Preprint of: Jennifer Sleeman and Tim Finin, Recognizing Entity Types in Heterogeneous Semantic Graphs, AAAI 2013 Fall Symposium on Semantics for Big Data, Arlington VA, Nov. 2013

# Entity Type Recognition for Heterogeneous Semantic Graphs

**Jennifer Sleeman and Tim Finin**
Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore. MD 21250 USA
{jsleem1,finin}@cs.umbc.edu

**Abstract**

We describe an approach to reducing the computational cost of identifying coreferent instances in heterogeneous semantic graphs where the underlying ontologies may not be informative or even known. The problem is similar to coreference resolution in unstructured text, where a variety of linguistic clues and contextual information is used to infer entity types and predict coreference. Semantic graphs, whether in RDF or another formalism, are semi-structured data with very different contextual clues and need different approaches to identify potentially coreferent entities. When their ontologies are unknown, inaccessible or semantically trivial, coreference resolution is difficult. For such cases, we can use supervised machine learning to map entity attributes via dictionaries based on properties from an appropriate background knowledge base to predict instance entity types, aiding coreference resolution. We evaluated the approach in experiments on data from Wikipedia, Freebase and Arnetminer and DBpedia as the background knowledge base.

## Introduction

Coreference resolution is the task of determining which instances in a collection represent the same real world entities. Without the use of filtering it is inherently an $O\left(n^2\right)$ process, though various techniques can be used to reduce this complexity (Mayfield et al. 2009; Sleeman and Finin 2010; Rao, McNamee, and Dredze 2010; Singh et al. 2011; Uryupina et al. 2011). When ontologies are not known or are not accessible, recognizing entity types can reduce the computation required since the number of instance pairs that must be checked is smaller. A closely related problem in information extraction is named entity recognition, the process of recognizing entities and their type (e.g., a person, location or organization) (Ratinov and Roth 2009; Nadeau and Sekine 2007).

When doing coreference resolution over RDF data or some other formalism, the entity types are usually explicitly given in a familiar ontology and their properties understood, enabling systems to reason about instance equality (Ferrara et al. 2008; Seddiqui and Aono 2010; Araujo et al. 2011). When this is not the case, i.e. when the ontologies are not accessible, not understood or several non-aligned ontologies are used, direct reasoning about instance equality is difficult, if not impossible. We believe that this situation will be common in many big data applications, where semantic annotations may be relatively simple. This is the problem we address in this paper.

A similar problem arises in work on database interoperability (Nottleman and Straccia 2007; Berlin and Motro 2002; Do, Melnik, and Rahm 2003) and ontology matching (Albagli, Ben-Eliyahu-Zohary, and Shimony 2012; Mitra, Noy, and Jaiswal 2005). In both, integrating heterogeneous data drawn from different repositories with different schemas is difficult simply because it is hard to establish that an attribute in one schema is the same (or nearly the same) as an attribute in another (Jaiswal, Miller, and Mitra 2010).

Linked open data (LOD) (Bizer 2009) has specifically addressed the issue of linking heterogeneous structured data in RDF to enable interoperability. In order to add an RDF dataset to a LOD collection, we represent the information as RDF and then link its elements (classes, properties and individuals) to known elements elsewhere in the collection. Though the "LOD cloud" collection has grown significantly, the total number of linked datasets is still relatively small (about 300) (Bizer, Jentzsch, and Cyganiak 2011) and the degree of interlinking often modest. Given the amount of data both available online and not available online, this number indicates that most repositories are still not linked to significant LOD collections and it is likely that these repositories use custom schemas.

## Related Work

Nikolov et al. (Nikolov, Uren, and Motta 2010) describe the problem of mapping heterogeneous data as it relates to coreference resolution, where often "existing repositories use their own schemas". They discuss how this makes coreference resolution difficult, since similarity evaluation is harder to perform when attribute mappings are unclear. They take advantage of linked data and knowledge of relationships between instances to support schema-level mappings. However, if a repository is not linked to an appropriate LOD collection, then this method is not feasible. We address this issue of custom schemas and their impact on coreference resolution by mapping attributes to a known set

of attributes for various entity types.

