Towards a Linking between WordNet and Wikidata

John P. McCrae
Data Science Institute
NUI Galway
Ireland
john@mccr.ae

David Cillessen
Data Science Institute
NUI Galway
Ireland
D.CILLESSEN1@nuigalway.ie

Abstract

WordNet is the most widely used lexical resource for English, while Wikidata is one of the largest knowledge graphs of entity and concepts available. While, there is a clear difference in the focus of these two resources, there is also a significant overlap and as such a complete linking of these resources would have many uses. We propose the development of such a linking, first by means of the hapax legomenon links and secondly by the use of natural language processing techniques. We show that these can be done with high accuracy but that human validation is still necessary. This has resulted in over 9,000 links being added between these two resources.

1 Introduction

English WordNet (McCrae et al., 2019, 2020), derived from Princeton WordNet (Miller, 1995; Fellbaum, 2012, PWN)[1], is the most complete wordnet for English, while Wikidata[2] provides one of the largest collection of encyclopedic facts in machine readable form. Moreover, as Wikidata is an open resource to which anyone can contribute and data is published without any license, it is quickly becoming a central database to which knowledge graphs can link. As such, a linking between WordNet and Wikipedia would provide value to users of both resources, and potentially make it easier to extend WordNet in the future with new synsets. However, there are significant differences between the scope of the two projects, with WordNet specialising on providing information about the use of words in English, including verbs, adjectives and adverbs, whereas Wikidata describes entities, mostly by means of proper nouns, although lexical information is currently being added to Wikidata (Nielsen, 2020). Still, there is a significant overlap in terms of the proper and common nouns in WordNet and providing links to Wikidata would help to improve and extend the usage of WordNet.

A linking between the proper nouns in WordNet and Wikipedia was constructed by McCrae et al. (2018) and as a side part of this work we updated and contributed this list to Wikidata including manually remapping 156 links that had become stale. However, we also see that for most common nouns it is still possible to match most of the senses to a concept in Wikidata, for example of the eight senses of 'work' in WordNet, six of them can easily be mapped to a concept in Wikidata and only two abstract definitions 'activity directed toward making or doing something' and 'applying the mind to learning and understanding a subject (especially by reading)' are not obviously available in Wikidata. In fact, out of 122,147 noun lemmas in English WordNet 67,569 (55.3%) are represented by an entry in Wikidata and as such we believe that the majority of noun senses in WordNet should have a counterpart in Wikidata.

Given the size of the task of this linking, it is obvious that we should have some automatic help to improve the linking process; however, neither resource would accept fully automatic linking as has been applied in other resources such as BabelNet (Navigli and Ponzetto, 2010) and UBY (Gurevych et al., 2012). As such, in this paper we start the process of using automatic tools to construct the links between the datasets and manually validating. For the purpose of this paper, our first focus is on what we refer to as *hapax* links, that is links for which there is only a single sense for the lemma in WordNet and for which only one page in Wikidata has this lemma as the English title. We then consider how we could extend this further to the links where there is ambiguity in the lemma. Finally, we consider how this linking could

[1] We use 'WordNet' to refer to either resource

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

be used to contribute back to English WordNet and extend the existing categories there.

2 Related Work

Most of the focus to date has been on the development of automatic linking between WordNet and encyclopedic knowledge graphs based on Wikipedia such as DBpedia (Auer et al., 2007) or Wikidata. One of the most prominent examples of this is BabelNet (Navigli and Ponzetto, 2010), which mapped WordNet to Wikipedia using a word sense disambiguation algorithm, in which the surrounding elements in the synset graph and article text were used as context for disambiguation. In their work (Navigli and Ponzetto, 2012), the authors report an F-Measure of 82.7% in their linking, and while this is strong it cannot be considered to be a gold standard. Another approach has been through the use of Personalised PageRank (PPR) (Agirre and Soroa, 2009), which was first attempted by Toral et al. (Toral et al., 2009) and later improved by Meyer and Gurevych (Meyer and Gurevych, 2011) to create the UBY resource (Gurevych et al., 2012). A similar resource, YAGO (Suchanek et al., 2008), has also been constructed by means of automatic linking and while they report very high accuracy (97.7%) this referred to only a limited number of concepts that are linked. There have also been attempts to link WordNet to other resources including the SemLink (Bonial et al., 2013; Palmer et al., 2014) that have provided links to other lexical resources and ontologies. In contrast to these works, this work is developing a manual linking that aims to be usable as a gold standard.

