DCProposal: Online Unsupervised Coreference Resolution for Semi-Structured Heterogeneous Data*

Jennifer Sleeman
Department of Computer Science
University of Maryland Baltimore County
jsleem1@umbc.edu

June 23, 2012

Abstract.

A pair of RDF instances are said to corefer when they are intended to denote the same thing in the world, for example, when two nodes of type foaf:Person describe the same individual. This problem is central to integrating and inter-linking semi-structured datasets. We are developing an online, unsupervised coreference resolution framework for heterogeneous, semi-structured data. The online aspect requires us to process new instances as they appear and not as a batch. The instances are heterogeneous in that they may contain terms from different ontologies whose alignments are not known in advance. Our framework encompasses a two-phased clustering algorithm that is both flexible and distributable, a probabilistic multidimensional attribute model that will support robust schema mappings, and a consolidation algorithm that will be used to perform instance consolidation in order to improve recall measures over time by addressing data spareness.

Keywords: Coreference Resolution, Instance Matching, Heterogeneous Data, Unsupervised Machine Learning, Semantic Web, Online Algorithms

1 Introduction

Coreference resolution is widely researched both from the computational linguistics perspective and from the knowledge representation perspective. Germaine to this discussion is research related to knowledge representation. From this perspective, when performing coreference resolution, one tries to determine if an instance represents a real-world entity, typically defined in a knowledge base. Various techniques have been

*Advisor: Tim Finin

used to perform coreference resolution including both supervised and unsupervised methods. Most algorithms designed to perform coreference resolution assume a complete data set, many assume there is some knowledge of the schemas used a priori and often the topic of heterogeneity is neglected. In many complex computing environments, particularly among scientific and intelligence communities, data schemas may not be known a priori, data is more typically acquired over time in parts rather than all at once and often heterogeneous, i.e. originating from multiple sources. In order to support these complexities, coreference resolution algorithms need to account for this online behavior and need to support heterogeneous data. Furthermore, very little focus is given to using the coreferent instances to further improve subsequent matching. Particularly when working with sparely defined instances, object consolidation, i.e., the merging of groups of entities connected by conference relations, can increase recall over time.

Current methods for performing coreference resolution typically use an offline model that assumes the data to be processed is complete. One approach is to use supervised learning [1]. Our previous work [2, 3] performed coreference resolution for Friend of a Friend (FOAF) ontology instances using a supervised method, namely Support Vector Machines [4]. While supervised learning methods can produce results that are reasonably accurate, many of these methodologies do not consider heterogeneous data and are harder to adapt to an online model. Supervised classification methods require a set of labeled training examples to train the classifier, and acquiring the labeled data can be expensive, especially if human judgments are necessary. With heterogeneous data composing a set of training documents that is representative of the mix of ontology terms used can be a challenge, which could lead to inaccuracies in the classification process.

Static context environments are not typical, particularly among scientific, biomedical and intelligence domains, which are more likely to be streaming. Work by Gama et. al [5] describes supervised methods as having a tendency to be tested and used for static data models. Streaming models typically require faster response time [6] and are typically associated with larger data sets. The supervised model would quickly become inaccurate as the training data would not be representative of the space. In order for these methods to support an online model, they would have to retrain in-process or perform other modifications to support this type of model.

Heterogeneous data is not uncommon in complex computing environments, particularly those that use Semantic Web technology. For example, linked data [7], a web of datasets that is linked together and shared, has become a common means for making data available for others on the web. Linking entities allows information across sources to be combined and needs to be performed automatically to accommodate the scale of current and future LOD collections [8]. There is a strong need for a flexible coreference resolution solution that could provide essential mappings between entities.

Given the problem of online coreference resolution for heterogeneous data, an unsupervised or semi-supervised learning approach is required to support the dynamic nature of such an environment; in particular we will show that a two-phased clustering algorithm and knowledge base reasoning will provide both a flexible and scalable way to support

this model with accuracy rates that approach supervised and offline methods.

