Instance-driven Discovery of Ontological Relation Labels

Marieke van Erp, Antal van den Bosch, Sander Wubben, Steve Hunt
ILK Research Group
Tilburg centre for Creative Computing
Tilburg University
The Netherlands
{M.G.J.vanErp,Antal.vdnBosch,S.Wubben,S.J.Hunt}@uvt.nl

Abstract

An approach is presented to the automatic discovery of labels of relations between pairs of ontological classes. Using a hyperlinked encyclopaedic resource, we gather evidence for likely predicative labels by searching for sentences that describe relations between terms. The terms are instances of the pair of ontological classes under consideration, drawn from a populated knowledge base. Verbs or verb phrases are automatically extracted, yielding a ranked list of candidate relations. Human judges rate the extracted relations. The extracted relations provide a basis for automatic ontology discovery from a non-relational database. The approach is demonstrated on a database from the natural history domain.

1 Introduction

The rapid growth in the digitisation of data has caused many curators, researchers, and data managers of cultural heritage institutions (libraries, archives, museums) to turn to knowledge management systems. Using these systems typically causes them to think about the ontological structure of their domain, involving the identification of key classes in object data and metadata features, and importantly, their relations. The starting point of this process is often a more classical "flat" database matrix model of size $n \times m$, where n is the number of collection items, and m is a fixed number of database columns, typically denoting object metadata features, as cultural heritage institutions are generally well accustomed to using databases of that type. An ontology can be bootstrapped from such a database by first assuming that the database columns can be mapped onto the domain's ontological classes. The next step is then to determine which classes are related to each other, and by which relation. In this paper we present a method that partially automates this process.

To gather evidence for a relation to exist between two ontological classes, it is not possible to simply look up the classes in text. Rather, classes are realised typically as a multitude of terms or phrases. For example, the natural history class "species" is realised as many different instances of species names in text. The automatic discovery of relations between ontological classes thus requires at least a two-step approach: first, the identification of instances of ontological classes in text and their particular relations, and second, the aggregation of these analyses in order to find evidence for a most likely relation.

It is common in ontology construction to use predicative labels for relations. Although no regulations for label names exist, often a verb or verb phrase head is taken, optionally combined with a prepositional head of the subsequent verb-attached phrase (e. g., "occurs in", or "donated by"). In this study, we make the assumption that good candidate labels are frequent verbs or verb phrases found between instances from a particular pair of classes, and that this may sometimes involve a verb-attached prepositional phrase containing one of the two terms. In this paper we explore this route, and present a case study on the discovery of predicative labels on relations in an ontology for animal specimen collections. The first step, identifying instances of ontological classes, is performed by selecting pairs of instances from a flat $n \times m$ specimen database, in which the instances

Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education –LaTeCH – SHELT&R 2009, pages 60–68,
Athens, Greece, 30 March 2009.

are organised by the database columns, and there is a one-to-one relationship between the database columns and the classes in our ontology.

Any approach that bases itself on text to discover relations, is dependent on the quality of that text. In this study we opt for Wikipedia as a resource from which to extract relations between terms. Although the status of Wikipedia as a dependable resource is debated, in part because of its dynamic nature, there is some evidence that Wikipedia can be as reliable a source as one that is maintained solely by experts (Giles, 2005). Wikipedia is also an attractive resource due to its size (currently nearly 12 million articles in over 250 languages). Additionally, Wikipedia's strongly hyperlinked structure closely resembles a semantic net, with its untyped (but directed) relations between the concepts represented by the article topics. Since the hyperlinks in Wikipedia indicate a relations between two encyclopaedia articles, we aim at discovering the type of relation such a link denotes through the use of syntactic parsing of the text in which the link occurs.

The idea of using Wikipedia for relation extraction is not new (Auer and Lehmann, 2007; Nakayama et al., 2008; Nguyen et al., 2007; Suchanek et al., 2006; Syed et al., 2008). However, most studies so far focus on the structured information already explicit in Wikipedia, such as its infoboxes and categories. The main contributions of our work are that we focus on the information need emerging from a specific domain, and that we test a method of pre-selection of sentences to extract relations from. The selection is based on the assumption that the strongest and most reliable lexical relations are those expressed by hyperlinks in Wikipedia pages that relate an article topic to another page (Kamps and Koolen, 2008). The selection procedure retains only sentences in which the topic of the article, identified by matching words in the article title, links to another Wikipedia article. The benefit of the preselection of sentences is that it reduces the workload for the syntactic parser.

