Probabilistic Models for Collective Entity Resolution Between Knowledge Graphs

Jay Pujara*
Department of Computer Science
University of Maryland
College Park, MD 20742
jay@cs.umd.edu

Kevin Murphy
Google Inc.
1600 Amphitheatre Pkwy
Mountain View, CA 94043
kpmurphy@google.com

Xin Luna Dong
Google Inc.
1600 Amphitheatre Pkwy
Mountain View, CA 94043
lunadong@google.com

Curtis Janssen
Google Inc.
1600 Amphitheatre Pkwy
Mountain View, CA 94043
clj@google.com

Abstract

The growing popularity of structured knowledge bases such as knowledge graphs necessitates integrating multiple knowledge sources. A key component of this integration is entity resolution (ER), reconciling instances of a single entity occurring in different knowledge graphs. In contrast to the conventional ER problem setting, we consider the scenario where ER judgments for related entities are made collectively while also determining when a new entity should be added to the graph. Our approach uses hinge-loss Markov random fields to define a joint probability distribution over entity coreferences. We apply this model to two publicly-available knowledge graphs, MusicBrainz and Freebase where relational structure allows us to collectively resolve musical artists and albums, achieving an F1 of 0.84.

1 Motivation

Knowledge base construction is a problem of growing importance, and a number of projects are working to use human collaboration[1, 5] or automatic methods[6] to produce structured knowledge bases of entities and their attributes and relationships referred to as knowledge graphs (KGs). The rise of KGs motivates the problem of integrating information between KGs, which often includes subproblems such as entity resolution, schema mapping, and data fusion.

We are interested in the problem of integrating facts and entities from a source KG, such as MusicBrainz, into a target KG, such as Freebase. We assume that the predicates have been aligned; however, for each source entity e_1, we need to determine if it matches an existing target entity e_2, or if it is a new entity e_3. In our case, the entities correspond to musicians, albums, and tracks on albums, but the technique is quite general.

2 Approach and Preliminary Results

Our approach is to generate a set of candidate matches $\mathbf{C}(e_1)$ based on matching names, and then to pick $e_2 \in \mathbf{C}(e_1)$ such that $\text{sim}(e_1, e_2)$ is maximized. The problem is that the similarity of two entities depends on which other entities they are connected to, so the entity resolution decisions have to be made jointly. Collective entity resolution[4] techniques have shown much promise, and in our

*work performed while at Google Inc.

`baseline` :

$[1]\ \textsc{Match}(A_1, A_2) \rightarrow \textsc{Same}(A_1, A_2)$

`Collective` :

$[2]\ \textsc{Album}(A_1) \wedge \neg\textsc{Album}(A_2) \wedge \textsc{Match}(A_1, A_2) \rightarrow \neg\textsc{Same}(A_1, A_2)$

$[1]\ \textsc{Same}(A_1, A_2) \rightarrow \neg\textsc{Same}(A_1, A_3)$

$[10]\ \textsc{AlbumArtist}(A_1, M_1) \wedge \textsc{AlbumArtist}(A_2, M_2) \wedge \textsc{Same}(A_1, A_2) \rightarrow \textsc{Same}(M_1, M_2)$

`NewEntity` :

$[1]\ \textsc{Match}(A_1, A_2) \wedge \textsc{New}(A_2) \rightarrow \textsc{Same}(A_1, A_2)$

$[1]\ \textsc{Same}(A_1, A_2) \wedge \textsc{New}(A_3) \rightarrow \neg\textsc{Same}(A_1, A_3)$

Figure 1: Rules used for ER models

Method	AUPRC	F1
`baseline`	0.416	0.734
`Collective`	0.569	0.805
`NewEntity`	0.724	0.840

Table 1: Results for 11K artists and albums added from MusicBrainz to Freebase

work we introduce a straightforward method that combines structural relationships in knowledge graphs with probabilistic modeling.

We propose to use hinge-loss MRFs[3] to define a probability distribution over entity co-references. Hinge-loss MRFs elegantly incorporate continuous-valued attributes (such as similarities) and offer a convex objective for MAP state optimization. We use the PSL[2] framework for templating hinge-loss MRFs, which allows us to define models using first-order logic syntax, as shown in Figure 1. Each rule has an accompanying weight (in square brackets) and encodes a dependency between variables and constants. For example, the baseline relates the co-reference prediction (Same) to a similarity score (Match), while the second rule restricts this dependency to entities that share the type Album and has a higher weight.

Our initial experiments resolve entities between the knowledge graphs MusicBrainz[1] and Freebase[5]. We use a corpus of entities added to Freebase from MusicBrainz between 5/5/14 and 6/29/14 consisting of 332K candidate Freebase entities, 11K MusicBrainz entities, and 15.7M relations involving these entities. Trained human annotators provide ground truth for the 11K MusicBrainz entities. Freebase candidates were generated based on a normalized string match of names with match scores based on an entity frequency, as well as new entity placeholders given a fixed match score of 0.5.

We compare three models (summarized in Figure 1) that each define a set of weighted rules for entity resolution. The first `baseline` uses only the string-based match scores. `Collective` uses ontological data from the knowledge graph such as type information and album-artist relationships as well as enforcing functional mappings. `NewEntity` adds additional rules that favor adding new entities to the knowledge graph when no existing candidate strongly matches. Table 2 shows that using collective entity resolution using rules from the knowledge graph ontology improve performance over a non-collective baseline, and appropriately handling new entities provides a further boost.

3 Ongoing Work

We are currently improving our experiments by adding more sophisticated ontological rules and using weight learning to improve the performance of the model. Finally, we hope to extend our entity resolution to other relational knowledge, particularly uncertain extractions generated from webpages.

References

[1] MusicBrainz: The Open Music Encyclopedia. http://musicbrainz.org.

[2] Probabilistic Soft Logic. https://github.com/linqs/psl.

[3] S. H. Bach, B. Huang, B. London, and L. Getoor. Hinge-loss Markov Random Fields: Convex Inference for Structured Prediction. In *UAI*, 2013.

[4] I. Bhattacharya and L. Getoor. Collective Entity Resolution In Relational Data. *TKDD*, 2007.

[5] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *SIGMOD*, 2008.

[6] G. Weikum and M. Theobald. From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. In *PODS*, 2010.