IMPROVING CROSS-DOCUMENT CO-REFERENCE WITH SEMI-SUPERVISED INFORMATION EXTRACTION MODELS

Rushin Shah, Bo Lin, Kevin Dela Rosa, Anatole Gershman, Robert Frederking

Language Technologies Institute, School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., PA 15213, USA
{rnshah,bolin,kdelaros,anatoleg,ref}@cs.cmu.edu

ABSTRACT

In this paper, we consider the problem of cross-document co-reference (CDC). Existing approaches tend to treat CDC as an information retrieval based problem and use features such as TF-IDF cosine similarity to cluster documents and/or co-reference chains. We augmented these features with features based on biographical attributes, such as occupation, nationality, gender, etc., obtained by using semi-supervised attribute extraction models. Our results suggest that the addition of these features boosts the performance of our CDC system considerably. The extraction of such specific attributes allows us to use features, such as semantic similarity, mutual information and approximate name similarity which have not been used so far for CDC with traditional bag-of-words models. Our system achieves $F_{0.5}$ scores of 0.82 and 0.81 on the WePS-1 and WePS-2 datasets, which rival the best reported scores for this problem.

1. INTRODUCTION

The problem of co-reference resolution deals with extracting all noun phrases from a document (names, descriptions, pronouns), and clustering them according to the real-world entity they describe (each such cluster is referred to as a 'chain'). This problem is much harder than named entity extraction, because it requires sophisticated parsing techniques to correctly identify all noun phrases, and a deep semantic analysis of the document to cluster these mentions correctly. Most research in this field has focused on the problem of within-document co-reference (WDC), where all the noun phrases are extracted from the same document. However, there has been some recent research on cross-document co-reference (CDC), where the task is to cluster noun phrases from all the documents in a collection. This problem is harder than within-document co-reference for the following reasons:

- The same name may refer to different entities in different documents,
- Analogously, different names in different documents may refer to the same real-world entity

In particular, the Web People Search (WePS) series of evaluations has focused on a subset of cross-document co-reference, where the task is to cluster entire documents according to the person they refer to, rather than individual clusters of mentions within documents. Even with this additional constraint, the problem still retains many of the difficulties described above, and it is hoped that a method that solves the WePS task successfully would perform well for the general CDC task as well. Since the WePS data is the standard data used to report results for this task, we shall use it as well.

The remainder of the paper is organized as follows: In Section 2, we describe the approaches by the winning systems at WePS and previous work on the CDC problem. Section 3 describes the datasets we used. Section 4 illustrates the architecture of our baseline system. In Section 5 we describe our attribute extraction models, and in Sections 6 and 7 we describe our augmented system and overall results respectively. Finally, in Section 8, we present our conclusions.

2. RELATED WORK

[4] published an influential early paper on CDC that merged co-reference clusters (or "chains") across documents using a bag of words representation and chain similarity. [16] suggested an improved approach to person name disambiguation based on augmenting bag of words features with automatically extracted biographic information. [11] used a Maximum Entropy model to disambiguate entities with the same name and different descriptions.

These initial efforts suffered from the lack of standard datasets to conduct tests. The WePS evaluations [1], [2] and [3] for person name disambiguation have sought to address this problem. There have been 3 previous iterations of

WePS. The winning system in WePS-1 [7] and WePS-2 [8] used TF-IDF cosine similarity between documents as a distance metric and performed Hierarchical Agglomerative clustering. The next two systems in WePS-2 [5], [13] also used variants based on the standard TF-IDF model (such as proper noun TF-IDF only, or P-LSA, which can be thought of as a generative version of TF-IDF). The winning system in WePS-3 [15] did attempt to use features based on attributes of the persons being described in the document, but couldn't report significant benefits from doing so. By comparison, we show significant improvement from the use of such features ($F_{0.5}$ score improves from .77 to .81), and report results that are slightly better than the best systems in WePS-1 and WePS-2 respectively. We don't report scores over the WePS-3 test set since it is not publicly available yet. Also related to our work are semi-supervised information extraction (IE) systems, such as NELL [6] and KnowItAll [10] but these are focused on growing knowledge bases as opposed to solving the specific CDC problem and their accuracy is well below dedicated CDC systems.

3. DATASETS

The WePS datasets were prepared by selecting person names from different sources, such as the US Census, Wikipedia, etc., and collecting web pages for different people having the same name. We trained our system on the WePS-1 training set, which consisted of 49 names, an average of 10 entities per name, and 100 web pages per entity. We tested our system on the WePS-1 test set (17 names, 45 entities per name, 100 pages per entity) and WePS-2 test set (25 names, 18 entities per name, 100 pages per entity).