Since the system is intentionally kept lightweight, the extraction of relations from Wikipedia is sufficiently fast, and we observe that the results are sufficient to build a basic ontology from the data. This paper is organised as follows. In Section 2 we review related work. In Section 3 the data used in this work is described, followed by the system in Section 4 and an explanation of how we evaluated the possible relations our system discovered is presented in Section 5. We report on the results of our study in Section 6, and formulate our conclusions and points for further research in Section 7.

2 Related Work

A key property of Wikipedia is that it is for the greater part unstructured. On the one hand, editors are encouraged to supply their articles with categories. These categories can be subsumed by broader categories, thus creating a taxonomy-like structure. On the other hand, editors can link to any other page in Wikipedia, no matter if it is part of the same category, or any category at all. An article can be assigned multiple categories, but the number of hyperlinks provided in an average article typically exceeds the number of categories assigned to it.

The free associative hyperlink structure of Wikipedia is intrinsically different from the hierarchical top down architecture as seen in WordNet, as a hyperlink has a direction, but not a type. A Wikipedia article can contain any number of links, pointing to any other Wikipedia article. Wikipedia guidelines state however that wikilinks (hyperlinks referring to another Wikipedia page) should only be added when relevant to the topic of the article. Due to the fact that most users tend to adhere to guidelines for editing Wikipedia pages and the fact that articles are under constant scrutiny of their viewers, most links in Wikipedia are indeed relevant (Blohm and Cimiano, 2007; Kamps and Koolen, 2008).

The structure and breadth of Wikipedia is a potentially powerful resource for information extraction which has not gone unnoticed in the natural language processing (NLP) community. Preprocessing of Wikipedia content in order to extract non-trivial relations has been addressed in a number of studies. (Syed et al., 2008) for instance utilise the category structure in Wikipedia as an upper ontology to predict concepts common to a set of documents. In (Suchanek et al., 2006) an ontology is constructed by combining entities and relations between these extracted from Wikipedia through Wikipedia's category structure and WordNet. This results in a large "is-a" hierarchy, drawing on the basis of WordNet, while further relation enrichments come from Wikipedia's category

structure. (Chernov et al., 2006) also exploit the Wikipedia category structure to which concepts in the articles are linked to extract relations.

(Auer and Lehmann, 2007) take a different approach in that they focus on utilising the structure present in infoboxes. Infoboxes are consistently formatted tables in articles that provide summary information, such as information about area, population and language for countries, and birth dates and places for persons. Although infoboxes provide rich structured information, their templates are not yet standardised, and their use has not permeated throughout the whole of Wikipedia.

Although the category and infobox structures in Wikipedia already provide a larger coverage at the concept or term level than for instance WordNet, they do not express all possibly relevant semantic relations. Especially in specific domains, relations occur that would make the Wikipedia data structure unnecessarily dense if added, thus an approach that exploits more of the linguistic content of Wikipedia is desirable.

Such approaches can be found in (Nakayama et al., 2008) and (Nguyen et al., 2007). In both works full sections of Wikipedia articles are parsed, entities are identified, and the verb between the entities is taken as the relation. They also extract relations that are not backed by a link in Wikipedia, resulting in common-sense factoids such as 'Brescia is a city'. For a domain specific application this approach lacks precision. In our approach, we care more for high precision in finding relations than for recall; hence, we carefully pre-select ontological classes among which relations need to be found, and use these as filters on our search.

The usefulness of the link structure in Wikipedia has been remarked upon by (Völkel et al., 2006). They acknowledge that the link structure in Wikipedia denotes a potentially meaningful relation between two articles, though the relation type is unknown. They propose an extension to the editing software of Wikipedia to enable users to define the type of relation when they add a link in Wikipedia. Potentially this can enrich Wikipedia tremendously, but the work involved would be tremendous as well. We believe some of the type information is already available through the linguistic content of Wikipedia.

3 Data Preparation

3.1 Data

The data used in this work comes from a manually created, non-relational research database of a collection of reptiles and amphibians at a natural history museum. The information contained in the cells describes when a specimen entered the collection, under what circumstances it was collected, its current location, registration number, etc. We argue that the act of retrieving information from this flat database could be enhanced by providing a meta-structure that describes relations between the different database columns. If for instance a relation of the type "is part of" can be defined between the database columns *province* and *country*, then queries for specimens found at a particular location can be expanded accordingly.

Even though the main language of the database is Dutch, we still chose to use the English Wikipedia as the resource for retrieval of relation label candidates. Explicitly choosing the English Wikipedia has as a consequence that the relation labels we are bound to discover will be English phrases. Furthermore, articles in the English Wikipedia on animal taxonomy have a broader coverage and are far more elaborate than those contained in the Dutch Wikipedia. Since these database values use a Latin-based nomenclature, using the wider-coverage English Wikipedia yields a much higher recall than the Dutch Wikipedia. The values of the other columns mainly contain proper names, such as person names and geographic locations and dates, which are often the same; moreover, English and Dutch are closely related languages. Different names exist for different countries in each language, but here the inconsistency of the database aids us, as it in fact contains many database entries partially or fully in English, as well as some in German and Portuguese.

