# Wikilinks: A Large-scale Cross-Document Coreference Corpus Labeled via Links to Wikipedia

SAMEER SINGH
Computer Science, University of Massachusetts, Amherst MA 01003

AMARNAG SUBRAMANYA
Google Research, Mountain View CA 94043

FERNANDO PEREIRA
Google Research, Mountain View CA 94043

ANDREW MCCALLUM
Computer Science, University of Massachusetts, Amherst MA 01003

**Technical Report**
Department of Computer Science
University of Massachusetts, Amherst

**UMASS-CS-2012-015**
October, 2012
Contact Email: sameer@cs.umass.edu

**Abstract**

Cross-document coreference resolution is the task of grouping the entity mentions in a collection of documents into sets that each represent a distinct entity. It is central to knowledge base construction and also useful for joint inference with other NLP components. Obtaining large, organic labeled datasets for training and testing cross-document coreference has previously been difficult. This paper presents a method for automatically gathering massive amounts of naturally-occurring cross-document reference data. We also present the Wikilinks dataset comprising of 40 *million* mentions over 3 million entities, gathered using this method. Our method is based on finding hyperlinks to Wikipedia from a web crawl and using anchor text as mentions. In addition to providing large-scale labeled data without human effort, we are able to include many styles of text beyond newswire and many entity types beyond people.

# 1 Introduction

A main goal of information extraction from free text is to identify the entities mentioned, which is a precondition for extracting the propositional content of the text. Within a single document, *coreference resolution* finds the referents of expressions such as pronouns, demonstratives, or definite descriptions. On a collection of documents, *cross-document coreference* finds the sets of mentions for each distinct entity mentioned in the collection. Cross-document coreference is not only a useful output of information extraction in itself, but it also supports other information extraction tasks (Blume, 2005; Mayfield et al., 2009). Cross-document coreference has been an active area of research (Finin et al., 2009; Baron and Freedman, 2008; Gooi and Allan, 2004; Bagga and Baldwin, 1998), with recent work focussing on scaling it up to large collections (Singh et al., 2010, 2011; Rao et al., 2010).

The lack of accepted datasets for training and evaluation has been a significant obstacle to progress in cross-document coreference. Manual annotation of coreference is tedious, time-consuming, and error-prone as the annotator must provide a complete clustering of a large number of mentions into sets of corefering mentions. Previous annotation efforts attempt to bypass these difficulties by considering only pairwise decisions among mentions and by restricting the sets of mentions that the annotator has to consider. However, those simplifications have led to annotated corpora that are too small to support supervised training, reliable error estimation, or scalability assessment. Furthermore, the distributions of coreference set sizes in these corpora are not representative of real-world text, where a few *popular* entities have a large number of mentions, while a much larger number have few mentions. Finally, most of these corpora are annotated only for a few types of entities (such as persons), and are based on carefully edited text (such as newswire). Internet text is much more varied than that, both in the types of entities and how they are mentioned.

To address the above problems, this paper presents an automated method to identify a large number of entity mentions in web text, as well as a data set dramatically larger and more varied than previous available data. Our method finds a large collection of hyperlinks to English Wikipedia pages, and annotates the corresponding anchor text (in context) as mentions of the entity or concept described by the Wikipedia page. Due to the widespread use of Wikipedia as reference source on the web, such links are plentiful. Figure 1 shows an example of two links (mentions) from two different pages, that refer to the `Banksy` page (entity).

This paper coincides with the release of a data set, henceforth refered to as *Wikilinks*, containing over 40 million mentions labeled as referring to 3 million entities — a scale that challenges most existing approaches. Since the links on these web pages have been created by their authors, the quality of labels is high (no er-