4. BASELINE SYSTEM ARCHITECTURE

We now describe the architecture of our baseline CDC system. Given a name and a pair of web pages that both refer to some person with that name, we extract the following similarity features between the web pages:

- Bag of words based features: These include TF-IDF cosine similarity between vectors of all words in the documents, TF-IDF cosine similarity between vectors of only named entities in each document, and approximate string matching based analogues of these features such as SoftTFIDF similarity [9] between the vectors of each document. SoftTFIDF based features are robust to factors such as variations in name spellings and typographical errors.
- Link based features: We exploit the web structure of the documents and use the number of links in common between documents, or the presence of a link from one of the documents to another as an additional class of features to calculate similarity between documents.
- Topic based similarity: We also use Latent Dirichlet Allocation to model all web pages for a name as being generated from a topic model, with each entity corresponding to a topic, and calculate the KL divergence between topic distributions of two documents.

We then use all these features to train a support vector machine (SVM) based classifier that classifies a pair of documents that refer to the same name as referring to the same *entity* or not. We use the distance from the negative class boundary produced by this SVM as a distance metric between two documents, and then perform density clustering on the documents for each name.

We measure the performance of our system using the standard co-reference metrics B-cubed precision and recall. These metrics have an advantage over traditional clustering metrics such as purity and inverse purity since they account for singleton clusters and are hence harder to game.

5. ATTRIBUTE EXTRACTION MODELS

We use five different attribute extraction models, described in [12]:

- Contextual Pattern-Based Model (Context)
- Absolute and Relative Document-Position-Based Model (ADPB and RDPB)
- Transitivity-Based Model (Trans)
- Latent Wide-Document-Context Model (Latent)

These are semi-supervised models because we initially train them using a few seed examples and then iteratively label new examples and retrain. We extract attributes such as a person's date of birth, date of death, gender, occupation and nationality. In addition, we use a round-robin back off model to combine the results of these models.

5.1. Contextual Pattern Based Model

The contextual pattern-based model is a standard method for extracting biographical facts, first described in [17]. This method uses contextual patterns such as <Name> "is a" <Nationality> "football player". In this approach, the probability of a relationship *r*(Name, Attribute Value) is approximated by the probability of the occurrence of the corresponding contextual patterns in training data. We also include an important improvement made to this model by using partial un-tethered contextual patterns [12].

5.2. Document Position Based Models

The absolute and relative document position-based models (ADPB and RDPB) make use of the observation that certain biographical attributes tend to appear in characteristic positions in biographic texts. For example, the birth-date of an individual typically occurs near the beginning of a Wikipedia page.

5.3. Transitivity Based Model

The transitivity-based model is an implicit model which uses the neighboring person's attribute value to predict the target person's attribute value. It is based on the intuition that for certain attributes such as occupation, a person is likely to appear in the same document as other people who have the same attribute value. For example, Michal Jordan (the basketball player) is more likely to appear in a document with other basketball players, such as Kobe Bryant and Dirk Nowitzki, as opposed to professors.

5.4. Latent Document-Context Model

The latent document-context model predicts attributes using a topic modeling approach. For example, a biographic document containing terms like "songs", "albums", or "recorded" all collectively indicate that the person being the discussed in the document is a singer or some sort of musician. One advantage of this model is that it can be used to detect an attribute value that may not be explicitly mentioned in the article text.

5.5. Combined Model

We combine these models using a round-robin-based back-off model on each attribute where it is applicable. Given that each model produces a list of possible candidate values for the attribute, the back-off model takes the most probable value from each of the five lists in a specific order (based on the models' performance in training) and then repeats this process with the next most probable value from each list until it reaches the number of required values.

Model	DOB	DOD	Gen	Occ	Ntl
Baseline	.23	0	.76	.23	.57
Combined	**.38**	**.09**	**.95**	.28	**.95**
Context			**.95**	.28	
ADPB	.23	.04	.85	0	.47
RDPB	**.38**	**.09**	.80	**.43**	**.95**
Trans				.38	

Table 1: Top 1 extraction accuracy for each attribute extraction model for birth date (DOB), death date (DOD), gender (Gen), nationality (Ntl) and occupation (Occ).

Table 1 shows the performance of our models on the test data set, when only the top extracted value per model is considered. The latent model's results were omitted because the model did not perform well in our experiments. The baseline extractor selects the most frequent value that matches the attribute domain model for a given document.

We should note here that during evaluation, we also marked instances where the attribute simply did not exist in the testing data as incorrect, therefore our accuracy results are not very high. However, when the attributes do exist in a document, our back-off model is able to extract them with high precision. The back-off model retains the best accuracy for most of the attributes and consistently outperforms the baseline extractor. We also find that the implicit models (Trans and Latent) do not work well, perhaps due to our limited training set. In contrast, the relative position based model is the best model.

6. IMPROVED CDC SYSTEM

The extraction of the above mentioned attributes thus allows us to use features that would not have been possible to obtain with simple bag-of-words representations of documents. We add the following features to our CDC system based on the above mentioned attributes: Exact match between dates of birth and dates of death (if found), exact match between genders (if found), SoftTFIDF similarity between occupations, SoftTFIDF similarity between nationalities, and semantic similarity [14] between occupations. If an attribute is not found in one or both documents, we assign default values using averages for those attributes from our training set.