2 Related Work

Existing research that addresses heterogeneous data uses various approaches. Volt et al. [9] propose a framework called Silk that supports generating owl:sameAs links for linked data, they support distributed environments, and use aggregation functions for similarity scores. Seddiqui et al.[10] describes a process of using anchors, described as 'lookalike' concepts to perform instance matching. Work by Araujo et al.[11] includes supporting instance matching specifically for interlinking data sets within the Linked Open Data Cloud. Based on a two stage process that includes string matching for selection and disambiguation. Hu et al. [12] developed a coreference resolution process that generates a kernel based on the OWL vocabulary and ranks coreferent pairs based on confidence measures. Nikolov et al. [8] describes a schema-level approach to support their previous work [13] that specifically addressed instance level matching. They use an outside knowledge base to support coreference resolution and schema-level mappings that are both fuzzy in nature and overlapping. They also use instance level coreference knowledge from other repositories to support their schema level mappings. Work by Yatskevich et al. [14] combines semantic web and natural language processing to perform cross-document coreference resolution. Using "Similarity Flooding" [14] to compare instance graphs. Nikolov et al. [15] also address the issue of automatic instance linking for linked data in support of schema matching. They use instance level links to infer schema-level relations and use the schema-level mappings for instance matching, building on their earlier work [8]. Of the work described that uses schema analysis, many have shortcomings because they use at most two types of attribute-based analysis. We address these shortcomings by performing attribute analysis using five different dimensions.

There is very little research related to online coreference resolution. A large majority of research uses hierarchical approaches to cluster data in streaming environments, which is the approach we take to support online coreference resolution. Our work is inspired by Rao et al. [16] which highlights a cross document coreference resolution approach for streaming data. They use a streaming clustering algorithm based on a doubling clustering algorithm that has two stages, an update stage and a merge stage. Upon receiving a stream of extracted coreference chains and types, they perform intra-document coreference resolution. For an entity chain they choose an appropriate cluster based on similarity scores.

Research related to instance consolidation has a tendency to use a methodology that relies upon inverse functional properties. There are three main works that address object consolidation as it relates to instance matching. Hogan et al. [17] use inverse functional properties to determine instances in common and rewrite identifiers based on each equivalence chain. Shi et al. [18] describe object consolidation as 'smushing' and performs 'smushing' by taking advantage of the inverse functional property. They work at the attribute level and calculate attribute level similarity measures. Yatskevich

et al. [14] address consolidation of graphs by merging graphs if the instances belong to the same class, and if their string similarity is higher than a threshold. None of this research addressed conflict resolution or used the consolidated instances to increase the number of coreferent pairs. We have seen in our previous work, when working with sparse input, consolidation can improve recall.

3 Approach

Our research makes four major research contributions that work together to achieve an effective approach. We describe a probabilistic multi-dimensional attribute model and attribute mapping. The attribute model and attribute mapping enables us to perform attribute analysis that can improve over time. We also describe a new two-phased clustering algorithm that will be used to support online coreference resolution. We describe a new algorithm that will perform instance consolidation that will improve recall over time. Finally we highlight a new coreference resolution benchmark that we will develop as part of our research.

Research Contribution: Multi-dimensional Model: We are developing a probabilistic multi-dimensional attribute model to address the problem of heterogeneous data by deriving meaning from the data and schemas using various types of analysis. Previous research specifically addressing linked data used at most two types of attribute-based analysis that included string matching functions [14, 12, 19, 11], graph-based functions [14], contextual methods [10] and/or schema level mappings [8, 15]. Our previous work related to FOAF coreference resolution [2, 3] used both string similarity and ontological axioms. Using the ontological axioms provided a quick way to assert coreference for a limited number of pairs. String similarity was used to support a Support Vector Machine (SVM) classifier. The classification process was clearly slower than the rules-based approach. In my previous work related to attribute alignment [20] we were able to show competitive F-Measure scores (55%) when using the statistical distribution as a way to perform attribute alignment.

In Figure 1, we show five dimensions for attribute analysis. Dissimilarity and similarity metrics performs comparisons between attribute values both at the individual pair level and across vectors. For example, if we are comparing two attributes that represent a person's name, we would likely use a distance metric to determine how dissimilar the two strings are to each other. Structural properties takes into consideration the graph itself. Graph matching algorithms may be used to determine similarity among graphs. Statistical properties involve analytics that use knowledge of the distribution of values for an attribute. Ontological definition uses axioms defined in the ontology. For example if an attribute is inverse functional. Contextual information provides macro-level information that supports conceptual heterogeneity. Using neighborhood graphs is one way to develop context.

Attribute values vary based on data type. For example, an attribute can consist of *Days of the Week (Mo,Tu,We,Th,Fr,Sa,Su)* or *textual names (Joe, Bob, George)*. They can

be binary *(Male,Female)* or represent long blocks of text. In our previous work [20], we sampled attribute level data to distinguish categorical vs. non-categorical data and used a specific kernel based on this sampling. We saw a significant improvement in F-Measure scores by making this distinction. Distinguishing among attribute types can improve accuracy when using dissimilarity functions. For example, we could measure the distance between two geographic locations using a Euclidean distance [21] rather than using a distance metric that calculates the number of transitions from one string to another such as Levenshtein [22].

