DRETa: Extracting RDF from Wikitables

Emir Muñoz, Aidan Hogan, and Alessandra Mileo

Digital Enterprise Research Institute, National University of Ireland, Galway
{emir.munoz, aidan.hogan, alessandra.mileo}@deri.org

Abstract. Tables are widely used in Wikipedia articles to display relational information – they are inherently concise and information rich. However, aside from info-boxe s, there are no automatic methods to exploit the integrated content of these tables. We thus present DRETa: a tool that uses DBpedia as a reference knowledge-base to extract RDF triples from generic Wikipedia tables.

1 Introduction

Large amounts of data on the Web are presented in tables [3]. Interpreting and extracting knowledge from HTML Web tables is thus relevant for many areas, including: finance, public policy, user experience, health-care, and so forth. However, such tables are diverse in terms of representation, structure and vocabulary used; they often contain polysemous (or missing or otherwise vague) attribute labels, ambiguous free-text cell content and referents, cell spanning multiple rows and/or columns, split tables, obscured contextual validity, and so forth. Recovering the semantics of generic Web tables is thus extremely challenging.

Instead of interpreting generic Web tables, we have rather been focussing on (partially) interpreting the tables embedded in Wikipedia (henceforth "Wikitables"). In particular, we have created DRETa: a prototype for (semi-)automatically performing a best-effort extraction of RDF triples from Wikitables. Though many of the challenges remain the same, focussing on Wikitables has a number of distinct advantages over the more general Web table scenario: (1) Wikitable cells often contain links to Wikipedia articles that disambiguate the entities being talked about; (2) Wikitables contain a high ratio of rich encyclopaedic knowledge; (3) the context of a Wikitable can be mapped to the article in which it appears; (4) existing RDF knowledge-bases, that offer partial exports of Wikipedia content, can be used for reference and for mining legacy entity URIs and predicates. To maximise the precision of the triples extracted from Wikitables, DRETa exploits these unique advantages insofar as possible.

The DRETa system works by using a suitable reference knowledge-base—such as DBpedia [1], YAGO2 [4], Freebase [2], etc.—to extract triples from Wikitables. We only consider tables embedded in article bodies (class=wikitables in the HTML source), filtering info-boxes (already used by DBpedia and YAGO2) and tables-of-content. Our prototype system—available at http://deri-srvgal36.nuigalway.ie:8080/wikitables-demo-0.1.0/—currently uses DBpedia as the reference knowledge-base. Thus, the RDF triples

No.	Position	Player	No.	Position	Player
1	GK	David de Gea	19	FW	Danny Welbeck
2	DF	Rafael	20	FW	Robin van Persie
3	DF	Patrice Evra *(vice-captain)*	21	FW	Ángelo Henríquez
4	DF	Phil Jones	23	MF	Tom Cleverley
5	DF	Rio Ferdinand	24	MF	Darren Fletcher
6	DF	Jonny Evans	25	MF	Nick Powell

Fig. 1: Split table of current Manchester United F.C. squad members (abridged from http://en.wikipedia.org/wiki/Manchester_United_F.C.).

extracted by DRETa from Wikipedia's tables use the same URIs as DBpedia to identify entities (subject/object URIs) and relations (predicate URIs). Our system can be seen as enriching the reference knowledge-base with additional facts found in tables using the legacy relations from the knowledge-base itself.

2 Triple Extraction: A Motivating Example

We sketch the extraction process by way of a real-world example. Figure 1 presents a Wikitable abridged from the "Manchester United F.C." Wikipedia article, containing relations between players, their shirt number, country and position. There are also relations between players and the entity described by the article (their current club is Manchester United F.C.).

Aside from the **No.** columns, the cells of the table contain hyperlinks to other articles in Wikipedia, including countries, football positions, and individual players. For example, the flags link to articles for the country; GK links to the article for Goalkeeper_(associated_football). These links provide unambiguous referents to Wikipedia entities, which can in turn be mapped directly to DBpedia entities and descriptions. Currently, DRETa focuses on the extraction of relations between cells containing wiki-links and does not consider plain-string values. For example, it would not try to extract player numbers from the table.

Previous works on extracting RDF from such tables (e.g., [5]) propose a vertical, column-centric approach, where columns are seen as referring to types and/or relations that can then be extended to all rows. We rather adopt a horizontal, row-centric approach and look at the pre-existing relations between entities on the same row. For example, if we find that the predicate `dbp:position` holds between `dbr:David_de_Gea` and `dbr:Goalkeeper_(association_football)` (GK), we can suggest that the relation holds from all entities in the **Player** column to all entities on the same row in the **Position** column. Similarly, we consider the article in which the table is found to be a protagonist for the table. If we find the predicate `dbo:team` and `dbp:currentclub` holds from `dbr:David_de_Gea` to `dbr:Manchester_United_F.C.`, we can propose that the same relations hold from all entities in the **textsfPlayer** column to the protagonist (the article entity).

