Ontological Quality Control in Large-scale, Applied Ontology Matching

Catherine Legg and Samuel Sarjant ([clegg,sarjant]@waikato.ac.nz)

The University of Waikato, Hamilton, New Zealand

Introduction

Many large-scale ontology mapping approaches rely on label matching and other relatively simple syntactic features. We offer a suite of partially overlapping ontology mapping heuristics which allows us to hypothesise matches and test them against the knowledge in our source ontology [1, 3].

Methodology

- Mappings are created via a **stage-wise process**.
- Each stage outputs one or more **weighted results**, where weight is roughly proportional to **mapping confidence**.
- The order of the process is governed by a **priority queue**.
- Beginning with an ontological concept, we employ three stages:
- **Ontology-Wikipedia** mapping heuristics,
- **Wikipedia-Ontology** mapping heuristics,
- **Consistency Checking** heuristics.

Ontology-Wikipedia Mapping Heuristics

Title Matching: Return all articles with the **same name** as the concept (equally weighted).

Synonym Matching: Return all articles with **anchor text** (internal link text) equal to one of the concept's synonyms. Weights are proportional to the frequency of links to the article.

Context-Related Synonym Matching: Like **Synonym Matching**, but uses a **set of context articles** composed from the concept's ontological context (mapping context concepts to articles). Each output article weight is multiplied by **relatedness** — similarity of incoming and outgoing links [2].

Wikipedia-Ontology Mapping Heuristics

Title Matching: As above, but from article to concept.

Label Matching: Returns all concepts with the same name/synonym as the article's **incoming anchor text**. Weight is proportional to the frequency of the anchor text.

Basketball Example
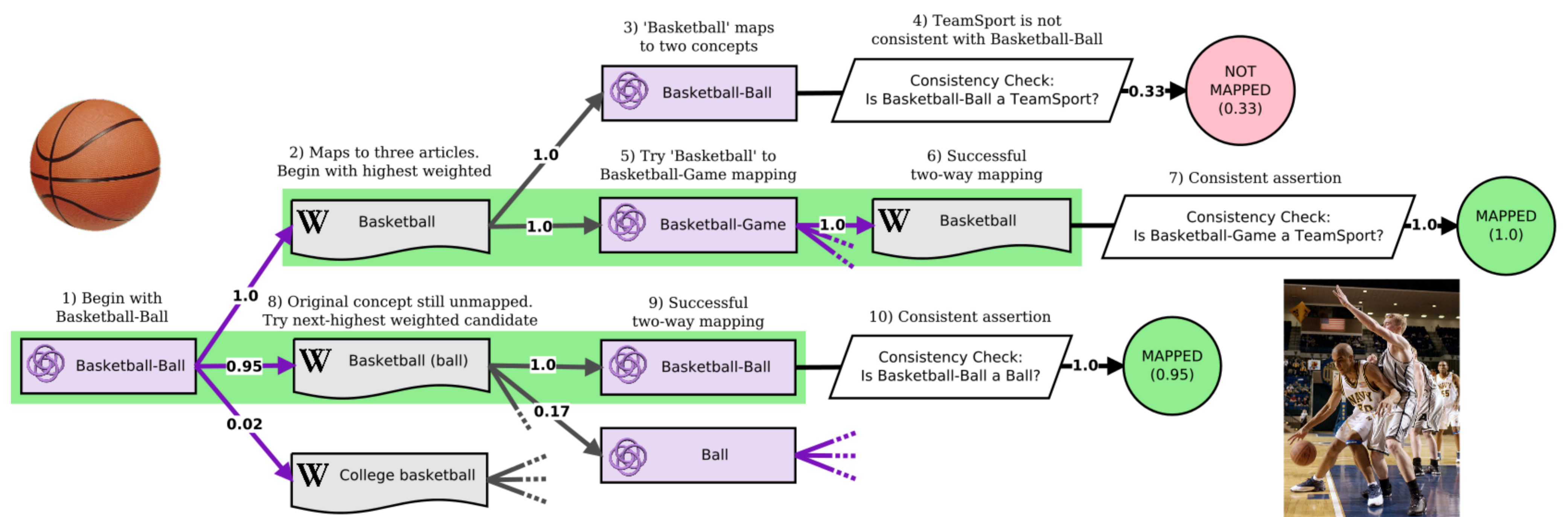
Consistency Checking Heuristics

- Consistency checking uses **assertions** extracted from the article as part of the mapping weight.
- E.g. *"X is/was/are/were a/an/the Y"* where **Y** can represent multiple weighted concepts (using the same mapping process to map article to concept).
- Using OpenCyc's **ontological disjointness information**, the mapping's weight is multiplied by the proportion of consistent assertions.

For example:

- *"**Bill Laswell** is an [[American|United States]] [[bassist]], [[record producer|producer]] and [[record label]] owner."*
- 75% of assertions are consistent:
- *BillLaswell* is a *UnitedStatesPerson*, *BassGuitarist*, *Producer*.
- '*BillLaswell* is a *RecordCompany*' is rejected because a *LivingThing* cannot be a *NonLivingThing*.

Conclusions

- The algorithm identified ~55,000 mappings between the ontology (of ~180,000 concepts) and Wikipedia.
- 93% accuracy from a manual evaluation of 300 mappings.

Future Work

- The modular mapping process can easily integrate further mapping heuristics.
- Consistency checking is more effective when more information is extracted as consistent groups are more easily distinguished.
- We have developed a social ontology interface to incorporate user contributions and feedback (bit.ly/GRRBcP)

References

[1] Medelyan, O. and Legg, C., **Integrating Cyc and Wikipedia: Folksonomy Meets Rigorously Defined Common-Sense**. Proceedings, Workshop on Wikipedia and AI, Chicago AAAI08, Chicago, United States of America, July 2008.

[2] Milne, D. and Witten, I. **An open-source toolkit for mining Wikipedia**. Artificial Intelligence(194), January 2013.

[3] Sarjant, S., Legg, C., Robinson, M., Medelyan, O., **All You Can Eat Ontology-Building: Feeding Wikipedia to Cyc**. IEEE/WIC/ACM International Conference on Web Intelligence (WI-09), Milan, Italy, September 2009.