Early work by Berlin et al. (Berlin and Motro 2002) addressed the problem of database mapping using machine learning. Their Automatch system used machine learning to build a classifier for schema-matching using domain experts to map attributes to a common dictionary. The approach performed well, achieving performance exceeding 70% measured as the harmonic mean of the soundness and the completeness of the matching process. We build on this idea, using the dictionary mapping concept which we generate from DBpedia through a process guided by information gain.

Work by (Reeve and Han 2005) provides a survey related to semantic annotation, which is more closely related to our work. They describe and benchmark methods designed for unstructured text complemented with the output of information extraction tools to construct mappings. This differs from our approach in that we start from the graphs themselves without the raw text and information extraction data and metadata. This is a key distinction since using the graphs alone is more limiting. The benchmark compared various annotation tools using annotation recall and annotation precision, which we also will use to measure our entity typing performance.

Recent research by Suchanek et al. (Suchanek, Abiteboul, and Senellart 2012) describe their approach, PARIS, for aligning ontologies. This work uses string equality and normalization measures and also takes the approach of only using positive evidence. Again our goal was to be domain-independent, such that one could use a dataset to build the dictionary of types they wish to recognize then apply our mapping process to map to these dictionaries. We do not rely on an outside knowledge base to do so, but rather use techniques more akin to traditional named entity recognition to perform the task. This distinguishes our work from much of the ontology mapping research.

## Problem Definition

We address two problems in this work: how to recognize entity types without the context provided by surrounding text and how to perform schema-level mapping when working with heterogeneous data in the absence of information that would be obtained from an information extractor. We constrain this problem by trying to specifically group instances of type person, location and organization. While we do not address how this affects the overall coreference resolution, we can show reduced work when using entity type information as a pre-filter for instance matching evaluation.

**Definition 0.1** *Given a set of instances $I$, if a pair of instances is coreferent then, $coref(I_1, I_2)$. Given $I_1$ has a set of attributes $(a_1, a_2, ...a_n)$ where $a \in A$ and $I_2$ has a set of attributes $(b_1, b_2, ...b_n)$ and $b \in B$, then $similarity(A, B)$ is used to establish coreferent instances, where highly similar attributes sets would mean there is a higher likelihood of $coref(I_1, I_2)$.*

In order to reduce the number of instances that need to be evaluated, we try to establish the semantic type(s) of each instance. Since we do not know the meaning of $a$ or $b$, we try to map $a$ and $b$ to a common dictionary set. We do this by first generating a set of attribute dictionaries. For each entity type $person, location, organization$ we define a set of attributes $(p_1, p_2, ...p_n)$, $(l_1, l_2, ...l_n)$, $(o_1, o_2, ...o_n)$ that represent each type. We then use this information to determine if $person(I_1)|location(I_1)|organization(I_1)$ and $person(I_2)|location(I_2)|organization(I_2)$. This information can inform the coreference resolution algorithm as to whether evaluating $I_1$ and $I_2$ is necessary. Based on this mapping and labeled instances, we train a classifier to recognize which mappings belong to which entity types and then build a model that can be used to classify non-labeled instances.

## Methodology

Our method starts by building attribute dictionaries for each entity type and then uses a set of label training data to map the attributes of the instances to the attributes in the dictionaries. Using a set of mappers, we then score the mapping. This in turn becomes the features used to build a supervised model. We then perform the same mapping for unlabeled test data and classify these instances using the supervised model. The result of this process are that instances are classified as a person, location or organization.