3 Hapax Linking

3.1 Methodology

One of the most obvious ways to get a good linking is to focus on the elements in the two resources that are *hapax legomenon* in the resource, that is that they only occur a single time in the resource. By this, we mean that for English WordNet, a synset only occurs in a single noun synset and for Wikidata the label is unique for this concept. As such, we first base our approach on identifying and linking these elements between the two resources based on an exactly matching hapax lemma in both resources. An initial analysis of this showed that there were quite a large number of links; however, we noticed that due to the large number of entities that are available in Wikidata there were often spurious links. In order to mitigate this, we took a couple of quick heuristics before evaluation

- Each Wikidata entity is identified with a 'Q' code that is assigned sequentially. A quick analysis suggested that 'Q' numbers over 10,000,000 generally referred to entities of such little significance that it was extremely unlikely they would be mentioned in English WordNet.
- We filtered out all entities whose definition contained "Wikipedia disambiguation page" or "Wikimedia disambiguation page" as these were not real-world entities in Wikidata.
- We also filtered out all entities whose definitions were of the form of 1-3 words followed by the word "by" and then 1-4 words. A very large number of entities matching this pattern were irrelevant entities such as "song/album/film" by "band/author/director".

In total, using this method we discovered 16,452 candidates for this hapax linking, which represents 19.5% of all noun synsets in WordNet.

3.2 Evaluation

In order to evaluate the quality of the hapax linking, and automatically check for any errors in the linking, we set up an evaluation program using a simple spreadsheet to evaluate the hapax links. We provided the evaluators with enough information to evaluate the quality of the linking, in particular: the lemma and Wikidata identifier, the definitions of the concept given in both resources and the (instance) hypernyms of the concepts in each resource. The results of this can be seen in Table 1, where we give four examples of the linkings extracted, where the first three were the first three rows randomly presented to our evaluators. The fourth row, 'Occam' gives an interesting example of a spurious match, where the philosopher is linked to a programming language named after the philosopher. As part of the annotation guidelines, annotators were instructed to consider matches as long as they were broadly correct, so for example 'prunus triloba' refers to a species of plants in Wikidata but as a tree in English WordNet, but as they clearly refer to the same plant they are considered matching even though ontologically a species is not a tree.

Wikidata ID	Lemma	Wikidata Definition	WordNet Definition	Wikidata Hypernyms	WordNet Hypernyms
Q2663273	boasting	to speak with excessive pride and self-satisfaction about one's achievements, possessions, or abilities [...]	speaking of yourself in superlatives	*none*	speech act
Q514686	aphonia	medical condition leading to loss of voice	a disorder of the vocal organs that results in the loss of voice	voice disorder	defect of speech, speech disorder, speech defect
Q105719	Jean Harlow	American film actress	United States film actress who made several films with Clark Gable (1911-1937)	human	actress
Q838062	Occam	Concurrent programming language	English scholastic philosopher and assumed author of Occam's Razor (1285-1349)	programming language; procedural programming language	philosopher
Q2727171	prunus triloba	species of plant	deciduous Chinese shrub or small tree with often trilobed leaves grown for its pink-white flowers	Prunus	almond tree

Table 1: Examples of the Hapax linking and the information give to annotators to evaluate the results.