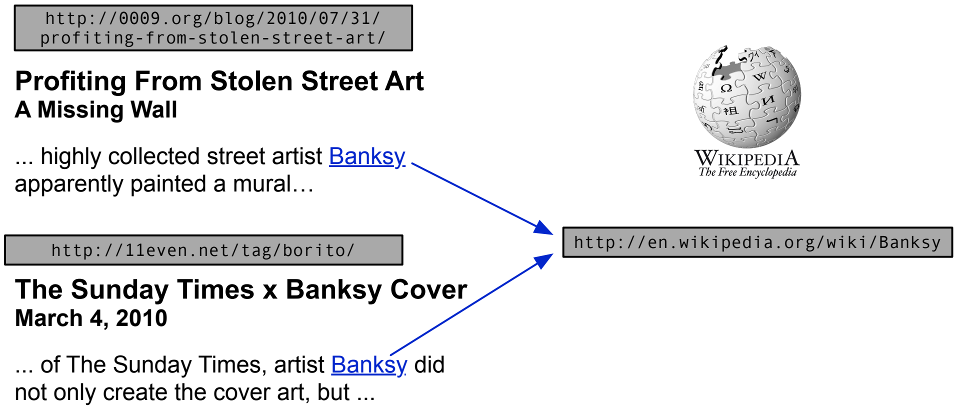
http://0009.org/blog/2010/07/31/profiting-from-stolen-street-art/
Profiting From Stolen Street Art
A Missing Wall
... highly collected street artist Banksy apparently painted a mural…
http://11even.net/tag/borito/
The Sunday Times x Banksy Cover
March 4, 2010
... of The Sunday Times, artist Banksy did not only create the cover art, but ...
Wikipedia
The Free Encyclopedia
http://en.wikipedia.org/wiki/Banksy

Figure 1: Links to Wikipedia as Entity Labels

rors were found upon manual inspection of 100 randomly sampled mentions). The dataset contains a large number of rare entities along with a few popular entities. Last, the dataset contains many types of entities, including people, locations, organizations, and general concepts, making it an excellent resource for a number of text understanding tasks including cross-document coreference.[1]

## 2 Cross-Document Coreference

Cross-document coreference classifies entity mentions in documents according to the entities to which they refer. Neither the identities nor the number of distinct entities are known in advance. This is a challenging task since the set of possible classifications is huge for a nontrivial corpus. On corpora of practical interest, the numbers of mentions and entities are in the millions. Furthermore, naming ambiguity is common as the same string can refer to multiple entities in different documents, and distinct strings may refer to the same entity in different documents.

Cross-document coreference has been the subject of many studies, which we cannot survey here exhaustively. Many studies use clustering algorithms with custom scoring functions between pairs of (mention) contexts, such as cosine distance between all pairs (Bagga and Baldwin, 1998), between "ambiguous" pairs (Ravin and Kazi, 1999), multiple similarity metrics (Gooi and Allan, 2004), and second-order co-occurrences of words (Pedersen et al., 2006). Niu et al. (2004) annotate small datasets to learn the context similarity, while others utilize external knowledge such as the web (Mann and Yarowsky, 2003) and Wikipedia (Cucerzan,

[1] We conducted some such experiments on an earlier version of this corpus in Singh et al. (2011).

2007). A number of studies describe hand-tuned weights, dictionaries, and heuristics for disambiguation (Blume, 2005; Baron and Freedman, 2008; Popescu et al., 2008). Daumé III and Marcu (2005) propose a generative approach to supervised clustering, and Haghighi and Klein (2010) use entity profiles to assist within-document coreference. More recently, studies have focused on scalability. Rao et al. (2010) propose a deterministic method that greedily assigns entities to a stream of mentions. Singh et al. (2010) perform efficient sampling by using domain knowledge to avoid all pairwise comparisons. Singh et al. (2011) introduce a hierarchical graphical model, and propose a distributed inference method. Wick et al. (2012) extend this hierarchical graphical model with arbitrary layers and explicit representation of entity attributes.

## 3 Related Datasets

Several labeled datasets have been constructed for training and evaluation of coreference resolution methods, which we now review.

**Manually Labeled:** Supervision for coreference is different from other language-processing and information extraction tasks because the hypothesis space is exponential in the number of mentions. Existing work reduces this prohibitive labeling cost by restricting the annotation decisions to those between pairs of *similar* mentions. Bentivogli et al. (2008) introduce a small dataset containing high ambiguity (209 names over 709 entities). Day et al. (2008) introduce a tool for annotating coreference data, but the resulting corpus offers little ambiguity. Bagga and Baldwin (1998) created the *John Smith* corpus, containing 197 mentions of "John Smith" from New York Times, labeled manually to obtain 35 entities. Further, this dataset also does not contain any string variations. Cross-document coreference was included in Automatic Content Extraction (ACE 2008) as *Global Entity Detection and Recognition* (Strassel et al., 2008). However, due to the difficulty of the annotation task, only 400 documents were labeled for cross-document coreference. Further, the mentions from this text refer to entities in a canonical way, e.g., newswire containing fully-qualified mentions as first references to each entity. Green et al. (2011) annotate 216 entities that appear in Arabic and English documents from this dataset, creating a cross-lingual annotated corpus.

