen out of the twenty name conflated pairs LDA has also improved upon *Pedersen*. The improvements range from +1.29 to +19.18.

Unfortunately, we are not deeply familiar with Romanian to provide a detailed analysis of the contexts and the errors that occurred. However, we noticed that for English, Spanish and Bulgarian often the same context containing two or three of the conflated names is used multiple times. Imagine that

Category	Name	Distribution
ENGLISH		
person/politician	Bill Cinton – Tony Blair	1900+1900=3800
person/politician	Bill Clinton – Tony Blair – Ehud Barak	1900+1900+1900=5700
organization	IBM – Microsoft	2406+3401=5807
location/country	Mexico – Uganda	1256+1256=2512
location/country&state	Mexico – India – California – Peru	1500+1500+1500+1500=6000
SPANISH		
person/politician	Yaser Arafat – Bill Clinton	1004+2340=3344
person/politician	Juan Pablo II – Boris Yeltsin	1447+1450=2897
organization	OTAN (NATO) – EZLN	1093+1093=2186
location/city	New York – Washington	1517+2418=3935
location/city&country	New York – Brasil – Washington	1517+1748+2418=5863
ROMANIAN		
person/politician	Traian Basescu – Adrian Nastase	1804+1932=3736
person/politician	Traian Basescu – Ion Illiescu – Adrian Nastase	1948+1966+2301=6215
organization	Romanian Democratic Party – Socialist Party	2037+3264=5301
location/city	Brasov – Bucarest	2310+2559=4869
location/country	France – USA – Romania	1370+2396+3890=7656
BULGARIAN		
person/politician	Petar Stoyanov – Ivan Kostov – Georgi Parvanov	318+524+811=1653
person/politician	Nadejda Mihaylova – Nikolay Vasilev – Stoyan Stoyanov	645+849+976=2470
organization	Bulgarian Socialist Party – Union Democratic Forces	2921+4680=7601
location/country	France – Germany –Russia	1726+2095+2645=6466
location/city	Varna – Bulgaria	1240+1261=2501

Table 1: *Data Set Characteristics of the Name Discrimination Task.*

Name	Baseline	Pedersen	LDA	Δ_B	Δ_P
ENGLISH					
Bill Cinton – Tony Blair	50.00%	80.95%	81.13%	**+31.13**	**+0.18**
Bill Clinton – Tony Blair – Ehud Barak	33.33%	47.93%	67.19%	**+33.86**	**+19.26**
IBM – Microsoft	58.57%	63.70%	65.44%	**+6.87**	**+1.74**
Mexico – Uganda	50.00%	59.16%	78.34%	**+28.35**	**+19.18**
Mexico – India – California – Peru	25.00%	28.78%	46.43%	**+21.43**	**+17.65**
SPANISH					
Yaser Arafat – Bill Clinton	69.98%	77.72%	83.67%	**+13.69**	**+5.95**
Juan Pablo II – Boris Yeltsin	50.05%	87.75%	52.36%	**+2.31**	-35.39
OTAN (NATO) – EZLN	50.00%	69.81%	96.89%	**+46.89**	**+27.08**
New York – Washington	61.45%	54.66%	66.73%	**+5.28**	**+12.07**
New York – Brasil – Washington	42.55%	42.88%	59.28%	**+16.73**	**+16.40**
ROMANIAN					
Traian Basescu – Adrian Nastase	51.34%	51.34%	58.51%	**+7.17**	**+7.17**
Traian Basescu – Ion Illiescu – Adrian Nastase	37.02%	39.31%	47.69%	**+10.67**	**+8.38**
Romanian Democratic Party – Socialist Party	61.57%	77.70%	61.57%	0.00	-16.13
Brasov – Bucarest	52.56%	63.67%	64.96%	**+12.40**	**+1.29**
France – USA – Romania	50.81%	52.66%	55.39%	**+4.58**	**+2.73**
BULGARIAN					
Petar Stoyanov – Ivan Kostov – Georgi Parvanov	49.06%	58.68%	57.96%	**+8.90**	-0.72
Nadejda Mihaylova – Nikolay Vasilev – Stoyan Stoyanov	39.51%	59.39%	53.97%	**+14.46**	-5.42
Bulgarian Socialist Party – Union Democratic Forces	61.57%	57.31%	61.76%	**+0.19**	**+4.45**
France – Germany –Russia	40.91%	41.60%	46.74%	**+5.83**	**+5.14**
Varna – Bulgaria	50.42%	50.38%	51.78%	**+1.36**	**+1.40**

Table 2: *Results on the Multilingual and Multi-category Name Discrimination Task.*

there is a single context in which both names *Nadejda Mihaylova* and *Stoyan Stoyanov* are mentioned. This context is used to create two name conflated examples. In the first case only the name *Nadejda Mihaylova* was hidden with the *Nadejda Mihaylova – Nikolay Vasilev – Stoyan Stoyanov* label while the name *Stoyan Stoyanov* was preserved as it is. In the second case, the name *Stoyan Stoyanov* was hidden with the label *Nadejda Mihaylova – Nikolay Vasilev – Stoyan Stoyanov* while the name *Nadejda Mihaylova* was preserved. Since the example contains two name conflations of the same context, it becomes very difficult for any algorithm to identify this phenomenon and discriminate the names correctly.