Our approach includes a way to map attributes from different domains to a common set for people, locations and organizations. First using the DBpedia ontology (DBpedia 2013), we gather attributes appropriate for each type. We then supplement this list with attributes from other common ontologies. Currently we include attributes from the Friend of a Friend ontology (Brickley and Miller 2010). Then using the 2011 DBpedia infobox dataset (DBpedia 2011), we calculate information gain, a way to reduce entropy, for the various attributes. Given a sample set $S$ and attribute set $A$, the output is used for weighting attributes.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{1}$$

$$Entropy = -\sum_{i}^{N} p(x_i) log_2 p(x_i) \tag{2}$$

The concept of mapping to a common set of attributes is similar to research that relates to database mapping and ontology alignment(Nottleman and Straccia 2007; Berlin and Motro 2002; Mitra, Noy, and Jaiswal 2005; Albagli, Ben-Eliyahu-Zohary, and Shimony 2012). A selection of this work is discussed in more detail in Section . To map instance attributes to the common set we use a set of 'mappers': the first mapper analyzes the label names by using a Levenshtein (Levenshtein 1966) distance measure, the second uses WordNet(Miller 1995) to develop a better understanding of the labels and values and the third is a pattern mapper that looks at the structure of the value for each attribute.

For example, an email address follows a common pattern that can be recognized with a regular expression and used to find other email addresses even if the property label does

Table 1: Example of Attributes with low Information Gain.

| Attribute | | |
|---|---|---|
| image | othername | website |
| district | name | fullname |
| area | province | nickname |
| imagename | lat | religion |
| postcode | longitude | |

not specify or suggest "email". Each mapper outputs its own measure, that measure is used as feature in the feature vector. This information is then used by the Support Vector Machine (SVM) (Joachims 2002) to classify the instance. The classifier is used to predict the entity types of the test data.

## Experimentation

With all experiments, we randomly selected a subset of instances with an equal distribution of persons, locations and organizations except when working with the Arnetminer dataset (Tang, Zhang, and Yao 2007; Tang et al. 2008), which is a dataset about people. We tested using the Freebase (Bollacker et al. 2008) dataset using 2000 instances with 403 features, the Wikipedia data (Auer et al. 2007) using 3000 instances with 403 features and the Arnetminer dataset contains 4000 instances with 403 features however all of the instances are 1 class. We used Weka (Hall et al. 2009) to run the experiments.

We use the standard precision and recall metrics for measuring performance:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (3)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (4)$$

The first experiment was over data from the Freebase dataset and evaluated the general performance of the entity typing using ten-fold validation. With the second experiment, we examined the effects of filtering and how well the entity typing performed using one dataset for training (Freebase) and the other (Wikipedia) for testing. For the third experiment, we used a harder to classify dataset, one which has both sparse information related to the known attributes and also contains a lot of noisy information that could not be used. This dataset contains instances of people only but we think offers a good way to see how our entity typing would perform given noisy and sparse information.

## Evaluation

After calculating information gain using the DBpedia 2011 Infobox dataset, we weighted our attributes. Table 1 shows examples of attributes that had low information gain. In the first experiment, shown in figure 1, we performed a ten-fold validation using the Freebase dataset for each entity type. The location entity type was most successfully classified and organization had lower success.

The second experiment, as seen in figure 2, measured the effects of filtering on each type,using the Freebase dataset to build the model and the Wikipedia and Infobox dataset to test the model. In the case of the person entity type, we saw our filtering had a minor negative impact for the Wikipedia data set but in general, consistently we saw improvements when using filtering on various tests.

In the third experiment, as seen in figure 3, we tested the Arnetminer dataset against the model created by Freebase. The Arnetminer dataset contains only people and is harder to classify due to sparsity of the data. Often there is only an email or a first and last name. As we reduced the size of the model, the accuracy improved. We believe that this is simply due to the fact that the larger the dataset the higher chance of having more instances that are negative in common with these spares instances and suggests that possibly we need to consider additional mappers that consider the size of the graphs.

We tested our datasets using alternative classification techniques and found the SVM performed consistently well in comparison with the other classifiers and hence this confirmed that the SVM was a good choice for this particular problem. We experimented with non-linear kernels and, finding no improvement, used a linear kernel.