**Automatically Labeled:** Due to the difficulty of manual annotation, several automatic methods for creating cross-document coreference datasets have been proposed. For instance, Niu et al. (2004) use approximate labels to generate small datasets for partial supervision. However the datasets are too small and noisy to be useful for training, or for reliable evaluation. Another popular technique for automatically generating labels for coreference has been to use *Person-X* evaluation (Gooi and Allan, 2004), in which unique person-name strings are treated as the true entity labels for the mentions, replacing the actual mention strings with an “X.” Although easy to generate, the difficulty of disambiguation is artificially inflated, making it a poor choice for evaluating real-world scenarios. It also suffers from the lack of mention variations. Singh et al. (2010) create a cross-document corpus by automatically aligning newswire mentions with Wikipedia pages; although the resulting dataset is large, newswire provides little ambiguity, with simple baselines obtaining high accuracy. Our corpus is similar to that of Spitkovsky and Chang (2012), in which Wikipedia entities are described by the anchor texts from the web pages that refer to it; however our corpus also provides URLs and the offsets that can be used to extract the context.

**Entity Matching:** The task of resolving mentions against a set of entities in a database is known as *entity matching*. The labeled data consists of mentions and the database records to which they refer. For mentions from text sources, this data may be used for cross-document coreference by using record linkages as entity labels. A primary dataset is the optional linking task in TAC Knowledge Base Construction (Ji et al., 2010). Another related task is that of web people disambiguation (WePS) (Artiles et al., 2010). However, these datasets differ considerably from the real-world use cases of cross-document coreference, in that they involve small numbers of mentions, the distribution of entity sizes does not rival that in large web collections, and unnatural restrictions to one or few entity types.

# 4 Annotation Pipeline

We now describe our method of gathering labeled cross-document coreference data. This method is used to create Wikilinks, and any researcher with access to a web crawl should be able to follow the same method to build an updated dataset.

## 4.1 Links to Wikipedia

The first step of the pipeline consists of identifying the set of non-Wikipedia web pages that contain links to Wikipedia. To obtain mentions that belong to the English language, we only consider web pages that have links to English Wikipedia. We use an existing recent Google search index to discover these, but other existing indices (such as Common Crawl Foundation (2011)) or search engine APIs could be used instead.

```
URL <url>\n
MENTION <anchor>\t<offset>\t<wiki url>\n
MENTION <anchor>\t<offset>\t<wiki url>\n
MENTION <anchor>\t<offset>\t<wiki url>\n
TOKEN <token>\t<offset>\n
TOKEN <token>\t<offset>\n
\n
\n
```

Figure 2: **Data format:** for a web page containing 3 mentions, and 2 rare tokens.

### 4.2 Filtering Mentions

As many exact (or near exact) copies of Wikipedia exist online, we discard web pages in which $> 70\%$ of the sentences come from a single Wikipedia page. Furthermore, we only consider links that appear in free text, skipping links that fall in tables, near images, or in obvious boilerplate material. In addition, we only consider links for which either (i) at least one token in the anchor matches a token in the title of the Wikipedia page, or (ii) the anchor text matches an alias for the target Wikipedia page, where the aliases are given by the set of anchor texts for that page within Wikipedia. We do not restrict the entities, and our annotation pipeline thus generates entities of the standard types (person, location, and organization) as well as many others, including many concepts described by common noun phrases.[2] In the next section we describe the results of several baselines showing that coreference on this dataset is still remarkably difficult.