So far the annotation has been completed up to 1,997 entities and of those 1,920 have been accepted (96.1%) indicating that the hapax linking is overall very reliable. The annotators quickly noted that some Wikidata classes contained many entities not found in English WordNet, in particular 'album', 'band', 'single', 'video game', 'film', 'television series', 'family name', 'written work', 'song' and 'television program'. These elements account for 35 of the false links and if they were excluded the overall accuracy of the hapax linking would be 98.4%. In addition, we also evaluated the inter-annotator agreement of the linking using two annotators over 497 evaluations and a Cohen's kappa score of 81.4% was obtained indicating strong agreement between the annotators. In fact, 8 of the 11 disagreements between the annotators were errors by the annotators and only 3 were due to the nature of the task. This suggests that the annotators are able to make clear judgements in the vast majority of cases.

3.3 Publishing

The links have been made available through Wikidata by means of the property `P5063`, which links the elements to the GWA InterLingual Index (ILI) (Bond et al., 2016). These were contributed to the Wikidata project by means of QuickStatements. In addition, the data is made available as a comma-separated value list on the English WordNet project.

4 Towards a complete linking

The hapax linking above, while it has a very high accuracy is also not sufficient in order to create a complete linking between two resources, as such we have attempted to evaluate how easily this can be extended to a complete linking of the two re-

Q7366	song
Q7889	video game
Q11424	film
Q101352	family name
Q134556	single
Q207628	musical composition
Q215380	musical group
Q222910	compilation album
Q386724	work
Q482994	album
Q3305213	painting
Q5398426	television series
Q5741069	rock band

Table 2: List of classes in Wikidata that do not frequently occur in English WordNet

sources using the Naisc system (McCrae and Buitelaar, 2018), so that we can also link entities where there is some ambiguity in the potential matching labels.

4.1 Extending the linking with Naisc

The first step in creating the linking is to extract the relevant facts about the entities from WordNet and Wikidata. From English WordNet, we extracted the definitions and labels as well as the synset links, and similarly for Wikidata we extracted the English labels and definitions, as well as the links between synsets. As the size of Wikidata was very large, we limited this extraction to entities whose terms occurred in English WordNet and hypernyms of these terms. As previously, we filtered these entities using heuristics, namely the "X by Y" pattern, disambiguation pages and discarding Q IDs over 10,000,000 as before. In addition, we also developed a reject list and removed all elements that were hyponyms of this list, which is shown in Table 2. We then applied the Naisc methodology consisting of the following analysis

- The system identified the hapax links as in the previous step and accepted them automatically due to the high precision of these links established in the previous step. This created a merged graph containing the links between the Wikidata concepts, the links between the English WordNet synsets and the hapax links.
- The definitions were compared using the Jaccard similarity of the two definitions both at word-level and character-level, as in previous word similarity approaches (McCrae and Buitelaar, 2018).
- In addition, we analysed the similarity of each element according the Personalised PageRank (PPR) algorithm (Page et al., 1999), using the Fast-PPR implementation (Lofgren et al., 2014), as in Meyer and Gurevych (Meyer and Gurevych, 2011).
- This generated three scores, which were normalized in the range [0,1], by means of percentile ranking, so that the score which corresponds to the lowest of the top 10% of scores was mapped to 0.1.
- A simple average of the three scores (character-level Jaccard, word-level Jaccard and PPR) was used to rank each potential match.
- We used a bijective assumption, that each entity in WordNet matches only a single element in Wikidata, and as such the problem can be cast as an *assignment problem* (Munkres, 1957), which can be solved with the Hungarian algorithm (Kuhn, 1955). However, due to the very large size of the datasets, we instead used a simple greedy approach.

4.2 Evaluation of the Extended linking

The evaluation of the linking was completed by two annotators who evaluated 100 links predicted by the system. They agreed on an accuracy between 65-66% with a Cohen's Kappa of 0.934 of the automatic linking. The primary disagreements were on two examples "snack bar" defined as "inexpensive food counter" or a "usually inexpensive bar" and "brother" defined as "Hong Kong internet slang" or "used as a term of address for those male persons engaged in the same movement". Divided by the prediction scores, those links predicted with a confidence of less than 60% by the system were all incorrect (0.0% accuracy), those with a 60-80% accuracy were correct 23/39 times (59.0% accuracy) and those with a greater than 80% confidence were correct 42/49 times (85.7% accuracy). These statistics indicate that the system's confidence was a good predictor of the accuracy of links.[3]