Our results compare favorably with those described in a survey (Reeve and Han 2005) in which six semantic annotation approaches were compared on a benchmark. They used an information extraction system to support their annotations and obtained F-measure scores ranging from 24.9% to 92.9%. Given this performance range and given that we are not supported by information extraction tools, we consider our results for this baseline approach to be encouraging.

## Conclusions

When heterogeneous data is in the form of a semantic graph and the schemas cannot be used, by mapping attributes to a common dictionary, entity typing can be performed for heterogeneous graphs. By implementing this method as a pre-filter to coreference resolution, the number of instances that need to be evaluated for coreferent relationships is reduced to strictly the instances of the same type. This typing paired with other filtering techniques can help with the $O\left(n^2\right)$ complexity of coreference resolution.

## References


Albagli, S.; Ben-Eliyahu-Zohary, R.; and Shimony, S. E. 2012. Markov network based ontology matching. *Journal of Computer and System Sciences 78.1 (2012)* 105–118.

Araujo, S.; Hidders, J.; Schwabe, D.; and de Vries, A. P. 2011. Serimi- resource description similarity, rdf instance matching and interlinking. In *CoRR*, volume 1107.1104.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. Dbpedia: a nucleus for a web of open data. In *Proc. 6th Int. Semantic Web Conf.*, 722–735. Berlin, Heidelberg: Springer-Verlag.

Berlin, J., and Motro, A. 2002. Database schema matching using machine learning with feature selection. In *Proceed-*

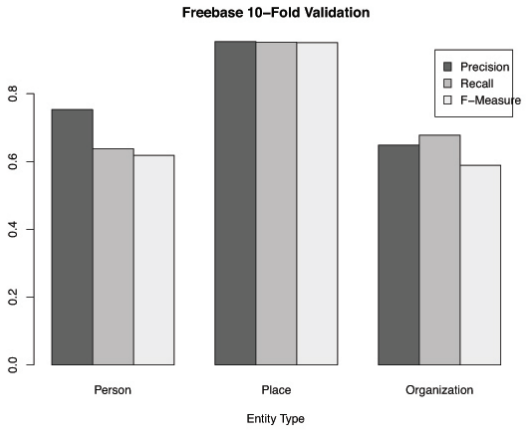


Figure 1: Freebase ten-fold Validation

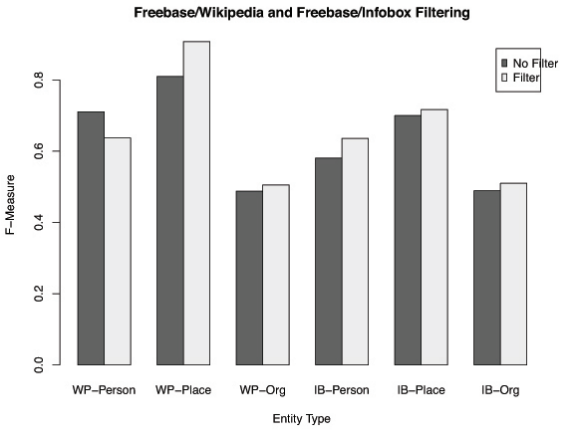


Figure 2: Effects of Information Gain Filtering

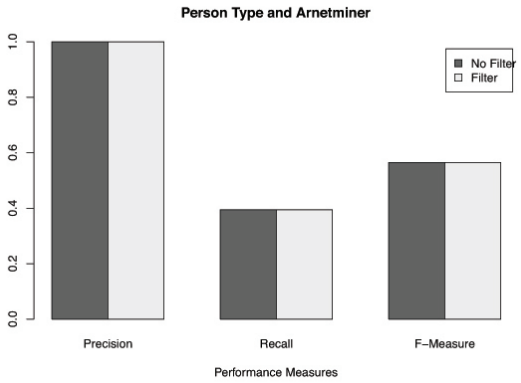


Figure 3: ArnetMiner Instances

*ings of the Conf. on Advanced Information Systems Engineering*, 452–466. Springer.

Bizer, C.; Jentzsch, A.; and Cyganiak, R. 2011. State of the lod cloud. http://lod-cloud.net/state/.