### 4.3 Data Release

As a part of the data release, we will provide the URLs of all the pages that contain labeled mentions, the actual mentions (i.e., the anchor text), the target Wikipedia link (entity label), and the byte offsets of the links on the page. In addition, the byte offsets of the top 5 least frequent words on the page will also be made available. These are being provided since pages often change, or are removed, and may not be the same as when they were visited by our crawl. These byte offsets provide a useful page version check. Further, if the pages do not match, the user can either simply discard the page and its links, adjust the offsets based on heuristics over rare tokens, or directly search the page for links to Wikipedia. The dataset is avail-

[2] If a specific subset of entity types is desired, the entities can be filtered using the linked Wikipedia page.

able at http://code.google.com/p/wiki-links in the format given by Figure 2.

### 4.4 Accompanying Software

Along with the dataset, we will also be releasing code for: (a) downloading all the web pages; (b) extracting the mentions from the web pages, with functionality to discover these links when the byte offsets do not match; (c) extracting the context around these mentions to partially reconstruct the paragraph and the document; and (d) computing evaluation metrics over predicted entities. The code is available at http://www.iesl.cs.umass.edu/data/wiki-links. Also at this URL is a version of the data that includes the extracted context words; this dataset can be directly used without requiring users to download the webpages themselves.

## 5 Applications

Wikipedia links annotation is useful for feature extraction and for algorithm evaluation of a number of NLP and information extraction tasks.

### 5.1 Cross-Document Coreference

As mentioned earlier, the primary objective of the corpus is as a labeled resource for cross-document coreference. Much of the recent work in cross-document coreference has relied on hand-tuned parameters (Rao et al., 2010; Singh et al., 2011; Wick et al., 2012); this corpus will allow automatic learning of the parameters of these models. The annotation can also be used to evaluate and compare different techniques by running them on the set of mentions in the dataset, and comparing the annotations.

### 5.2 Within-document Coreference

This corpus may also be used to improve within-document coreference resolution. The main challenge in within-document coreference is to resolve pronouns and pronominals against the set of already identified fully-qualified entities in the corpus. For each entity in our corpus, we provide a large collection of mentions from free text on the web, which provides valuable information about the types of context that entity appears in. By matching the context of the pronoun/pronominal against the context patterns (extracted from our corpus) of the identified entities, within-document coreference can result in significant gains. There have been similar approaches that use the text and the link structure of the Wikipedia article to achieve this (Finin et al., 2009; Hoffart et al., 2011), but this corpus should help further since the context patterns in free text on the web are likely to be more varied and similar to most real-world applications than Wikipedia articles.

## 5.3 Entity Matching

Entity matching is the task of identifying to which entity from a knowledge base a given mention in free text refers. If the knowledge base is Wikipedia, then this corpus can be directly used as a labeled resource to learn the parameters and for evaluation. It can also be used to disambiguate between entities by comparing the context of the mention with the ones in the corpus. For other knowledge bases, this corpus can provide valuable context information for Wikipedia entities which may have some overlap with the entities in the knowledge base of interest.

## 5.4 Entity Tagging

Entity tagging, or named-entity recognition, is an important task in most information extraction pipelines. Due to the difficulty of labeling, most large-scale annotations only perform coarse level labeling of entity mentions into a few categories such as person, location and organization. Often these labels are not sufficiently informative for downstream tasks, as many of them require more granular description of the entities in the text. Our corpus, by allowing linking of phrases in free text to entities in Wikipedia, can project the complete ontology of Wikipedia onto the mention. For example, for a mention "Barack Obama" we have linked to `Barack Obama`, we also obtain a number of categories such as `Living People`, `African-American academics`, `Current national leaders`, and so on at the finest level, but also `American`, `President of the United States`, `Non-fiction writer`, and `Lawyer` at slightly coarser levels of the ontology. Further, each of these fine level categories is accompanied by a large number of annotations. This large amount of text and accompanying labels can be used to train highly specific, accurate entity-taggers.

## 5.5 Relation Extraction

Recent work in relation extraction has focussed on performing large-scale, cross-document relation extraction. As a pre-processing step, these approaches often rely on large-scale entity matching and entity tagging (Riedel et al., 2010; Hoffmann et al., 2011), however entity linking is often simple string matching, while an off-the-shelf entity tagger with 3 or 4 entity types is often used. Instead, an entity tagger that uses this corpus for training can help discover a much larger set

| **Dataset** | **#Mentions** | **#Entities** |
|---|---:|---:|
| Bagga and Baldwin (1998) | 197 | 35 |
| Bentivogli et al. (2008) | 43,704 | 709 |
| Day et al. (2008) | <55,000 | 3,660 |
| Artiles et al. (2010) | 57,357 | 300 |
| **Our Dataset** | 40,323,863 | 2,933,659 |

Table 1: Cross-Document Coreference Corpus Sizes

of relation mentions of interest, and can have a significant impact on the recall. Similarly, obtaining access to a larger set of finer-detail entity tags can help identify rare relations with very specific argument constraints, potentially improving the precision for relation extraction.

