Online Unsupervised Coreference Resolution for Semi-structured Heterogeneous Data*

Jennifer Sleeman

Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County, Baltimore, MD 21250 USA
jsleem1@cs.umbc.edu

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 accuracy rates over time by addressing data spareness.

1 Introduction

When performing coreference resolution, as it relates to knowledge representation, one tries to determine if an instance represents a real-world entity, typically defined in a knowledge base. Various techniques have been used to perform coreference resolution including both supervised and unsupervised methods, however many approaches tend to function based on a batch data set, assume the schemas are accessible a priori and often neglect the topic of heterogeneity. 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 the effects of temporal object consolidation, i.e., the merging of groups of entities over time, connected by coreferent relations.

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.

* Advisor: Tim Finin.

P. Cudré-Mauroux et al. (Eds.): ISWC 2012, Part II, LNCS 7650, pp. 457–460, 2012.

2 Related Work

Though there is a significant amount of research in this area[15,14,10,8,7], we highlight a few more recent works. Araujo et al.[1] support instance matching specifically for interlinking data sets within the Linked Open Data Cloud. This work is consistent with others in that it assumes a static environment. Hu et al. [4] uses language axioms to generates a kernel based on the OWL vocabulary and ranks coreferent pairs based on confidence measures. Using language axioms can be a limitation, often data does not strictly conform to language axioms and in many cases, schemas are not accessible. In our previous work[12] only a small portion of our data contained axioms that could be used for this type of analysis. Rao et al.[9] highlight a cross document coreference resolution approach for streaming data that uses a clustering algorithm based on a doubling clustering algorithm which is similar to our approach; we however use a two-phased approach to clustering to reduce the computational costs. Song et al.[13] describe an approach to candidate selection that learns attributes that occur most frequently across their data set and a matching algorithm to designate coreferent pairs. Though supportive of heterogeneous data, the candidate selection process is limited by the key designation which could underperform when working with sparse data. It is also not clear how this approach could support temporal changes. Both Hogan et al.[3] and Shi et al.[11] do not address conflicts and rely upon inverse functional properties to perform object consolidation, which could be problematic since inverse functional properties are not always present. Our work does not rely on inverse functional properties, we address conflicts and we are specifically evaluating how consolidated instances will improve the accuracy of subsequent coreference resolution over time.

3 Approach

Our research makes four major research contributions that work together to achieve an effective approach to perform online coreference resolution. We will build a system that will bring together these contributions.

Research Contribution: Multi-dimensional Model: We are developing a probabilistic multi-dimensional attribute model that will support heterogeneous data by deriving meaning from the data and schemas using five dimensions. Dissimilarity and similarity functions are used to compare 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 function to determine how dissimilar the two strings are to each other. Structural properties take into consideration the graph itself. Statistical properties involve analytics that use knowledge of the distribution of values for an attribute. Ontological definitions use axioms defined in the ontology. Contextual information provides macro-level information that supports conceptual heterogeneity, for example using neighborhood graphs.

We are currently experimenting with a Bayesian model to represent these five dimensions. We are implementing this model to support our second phase of clustering to determine which instances should be part of the same cluster, rather than using a single distance measure. We also use attribute mapping to classify attribute types for

subsequent processing and for specializing the five dimensions. As the attribute model is used over time, we plan to develop optimal models based on data types. For example, we could measure the distance between two geographic locations using a Euclidean distance [2] rather than using a distance function that calculates the number of transitions from one string to another such as Levenshtein [5].

Research Contribution: Two-Phased Clustering: We are developing a new clustering algorithm that performs clustering in two phases. The first phase acts as a filter resulting in neighborhoods of related instances and the second phase performs the clustering of coreferent instances. 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 second phase of clustering, as shown in previous work using a canopy approach [6]. We are building the first phase to work at a complexity under $O\left(n^2\right)$ that will roughly partition instances into neighborhoods of likeness. Currently we use a bag of words model and a canopy-like approach [6]. The second phase of clustering is applied to each partition and will use our defined attribute model to perform coreferent-based clustering of each neighborhood cluster. Currently we use agglomerative hierarchical clustering with distance metrics only, and we are developing our new algorithm to support the integration of our attribute model.

Research Contribution: Instance Consolidation: In our model, to support temporal changes, the concept of an instance is abstractly defined as a single instance or a cluster of instances that are coreferent. Given our two-phased clustering work, the results are clusters where in each cluster, we symbolically link instances using a weighted measure to allow for cluster changes over time. Features among instances are weighted in order to support subsequent instance matching using dominate cluster features. We are currently experimenting with a number of feature reduction algorithms to support subsequent instance matching.

Research Contribution: Coreference Resolution Benchmark: A challenging problem related to testing coreference resolution systems is finding data that has enough positive test cases to formulate a valid test. For this reason we are developing a set of Semantic Web coreference resolution benchmarks that could be shared with the research community. The benchmarks will exercise the coreference resolution algorithm from different perspectives.

4 Evaluation

We will evaluate our clustering algorithm with respect to offline supervised methods as a way to show comparison F-Measure scores using both the Ontology Alignment Evaluation Initiative (OAEI) data set and our custom data sets. In addition, we will measure the effectiveness of this algorithm and how it can process data incrementally over time. We will also evaluate the effectiveness of using both attribute typing and a probabilistic model by performing precision and recall comparisons. We will evaluate consolidation by determining if the consolidated clusters improve the accuracy of the system over time.

5 Conclusion

Data is noisy, heterogeneous in nature, incrementally processed, large and often based on schemas that are not known a priori. To support these complexities we are developing algorithms that work together under a common framework including a probabilistic attribute model to address the aspects such as noisiness and heterogeneity, a two-phased clustering algorithm that supports an online model to address working with data that is incrementally processed over time and an instance consolidation algorithm that will improve matching over time and addresses data spareness.

References

1. Araujo, S., Hidders, J., Schwabe, D., de Vries, A.P.: Serimi resource description similarity, rdf instance matching and interlinking. CoRR, Vol. abs/1107.1104 (2011)
2. Weisstein, E.: Distance. From MathWorld–A Wolfram Web Resource (1999-2012) (accessed May 2012)
3. 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)
4. Hu, W., Qu, Y., Sun, X.: Bootstrapping object coreferencing on the semantic web. Journal of Computer Science and Technology 26(4), 663–675 (2011)
5. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals, vol. 10(8), pp. 707–710 (1966)
6. 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, pp. 169–178 (2000)
7. 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)
8. Nikolov, A., Uren, V., Motta, E., de Roeck, A.: Overcoming Schema Heterogeneity between Linked Semantic Repositories to Improve Coreference Resolution. In: Gómez-Pérez, A., Yu, Y., Ding, Y. (eds.) ASWC 2009. LNCS, vol. 5926, pp. 332–346. Springer, Heidelberg (2009)
9. Rao, D., McNamee, P., Dredze, M.: Streaming cross document entity coreference resolution. In: International Conference on Computational Linguistics (COLING). Coling 2010 Organizing Committee, pp. 1050–1058 (November 2010)
10. Seddiqui, M.H., Aono, M.: Ontology instance matching by considering semantic link cloud. In: 9th WSEAS International Conference on Applications of Computer Engineering (2010)
11. 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)
12. 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)
13. Song, D., Heflin, J.: Automatically Generating Data Linkages Using a Domain-Independent Candidate Selection Approach. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 649–664. Springer, Heidelberg (2011)
14. 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)
15. Yatskevich, M., Welty, C., Murdock, J.: Coreference resolution on rdf graphs generated from information extraction: first results. In: The ISWC 2006 Workshop on Web Content Mining with Human Language Technologies (2006)