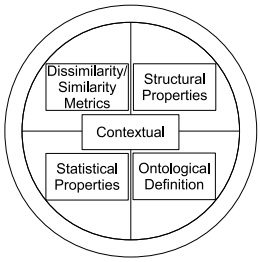

Figure 1: Attribute Dimensions

Research Contribution: Two-Phased Clustering: We are developing a new algorithm that performs two-phased clustering. The first phase acts as a filter and the second phase performs hierarchical clustering. Based on previous research, the hierarchical clustering method has advantages over partitional methods because it does not assume the data distribution, and centroids do not have to be defined a priori [23]. The complexity of clustering algorithms can range from $O\left(n^2\right)$ to $O\left(n^3\right)$. A first phase clustering that is computationally less expensive can reduce the size of the data that must be partitioned by the hierarchical clustering algorithm as shown in previous work using a similar approach [24]. The first phase, as depicted in Figure 2, acts as a filter, partitioning instances into clusters. This model is captured in a structure that can be used on a graph by graph basis. The second phase of clustering is applied to each partition as depicted in Figure 2. By partitioning instances in such a way, the second phase clustering could be distributed across computing resources. From a streaming data perspective, not all clusters need to be evaluated when a new instance is processed (using a greedy clustering method), and therefore not all instances must be compared to every other instance, improving the overall efficiency.

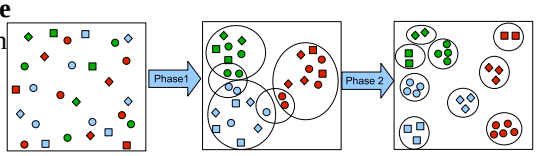

Figure 2: Two-Phased Clustering

Research Contribution: Instance Consolidation: Typically research in the area of coreference resolution focuses on the resolution aspect only and does not address consolidation. Our previous work related to coreference resolution of FOAF instances [2, 3] tested using a cluster-based model to perform consolidation of coreferent pairs. We used these clusters to increase the number of future coreferent pairs. Pairs designated as coreferent formed new clusters which were then evaluated as part of future coreference resolution. Our research showed that subsequent pairing did occur with coreferent clusters. Our approach builds on these ideas, with a consolidation algorithm that addresses the following: clustering of coreferent instances with the ability to uncluster, merging feature sets of coreferent cluster entities, and unmerging feature sets of coreferent

cluster entities.

Research Contribution: Coreference Resolution Benchmark: As part of this work, we plan to develop a coreference resolution benchmark for the Semantic Web. We will extract files from various RDF datasets with the goal of creating a data set that is rich with coreferent pairs and pairs that would be mistakenly asserted as coreferent due to common values for dominate attributes. This benchmark will be developed such that it can be shared with the research community. One of the most challenging problems related to testing coreference resolution systems is finding data that has enough positive test cases to formulate a valid test. By providing such a data set, we believe this will provide an invaluable resource to the research community.

4 Evaluation

Given the multi-dimensional attribute model, we want to prove that this method both supports heterogeneous data and improves precision when compared with other approaches that use up to two dimensions. The Ontology Alignment Evaluation Initiative (OAEI) offers a benchmark for instance matching that is used by other researchers in this domain and can be used for evaluation. In addition, we wish to show that attribute typing can also improve precision.

Our ultimate goal is to achieve high F-Measure scores when performing coreference resolution in an online environment. We will evaluate our two-phased clustering algorithm with respect to offline supervised methods as a way to show comparison F-Measure scores using both the OAEI data set and our custom data set. In addition, we will measure the effectiveness of this algorithms in how it can process data incrementally and over time. An important part of this problem is overcoming scalability issues particularly when processing large amounts of data. Part of the evaluation will address this requirement.

Coreferent pair consolidation can be measured by determining if the consolidated coreferent clusters increases recall, without decreasing precision given the pairs were not consolidated or consolidated using the standard approaches. For example, if we cluster two coreferent instances each having a sparse set of features that individually were not strong enough to match other instances but as a consolidated cluster match other instances, we show an increase in recall.

5 Conclusion

Instance matching algorithms need to address the complexities of today's computing environments. Data is noisy, heterogeneous in nature, incrementally processed, large and often based on schemas that are not known a priori. In order to support these complexities we are developing algorithms that work together under a common framework. We proposed a probabilistic multi-dimensional attribute model to address the aspects

of the data such as noisiness and heterogeneity. We proposed a two-phased clustering algorithm that supports an online model to address working with large data and data that is incrementally processed over time. Finally we proposed an instance consolidation algorithm that works within the context of an online model by improving matching over time and addressing data spareness.