Candidate triples that we extract are further associated with a number of features to help classify them as correct/incorrect, associating each triple with a confidence score. Details of the features are out-of-scope, but, for example, we hypothesise that the more rows a given relation holds for across entities in two fixed columns in the reference KB, the higher the likelihood that that relationship exists on all such rows. Other features, such as a match between the label of the candidate relation and a column header, can further strengthen confidence in the match. Using a selection of 750 random triples labelled by three judges, we employed offline various machine learning methods (SVM, Naïve Bayes, Bagging Decision Trees, Random Forest, Logistic) to train a range of binary classifiers that are then made available to the DRETa system for classifying (in)correct triples, and ultimately for ranking and filtering candidate triples at runtime.

3 DRETa system description

The current DRETa prototype works with DBpedia 3.8 and Wikipedia article names (for auto-completion) as last updated in May 2013. The user submits a Wikipedia article title (with the help of auto-complete) and optionally selects a classifier. The extraction process is then as follows: *i)* the selected article is downloaded and cached in memory for future queries; *ii)* all the tables are extracted, repaired (as applicable), and filtered; *iii)* for each table, mappings from wiki-links to KB entities are executed; *iv)* candidate relations for pairs of resources are collected from the reference KB and candidate triples proposed; *v)* the selected classifier is run to rank triples by confidence, *vi)* each candidate triple is tested against the KB to determine if it is a pre-existing triple or not.

Following our motivating example, Figure 2 presents DRETa's results for the Wikipedia article "Manchester United F.C.". After ca. 11 secs. we get 457 RDF triples extracted from tables contained in that article as a result. Some of these triples already exist in DBpedia (rows with an DBpedia icon visible). Others are novel: from the top-10 extracted triples sorted by confidence, for example, we extract triples for the birth-places of the footballers Rafael da Silva (Brazil) and Phil Jones (England), which were not previously known to DBpedia.

4 Conclusion

In this paper, we have presented DRETa: a prototype system for extracting RDF triples from Wikipedia tables. The process maps entities in table cells to entities in a reference knowledge-base and then looks for potential relations that hold between entities on the same row across two given columns, or that hold between entities in a single column and the article entity. Triples are associated with a set of features that help classify correct/incorrect triples. A selection of machine learning methods are used offline to train classifiers, where these classifiers can then be used to rank the confidence of triples. The prototype is available online at `http://deri-srvgal36.nuigalway.ie:8080/wikitables-demo-0.1.0/`.

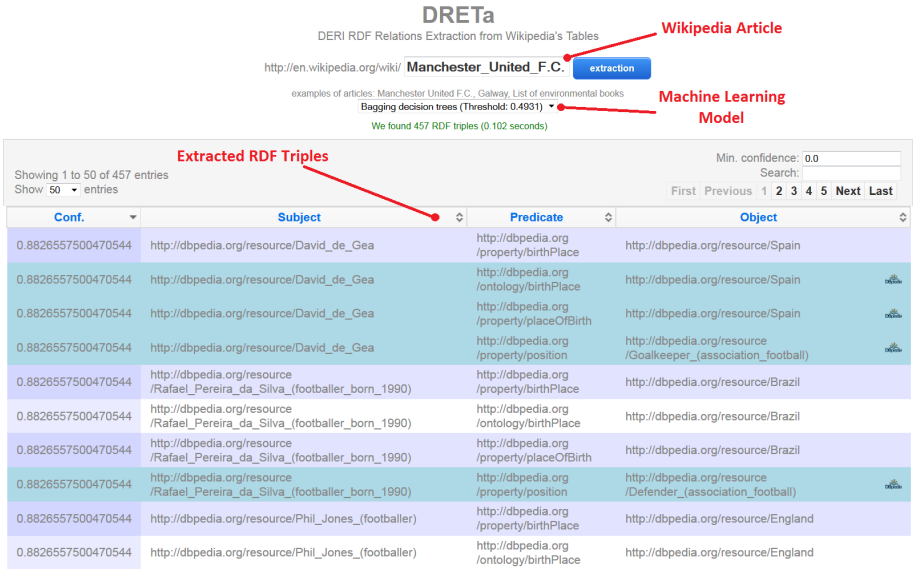

Fig. 2: DRETa demo interface showing the top triples extracted from tables in the "Manchester United F.C." article with associated confidence scores.

We are currently investigating methods to perform a high-quality "bulk" triplification of all tables in English Wikipedia. We have used the architecture of DRETa to extract 22 million triples from over one million tables (all tables) in English Wikipedia. We have already estimated a 52% precision measure for the raw candidate triples extracted by our methods and our next steps are to evaluate the extent to which machine learning techniques and different classifiers improve this baseline precision by filtering incorrect triples at various thresholds, and we will identify a gold standard to be able to estimate recall. We also wish generalize our methods with YAGO2 and Freebase as reference knowledge-bases, towards a more ambitious generalisation that employs entity-recognition tools (instead of wiki-links) for processing generic Web tables.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia – a crystallization point for the Web of Data. JWS 7(3), 154–165 (2009)
2. Bollacker, K.D., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Wang, J.T.L. (ed.) SIGMOD Conference. pp. 1247–1250. ACM (2008)
3. Cafarella, M.J., Halevy, A.Y., Wang, D.Z., Wu, E., Zhang, Y.: Webtables: exploring the power of tables on the web. PVLDB 1(1), 538–549 (2008)
4. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. 194, 28–61 (2013)
5. Mulwad, V., Finin, T., Syed, Z., Joshi, A.: Using linked data to interpret tables. In: COLD Workshop (November 2010)