### 5.6 Other Tasks

A number of tasks in NLP and information extraction use one of the above tasks as an input, and thus any improvement in the above tasks aids the downstream tasks. Although this remains future work, there are also a number of simple extensions of the methodology that can result in subsequent versions of the dataset that is appropriate for a number of different tasks. For example, comparisons of regularly interval updates to the dataset can help discover temporal patterns in how a particular entity on Wikipedia is referred to over time. By releasing this dataset to the researcher community, the authors are confident there will be many unexpected uses of these annotations.

## 6 Corpus Statistics

Table 1 compares our dataset, which involves $9.5$ million web pages, to existing datasets. The dataset is larger than existing sets by several orders of magnitude. Along with enabling accurate large-scale evaluation, this size also facilitates flexibility in specifying custom subsets of the data for specific domains and tasks. To study the corpus in finer detail, we plot the histogram of entity (mention set) sizes in Figure 3. The majority of the entities have fewer than $10$ mentions each, while just a few have more than $50,000$ mentions; the data is somewhat Zipfian, fitting a truncated power law better than it fits a lognormal or an exponential distribution (Clauset et al., 2009; Alstott, 2012).

To provide further insight into the corpus, we evaluate three very simple base-

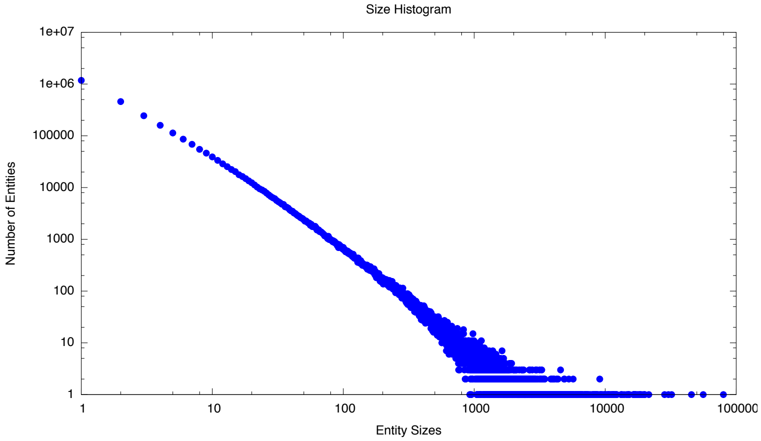

Figure 3: Distribution of the Entity Sizes

| **Baselines** | **Pairwise** | | | $\mathbf{B}^3$ | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| String Identical | 82.887 | 82.813 | 82.850 | 81.757 | 67.341 | 73.852 |
| Ignore Case | 81.150 | 85.865 | 83.441 | 80.219 | 72.954 | 76.414 |
| Match Heads | 3.690 | 87.765 | 7.083 | 33.209 | 76.698 | 46.349 |

Table 2: Results using simple baselines

line clustering algorithms on standard coreference metrics (shown in Table 2). The first baseline treats mentions with string-identical anchors to be coreferent, resulting in $4.3$ million predicted entities. This heuristic is often used in practice for cross-document coreference, since it achieves high accuracy in some domains. To improve recall, our second baseline merges mentions that are identical when ignoring case. To improve recall further, we explore another heuristic that merges mentions with identical heads, approximated by picking the rightmost token. Low accuracy by these baselines suggests the dataset is suitable for evaluation of more sophisticated coreference techniques.

## 7 Conclusions and Future Work

We introduce a large-scale corpus for cross-document coreference over free-form text. The dataset will be released publicly, along with the processing and evaluation code. The high-quality labels in the corpus allow training of cross-document coreference models, whereas its size encourages the study of scalable inference methods. This dataset also provides a valuable resource for other NLP tasks, such as named-entity tagging, within-document coreference, and entity matching.