According to a study conducted by (Pedersen et al., 2006), the conflated entities in the automatically collected data sets can be ambiguous and can belong to multiple semantic categories. For example, they mention that the city *Varna* occurred in the collection as part of other named entities such as the *University of Varna*, *the Townhall of Varna*. Therefore, by conflating the name *Varna* in the organization named entity *University of Varna*, the context starts to deviate the meaning of *Varna* as a city into the meaning of university. Such cases transmit additional ambiguity to the conflated name pair and make the task even harder.

Finally, our current approach does not use stopwords except for English. According to Pedersen et al. (2006) the usage of stop-words is crucial for this task and leads to a substantial improvement.

4.2 Web People Search

Recently, Artiles et al. (2009b) introduced the *Web People Search* task (WebPS), where given the top 100 web search results produced for an ambiguous person name, the goal is to produce clusters that contain documents referring to the same individual.

We have randomly selected from the WebPS-2 test data three names from the Wikipedia, ACL'08 and Census categories. Unlike the previous data, WebPS has (1) names with higher ambiguity from 3 to 56 entities per name, (2) only person names and (3) unstructured and semi-structured texts from the Web and Wikipedia[1]. Table 3 shows the number of entities (senses) (#E) and the number of documents for each ambiguous name (#Doc).

In contrast to the previous task where the number of topics is equal to the exact number of senses, in this task the number of topics is approximate to the number of senses[2]. In our experiments we set the number of topics to 40. We embarked on this experimental set up in order to make our results comparable with the rest of the systems in WebPS. However, if we use the exact number of name senses then LDA achieves higher results.

To evaluate the performance of our approach, we use the official WebPS evaluation script. We report BCubed Precision, Recall and F-scores for our LDA approach, two baseline systems and the ECNU (Lan et al., 2009) system from the WebPS-2 challenge. We compare our results against ECNL, because they use similar word representation but instead of relying on LDA they use a clustering algorithm. We denote in Table 3 the difference between the F-score performances of LDA and the ECNU system as Δ_{F_1}. We highlight the differences in bold.

Since a name disambiguation system must have good precision and recall results, we decided to compare our results against two baselines which represent the extreme case of a system that reaches 100% precision (called ONE-IN-ONE) or a system that reaches 100% recall (called ALL-IN-ONE). Practically ONE-IN-ONE corresponds to assigning each document to a different cluster (individual sense), while the ALL-IN-ONE baseline groups together all web pages into a single cluster corresponding to one name sense (the majority sense). A more detailed explanation about the evaluation measures and the intuition behind them can be found in (Artiles et al., 2007) and (Artiles et al., 2009b).

For six out of the nine names, LDA outperformed the two baselines and the ECNU system with 5 to 41% on F-score. Precision and recall scores for LDA are comparable except for *Tom Linton* and *Helen Thomas* where precision is much higher. The decrease in performance is due to the low number of senses (entities associated with a name) and the fact that LDA was tuned to produce 40 topics. To overcome this limitation, in the future we plan to work on estimating the number of topics automatically.

[1] We clean all *html* tags and remove stopwords.

[2] Researchers use from 15 to 50 number of clusters/senses.

Name	#E	#Doc	ONE-IN-ONE			ALL-IN-ONE			ECNU			LDA			Δ_{F_1}
			BEP	BER	F_1	BEP	BER	F_1	BEP	BER	F_1	BEP	BER	F_1	
Wikipedia Names															
Louis Lowe	24	100	1.00	.32	.48	.23	1.00	.37	.39	.78	.52	.63	.52	.57	**+5**
Mike Robertson	39	123	1.00	.44	.61	.11	1.00	.19	.14	.96	.25	.59	.62	.61	**+36**
Tom Linton	10	135	1.00	.11	.19	.54	1.00	.70	.68	.48	.56	.89	.22	.35	-21
ACL '08 Names															
Benjamin Snyder	28	95	1.00	.51	.67	.08	1.00	.15	.16	.79	.27	.59	.81	.68	**+41**
Emily Bender	19	120	1.00	.21	.35	.24	1.00	.39	.45	.60	.51	.78	.42	.55	**+4**
Hao Zhang	24	100	1.00	.26	.41	.21	1.00	.35	.45	.78	.57	.72	.36	.48	-9
Census Names															
Helen Thomas	3	127	1.00	.03	.06	.96	1.00	.98	.96	.24	.39	.97	.08	.15	-24
Jonathan Shaw	26	126	1.00	.32	.49	.10	1.00	.18	.18	.60	.34	.66	.51	.58	**+24**
Susan Jones	56	110	1.00	.70	.82	.03	1.00	.06	.13	.81	.22	.51	.79	.62	**+40**