References

[1] Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms using different performance metrics. Technical Report TR2005-1973 (2005)

[2] Sleeman, J., Finin, T.: A machine learning approach to linking foaf instances. In: Spring Symposium on Linked Data Meets AI, AAAI (January 2010)

[3] Sleeman, J., Finin, T.: Computing foaf co-reference relations with rules and machine learning. In: The Third International Workshop on Social Data on the Web, ISWC (November 2010)

[4] Joachims, T.: SVMLight: Support Vector Machine. University of Dortmund, http://svmlight.joachims.org/ (1999)

[5] Gama, J., Rodrigues, P.P., Castillo, G.: Evaluating algorithms that learn from data streams. In: the 24th Annual ACM Symposium on Applied Computing (SAC 2009), ACM Press (2009) 14961500

[6] Han, J., Kamber, M.: Data mining concepts and techniques. Elsevier (2006)

[7] Bizer, C.: The emerging web of linked data. IEEE Intelligent Systems **24**(5) (2009) 87–92

[8] Nikolov, A., Uren, V., Motta, E., Roeck, A.: Overcoming schema heterogeneity between linked semantic repositories to improve coreference resolution. In: A.Gomez-Perez and Y. Yu and Y. Ding eds.: The SemanticWeb, Fourth Asian Conference ASWC 2009. Volume 5926., Spring 2009 (December 2009) 332346

[9] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk - a link discovery framework for the web of data. In: Proc. 2nd Workshop on Linked Data on the Web, Madrid, Spain (April 2009)

[10] Seddiqui, M., Aono, M.: Ontology instance matching by considering semantic link cloud. In: 9th WSEAS International Conference on Applications of Computer Engineering. (2010)

[11] Araujo, S., Hidders, J., Schwabe, D., de Vries, A.P.: Serimi resource description similarity, rdf instance matching and interlinking. In: CoRR. Volume Vol. abs/1107.1104. (2011)

[12] Hu, W., Qu, Y., Sun, X.: Bootstrapping object coreferencing on the semantic web. Journal of Computer Science and Technology **26(4)** (2011) 663–675

[13] Nikolov, A., Uren, V., Motta, E., de Roeck, A.: Integration of semantically annotated data by the knofuss architecture. In: EKAW 2008. (2008)

[14] Yatskevich, M., Welty, C., Murdock, J.: Coreference resolution on rdf graphs generated from information extraction: first results. In: the ISWC 06 Workshop on Web Content Mining with Human Language Technologies. (2006)

[15] Nikolov, A., Uren, V., Motta, E.: Data linking: Capturing and utilising implicit schema level relations. In: International Workshop on Linked Data on the Web. (2010)

[16] Rao, D., McNamee, P., Dredze, M.: Streaming cross document entity coreference resolution. In: International Conference on Computational Linguistics (COLING), Coling 2010 Organizing Committee (November 2010) 1050–1058

[17] Hogan, A., Harth, A., Decker, S.: Performing object consolidation on the semantic web data graph. In: Proc. I3: Identity, Identifiers, Identification. Workshop at 16th Int. World Wide Web Conf. (February 2007)

[18] Shi, L., Berrueta, D., Fernandez, S., Polo, L., Fernandez, S.: Smushing rdf instances: are alice and bob the same open source developer? In: Proc. 3rd Expert Finder workshop on Personal Identification and Collaborations: Knowledge Mediation and Extraction, 7th Int. Semantic Web Conf. (November 2008)

[19] Song, D., Heflin, J.: Automatically generating data linkages using a domain-independent candidate selection approach. In: International Semantic Web Conference. (2011)

[20] Sleeman, J., Alonso, R., Li, H., Pope, A., Badia, A.: Opaque attribute alignment. In: Proceedings of the 3rd International Workshop on Data Engineering Meets the Semantic Web. (2012)

[21] E.Weisstein: Distance. From MathWorld–A Wolfram Web Resource (1999-2012) Accessed May 2012.

[22] Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. **10**(8) (1966) 707–710

[23] Rodriguess, P.P., Pedroso, J.P.: Hierarchical clustering of time series data streams. IEEE Transactions on Knowledge and Data Engineering **20**(5) (May 2008) 615–627

[24] McCallum, A., Nigam, K., Ungar, L.: Efficient clustering of high-dimensional data sets with application to reference matching. In: The Sixth International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD (2000) 169–178