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

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. ACM Int. Conf. on Management of Data*, 1247–1250. ACM.

Brickley, D., and Miller, L. 2010. Foaf vocabulary specification .98. http://xmlns.com/foaf/spec/.

DBpedia. 2011. Dbpedia data set. http://dbpedia.org-/Datasets.

DBpedia. 2013. Dbpedia. http://dbpedia.org/ontology/.

Do, H.-H.; Melnik, S.; and Rahm, E. 2003. Comparison of schema matching evaluations. *Web, Web-Services, and Database Systems* 221–237.

Ferrara, A.; Lorusso, D.; Montanelli, S.; and Varese, G. 2008. Towards a benchmark for instance matching. In *Int. Workshop on Ontology Matching, volume 431, 2008*.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11.

Jaiswal, A.; Miller, D. J.; and Mitra, P. 2010. Uninterpreted schema matching with embedded value mapping under opaque column names and data values. *Knowledge and Data Engineering, IEEE Transactions on 22.2* 291–304.

Joachims, T. 2002. *Learning to Classify Text using Support Vector Machines Methods, Theory, and Algorithms*. Springer.

Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8):707–710.

Mayfield, J.; Alexander, D.; Dorr, B.; Eisner, J.; Elsayed, T.; Finin, T.; Fink, C.; Freedman, M.; Garera, N.; Mayfield, J.; McNamee, P.; Mohammad, S.; Oard, D.; Piatko, C.; Sayeed, A.; Syed, Z.; and Weischedel, R. 2009. Cross-document coreference resolution: A key technology for learning by reading. In *Proc. AAAI Spring Symp. on Learning by Reading and Learning to Read*.

Miller, G. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Mitra, P.; Noy, N. F.; and Jaiswal, A. R. 2005. Omen: A probabilistic ontology mapping tool. In *Int. Semantic Web Conf.*, 537–547. Springer Berlin Heidelberg.

Nadeau, D., and Sekine, S. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30.1:3–26.

Nikolov, A.; Uren, V.; and Motta, E. 2010. Data linking: Capturing and utilising implicit schema level relations. In *Int. Workshop on Linked Data on the Web*.

Nottleman, H., and Straccia, U. 2007. Information retrieval and machine learning for probabilistic schema matching. *Information processing and management 43.3* 552–576.

Rao, D.; McNamee, P.; and Dredze, M. 2010. Streaming cross document entity coreference resolution. In *Int. Conf. on Computational Linguistics (COLING)*, 1050–1058.

Ratinov, L., and Roth, D. 2009. Design challenges and misconceptions in named entity recognition. In *The Thirteenth Conf. on Computational Natural Language Learning*. Association for Computational Linguistics.

Reeve, L., and Han, H. 2005. Survey of semantic annotation platforms. In *The 2005 ACM symposium on Applied computing*, 1634–1638. ACM.

Seddiqui, M., and Aono, M. 2010. Ontology instance matching by considering semantic link cloud. In *9th WSEAS Int. Conf. on Applications of Computer Engineering*.

Singh, S.; Subramanya, A.; Pereira, F.; and McCallum, A. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. *Association for Computational Linguistics*.

Sleeman, J., and Finin, T. 2010. Computing foaf coreference relations with rules and machine learning. In *The Third Int. Workshop on Social Data on the Web*. ISWC.

Suchanek, F. M.; Abiteboul, S.; and Senellart, P. 2012. Paris: Probabilistic alignment of relations, instances, and relations. In *38th Int. Conf. on Very Large Databases*. VLDB.

Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; and Su, Z. 2008. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, 990–998.

Tang, J.; Zhang, D.; and Yao, L. 2007. Social network extraction of academic researchers. In *ICDM'07*, 292–301.

Uryupina, O.; Poesio, M.; Giuliano, C.; and Tymoshenko, K. 2011. Disambiguation and filtering methods in using web knowledge for coreference resolution. In *the 24th Int. Florida Artificial Intelligence Research Society Conf.*