Table 3: *Results for Web People Search-2.*

5 Conclusion

We have shown how ambiguity in names can be modeled and resolved using a generative probabilistic model. Our LDA approach learns a distribution over topics which correspond to entities (senses) associated with an ambiguous name. We evaluate our novel approach on two tasks: *name discrimination* and *Web People Search*. We conduct a detailed evaluation on (1) Web, Wikipedia and news documents; (2) English, Spanish, Romanian and Bulgarian languages; (3) people, location and organization names. Our method achieves consistent performance and substantial improvements over baseline and existing state-of-the-art clustering methods.

In the future, we would like to model the biographical fact extraction approach of (Mann and Yarowsky, 2003) in our LDA model. We plan to estimate the number of topics automatically from the distributions. We want to explore variants of our current model. For example, currently all words are generated by multiple topics (senses), but ideally we want them to be generated by a single topic. Finally, we want to impose additional constraints within the topic models using hierarchical topic models.

Acknowledgments

We acknowledge the support of DARPA contract FA8750-09-C-3705 and NSF grant IIS-0429360.

References

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*, pages 64–69.

Javier Artiles, Enrique Amigó, and Julio Gonzalo. 2009a. The role of named entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 534–542.

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2009b. WePS 2 evaluation campaign: overview of the web people search clustering task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Javier Artiles, Andrew Borthwick, Julio Gonzalo, Satoshi Sekine, and Enrique Amigó. 2010. WePS-3 evaluation campaign: Overview of the web people search clustering and attribute extraction ta. In *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF)*.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 79–85.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Razvan C. Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*.

Ying Chen and James H. Martin. 2007. Cu-comsem: Exploring rich features for unsupervised web personal name disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 125–128, June.

Ying Chen, Sophia Yat Mei Lee, and Chu-Ren Huang. 2009. Polyuhk: A robust information extraction system for web personal names. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Sydney, Australia, July. Association for Computational Linguistics.

Filip Ginter, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2004. New techniques for disambiguation in natural language and their application to biological text. *J. Mach. Learn. Res.*, 5:605–621, December.

Thomas L Griffiths and Mark Steyvers. 2002. A probabilistic approach to semantic representation. In *Proceedings of the Twenty-Fourth Annual Conference of Cognitive Science Society*.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1(Suppl 1):5228–5235.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393.

Zellig Harris. 1954. Distributional structure. 10(23):146–162.

Vasileios Hatzivassiloglou, Pablo A. Duboue, and Andrey Rzhetsky. 2001. Disambiguating proteins, genes, and rna in text: A machine learning approach. In *Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology*.

Man Lan, Yu Zhe Zhang, Yue Lu, Jian Su, and Chew Lim Tan. 2009. Which who are they? people attribute extraction and disambiguation in web search results. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Gideon S. Mann and David Yarowsky. 2003. Unsupervised personal name disambiguation. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 33–40.

Matthias Blume Matthias. 2005. Automatic entity disambiguation: Benefits to ner, relation extraction, link analysis, and inference. In *International Conference on Intelligence Analysis*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

Hien T. Nguyen and Tru H. Cao. 2008. Named entity disambiguation: A hybrid statistical and rule-based incremental approach. In *Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web*, ASWC '08, pages 420–433.

Ted Pedersen and Anagha Kulkarni. 2007. Unsupervised discrimination of person names in web contexts. In *Computational Linguistics and Intelligent Text Processing, 8th International Conference, CICLing 2007*, pages 299–310.

Ted Pedersen, Amruta Purandare, and Anagha Kulkarni. 2005. Name discrimination by clustering similar contexts. In *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005*, pages 226–237.

Ted Pedersen, Anagha Kulkarni, Roxana Angheluta, Zornitsa Kozareva, and Thamar Solorio. 2006. An unsupervised language independent method of name discrimination using second order co-occurrence features. In *Computational Linguistics and Intelligent Text Processing, 7th International Conference, CICLing 2006*, pages 208–222.

Octavian Popescu and Bernardo Magnini. 2007. Irst-bp: Web people search using name entities. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007*, pages 195–198. Association for Computational Linguistics.

Joseph Reisinger and Marius Pasca. 2009. Latent variable models of concept-attribute attachment. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 620–628. Association for Computational Linguistics, August.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434. Association for Computational Linguistics, July.

Yang Song, Jian Huang, Isaac G. Councill, Jia Li, and C. Lee Giles. 2007. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries*, pages 342–351.

Xiaojun Wan, Jianfeng Gao, Mu Li, and Binggong Ding. 2005. Person resolution in person search results: Webhawk. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 163–170.