[3]These scores were not shown to the annotators in the manual evaluation

5 Discussion

One of the key objectives of this project is to enable the extension of WordNet with more entities achieving a similar goal to that of Bond and Bond (2019) of developing wordnets of geographic place names, but for more categories than just place names. Given that we have 9,149 links now confirmed between WordNet and Wikidata, we can make inference about likely extra entities that could be added to WordNet. For example, if we know that 'Paris' (`i83645`) is an instance of 'national capital' (`i82619`) and we have now linked this to Wikidata (`Q90`) which asserts that this is an instance of 'capital' (`Q5119`), then we could establish the link between the categories for 'national capitals' and 'capitals' and add capitals that are missing from WordNet, such as 'Juba' (`Q1947`). We are currently investigating the potential to create an extended WordNet from this linking, however there are challenges due to the difference in structure between WordNet and Wikidata. For example, 'George Washington' (`i97352/Q23`) is asserted as an instance of 'general' (`i90718`) and 'President of the United States' (`i92216`) in WordNet but only as a 'human' (`Q5`) in Wikidata. Instead, Wikidata uses different properties, namely 'occupation' (`P106`) and 'position held' (`P39`) to assert the facts expressed in WordNet. It is unclear how best these inconsistencies should be resolved in the context of WordNet.

6 Conclusion

In this work we have analysed the task of linking the noun hierarchy of WordNet with Wikidata. We found that the approach relying on hapax linking can be achieved with very high accuracy, although this does still produce occasional errors. However, for ambiguous senses the task of linking is still much harder and the automatic methods need to be further refined to produce high quality results. As a result of this we have increased the amount of links between Wikidata and WordNet to nearly 10,000 and have made them available in Wikidata and English WordNet[4]. We hope that this can be a seed to further the integration of the two projects and close the gap between the lexical and encyclopedic information in the two resources.

[4] https://github.com/globalwordnet/english-wordnet

Acknowledgements

This work is supported by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexicographic Infrastructure) and by a research grant from Science Foundation Ireland, co-funded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer.

Francis Bond and Arthur Bond. 2019. GeoNames Wordnet (gnwn): extracting wordnets from GeoNames. In *Proceedings of the 10th Gloabl WordNet Conference*, pages 387–393.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9–17.

Christiane Fellbaum. 2012. Wordnet. *The encyclopedia of applied linguistics*.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. UBY-a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590.

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Peter A Lofgren, Siddhartha Banerjee, Ashish Goel, and C Seshadhri. 2014. FAST-PPR: scaling personalized pagerank estimation for large graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1436–1445.

John P. McCrae. 2018. Mapping WordNet Instances to Wikipedia. In *Proceedings of the 9th Global WordNet Conference*.

John P. McCrae and Paul Buitelaar. 2018. Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1):109–123.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – An Open-Source WordNet for English. In *Proceedings of the 10th Global WordNet Conference – GWC 2019*.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In *Proceedings of the Multimodal Wordnets Workshop at LREC 2020*, pages 14–19.

Christian M Meyer and Iryna Gurevych. 2011. What psycholinguists know about chemistry: Aligning Wiktionary and WordNet for increased domain coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 883–892.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Finn Nielsen. 2020. Lexemes in Wikidata: 2020 status. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 82–86, Marseille, France. European Language Resources Association.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Martha Palmer, Claire Bonial, and Diana McCarthy. 2014. Semlink+: Framenet, verbnet and event ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 13–17.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A large ontology from Wikipedia and WordNet. *Journal of Web Semantics*, 6(3):203–217.

Antonio Toral, Oscar Ferrandez, Eneko Agirre, and Rafael Munoz. 2009. A study on Linking Wikipedia categories to Wordnet synsets using text similarity. In *Proceedings of the international conference RANLP-2009*, pages 449–454.