## Acknowledgments

The authors would like to thank Brian Martin, Harshal Pandya, and John R. Frank for providing valuable assistance in downloading and processing of the webpages.

## References

Jeff Alstott. powerlaw Python package. (Version 3.1), 2012. URL http://pypi.python.org/pypi/powerlaw.

J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. Weps-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks. In *Conference on Multilingual and Multimodal Information Access Evaluation*, 2010.

Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *International Conference on Computational Linguistics*, pages 79–85, 1998.

A. Baron and M. Freedman. Who is who and what is what: experiments in cross-document co-reference. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 274–283, 2008.

Luisa Bentivogli, Christian Girardi, and Emanuele Pianta. Creating a gold standard for person cross-document coreference resolution in italian news. In *LREC Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, 2008.

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

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. doi: 10.1137/070710111.

Common Crawl Foundation. Common Crawl Dataset, November 2011. URL http://www.commoncrawl.org/.

Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 708–716, 2007.

Hal Daumé III and Daniel Marcu. A Bayesian model for supervised clustering with the Dirichlet process prior. *Journal of Machine Learning Research (JMLR)*, 6:1551–1577, 2005.

David Day, Janet Hitzeman, Michael Wick, Keith Crouch, and Massimo Poesio. A corpus for cross-document co-reference. In *International Conference on Language Resources and Evaluation (LREC)*, 2008.

Tim Finin, Zareen Syed, James Mayfield, Paul McNamee, and Christine Piatko. Using Wikitology for cross-document entity coreference resolution. In *AAAI Spring Symposium on Learning by Reading and Learning to Read*, 2009.

Chung Heong Gooi and James Allan. Cross-document coreference on a large scale corpus. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, pages 9–16, 2004.

Spence Green, Nicholas Andrews, Matthew R. Gormley, Mark Dredze, and Christopher Manning. Cross-lingual coreference resolution: A new task for multilingual comparable corpora. Technical Report 6, Human Language Technology Center of Excellence, Johns Hopkins University, 2011.

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

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 782–792, 2011.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 541–550, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

H. Ji, R. Grishman, H.T. Dang, K. Griffitt, and J. Ellis. Overview of the tac 2010 knowledge base population track. In *Text Analysis Conference (TAC)*, 2010.

Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, pages 33–40, 2003.

J. Mayfield, D. Alexander, B. Dorr, J. Eisner, T. Elsayed, T. Finin, C. Fink, M. Freedman, N. Garera, P. McNamee, et al. Cross-document coreference resolution: A key technology for learning by reading. In *AAAI Spring Symposium on Learning by Reading and Learning to Read*, 2009.

Cheng Niu, Wei Li, and Rohini K. Srihari. Weakly supervised learning for cross-document person name disambiguation supported by information extraction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, page 597, 2004.

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

Octavian Popescu, Christian Girardi, Emanuele Pianta, and Bernardo Magnini. Improving cross-document coreference. *Journées Internationales d'Analyse statistique des Données Textuelles*, 9:961–969, 2008.

Delip Rao, Paul McNamee, and Mark Dredze. Streaming cross document entity coreference resolution. In *International Conference on Computational Linguistics (COLING)*, pages 1050–1058, Beijing, China, August 2010. Coling 2010 Organizing Committee.

Yael Ravin and Zunaid Kazi. Is Hillary Rodham Clinton the president? disambiguating names across documents. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9–16, 1999.

Sebastian Riedel, Limin Yao, and Andrew McCallum. Collective cross-document relation extraction without labelled data. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2010.

Sameer Singh, Michael L. Wick, and Andrew McCallum. Distantly labeling data for large scale cross-document coreference. *Computing Research Repository (CoRR)*, abs/1005.4298, 2010.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Association for Computational Linguistics: Human Language Technologies (ACL HLT)*, 2011.

Valentin I. Spitkovsky and Angel X. Chang. A cross-lingual dictionary for english wikipedia concepts. In *International Conference on Language Resources and Evaluation (LREC)*, 2012.

Stephanie Strassel, Mark Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Michael Wick, Sameer Singh, and Andrew McCallum. A discriminative hierarchical model for fast coreference at large scale. In *Association for Computational Linguistics (ACL)*, 2012.