7. RESULTS

Table 2 shows the performance of our baseline and improved CDC systems as well as the top performing system in WePS-1 over the WePS-1 dataset.

System	P	IP	$F_{0.5}$
Improved	**.84**	.80	**.82**
Baseline	.79	.80	.79
Chen et al., 2007 [7]	.72	**.88**	.78

Table 2: Performance of our system and other systems over the WePS-1dataset

System	Pre	Rcl	$F_{0.5}$
Chen et al., 2009 [8]	.87	.79	**.82**
Improved	.84	.78	.81
Balog et al., 2009 [5]	.85	**.80**	.81
Ikeda et al., 2009 [13]	**.93**	.73	.81
Baseline	.78	.77	.77
Long and Shi, 2010 [15]	-	-	.73
Romano et al., 2009 [18]	.82	.66	.72

Table 3: Performance of our system and other systems over the WePS-2 dataset

Table 3 shows the performance of our system, the top four WePS-2 systems and the top WePS-3 system over the WePS-2 dataset. Note that WePS-1 used Purity (P), Inverse Purity (IP) and their $F_{0.5}$ score as evaluation metrics while WePS used B-cubed precision (Pre) and recall (Rcl) and their $F_{0.5}$ score.

We found that our system shows significant improvements with the addition of attribute-based features, even though there are many documents which do not contain the desired attributes at all. Our final system does better than the best WePS-1 system on the WePS-1 dataset. It also rivals the highest scoring WePS-2 systems over the WePS-2 dataset. Also, the winning WePS-3 system reported their scores on the WePS-2 dataset and our system performs better on that dataset.

8. CONCLUSION AND FUTURE WORK

We have shown that the application of semi-supervised information extraction (IE) methods can aid the performance of cross-document co-reference (CDC) systems as opposed to using bag-of-words models alone, and it can also allow the use of more diverse types of features. This holds true even when the attribute extraction methods do not offer great recall. We presented a CDC system that rivals the state-of-art systems over standard datasets. In the future, we plan to explore ways of increasing the recall of our attribute extraction models without hurting precision, as well as exploring the use of other types of features based on attributes.

9. ACKNOWLEDGMENTS

We would like to thank Dr. William Cohen at Carnegie Mellon University for his support and guidance.

10. REFERENCES

[1] E. Amigó, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo, "WePS-3 Evaluation Campaign: Overview of the On-line Reputation Management Task," in Proc. Conference on Multilingual and Multimodal Information Access Evaluation (CLEF), 2010.

[2] J. Artiles, J. Gonzalo, and S. Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). ACL, 2007

[3] Javier Artiles, Julio Gonzalo and Satoshi Sekine. “WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task,” In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

[4] 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 the 17th International Conference on Computational Linguistics (COLING-ACL’98), pages 79-85.

[5] K. Balog, J. He, K. Hofmann, V. Jijkoun, C. Monz, M. Tsagkias, W. Weerkamp, and M. de Rijke. The university of amsterdam at weps2. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr. and T.M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In Proceedings of the Conference on Artificial Intelligence (AAAI), 2010

[7] Ying Chen and James H. Martin, “CU-COMSEM: Exploring Rich Features for Unsupervised Web Personal Name Disambiguation,” In Proceedings of Semeval 2007, Association for Computational Linguistics, 2007.

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

[9] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. In Proc. of IJCAI Workshop on Information Integration on the Web

[10] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld and A. Yates, Web-scale information extraction in KnowItAll, Proceedings of the 13th International World Wide Web Conference (WWW-04), New York (2004), pp. 100–110.

[11] Michael B. Fleischman and Eduard Hovy 2004. Multidocument person name resolution. In Proceedings of ACL-42, Reference Resolution Workshop.

[12] Nikesh Garera and David Yarowsky (2009). Structural, Transitive and Latent Models for Biographic Fact Extraction. Proceedings of the 12th Conference of the European Chapter of the ACL.

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

[14] J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In

International Conference Research on Computational Linguistics (ROCLING X), September 1997

[15] C. Long and L. Shi. Web person name disambiguation by relevance weighting of extended feature sets. In Third Web People Search Evaluation Forum (WePS-3), CLEF 2010

[16] Gideon S. Mann and David Yarowsky 2003. Unsupervised Personal Name Disambiguation In Proceedings of the seventh conference on Natural language learning at HLT-NAACL, pages 33-40.

[17] D. Ravichandran and E. Hovy. Learning surface text patterns for a question answering system. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), 2002.

[18] L. Romano, K. Buza, C. Giuliano, and L. Schmidt-Thieme. Xmedia: Web people search by clustering with machinely learned similarity measures. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.