The database contains 16,870 records in 39 columns. In this work we focus on 20 columns; the rest are discarded as they are either extrinsic features not directly pertaining to the object they describe, e.g., a unique database key, or elaborate textual information that would require a separate processing approach. The columns we focus on describe the position of the specimen in the zoological taxonomy (6 columns), the geographical location in which it was found (4 columns), some of its physical properties (3 columns), its collector

Column Name	Value
Taxonomic Class	Reptilia
Taxonomic Order	Crocodylia Amphisbaenia
Taxonomic Genus	Acanthophis Xenobatrachus
Country	Indonesia Suriname
Location	city walls near Lake Mahalona
Collection Date	01.02.1888 02.01.1995
Type	holotype paralectotype
Determinator	A. Dubois M. S. Hoogmoed
Species defined by	(Linnaeus, 1758) (LeSueur, 1827)

Table 1: Example classes from test data

and/or determiner, donator and associated date (4 columns), and other information (3 columns). The values in most columns are short, often consisting of a single word. Table 1 lists some example database values.

3.2 Preprocessing

As the database was created manually, it was necessary to normalise spelling errors, as well as variations on diacritics, names and date formats. The database values were also stripped of all non-alphanumeric characters.

In order to find meaningful relations between two database columns, query pairs are generated by combining two values occurring together in a record. This approach already limits the number of queries applied to Wikipedia, as no relations are attempted to be found between values that would not normally occur together. This approach yields a query pair such as *Reptilia Crocodylia* from the taxonomic class and order columns, but not *Amphibia Crocodylia*. Because not every database field is filled, and some combinations occur more often, this procedure results in 186,141 query pairs.

For this study we use a database snapshot of the English Wikipedia of July 27, 2008. This dump contains about 2.5 million articles, including a vast amount of domain-specific articles that one would typically not find in general encyclopaedias. An index was built of a subset of the link structure present in Wikipedia. The subset of links included in the index is constrained to those links occurring in sentences from each article in which the main topic of the Wikipedia article (as taken from the title name) occurs. For example, from the Wikipedia article on *Anura* the following sentence would be included in the experiments[1]:

The frog is an [[amphibian]] in the order Anura (meaning "tail-less", from Greek an-, without + oura, tail), formerly referred to as Salientia (Latin saltare, to jump)

whereas we would exclude the sentence:

An exception is the [[fire-bellied toad]] (Bombina bombina): while its skin is slightly warty, it prefers a watery habitat.

This approach limits the link paths to only those between pages that are probably semantically strongly connected to each other. In the following section the computation of the link paths indicating semantic relatedness between two Wikipedia pages is explained.

3.3 Computing Semantic Relatedness

Relation discovery between terms (instantiations of different ontological classes) that have a page in Wikipedia is best performed after establishing if a sufficiently strong relation between the two terms under consideration actually exists. To do this, the semantic relatedness of those two terms or concepts needs to be computed first. Semantic relatedness can denote every possible relation between two concepts, unlike semantic similarity, which typically denotes only certain hierarchical relations (like hypernymy and synonymy) and is often computed using hierarchical networks like WordNet (Budanitsky and Hirst, 2006).

A simple and effective way of computing semantic relatedness between two concepts c_1 and c_2 is measuring their distance in a semantic network. This results in a semantic distance metric, which can be inversed to yield a semantic relatedness metric. Computing the path-length between terms c_1 and c_2 can be done using Formula 1 where P is the set of paths connecting c_1 to c_2 and Np is the number of nodes in path p.

[1]The double brackets indicate Wikilinks

$$rel_{path}(c_1, c_2) = argmax_{p \in P} \frac{1}{N_p} \quad (1)$$

We search for shortest paths in a semantic network that is constructed by mapping the concepts in Wikipedia to nodes, and the links between the concepts to edges. This generates a very large network (millions of nodes and tens of millions of edges), but due to the fact that Wikipedia is scale-free (Barabasi and Albert, 1999) (its connectedness degree distribution follows a power-law), paths stay relatively short. By indexing both incoming and outgoing links, a bidirectional breadth-first search can be used to find shortest paths between concepts. This means that the search is divided in two chains: a forward chain from c_1 and a backward chain to c_2. As soon as the two chains are connected, a shortest path is found.

4 Extracting Relations from Wikipedia

Each query pair containing two values from two database columns are sent to the system. The system processes each term pair in four steps. A schematic overview of the system is given in Figure 1.

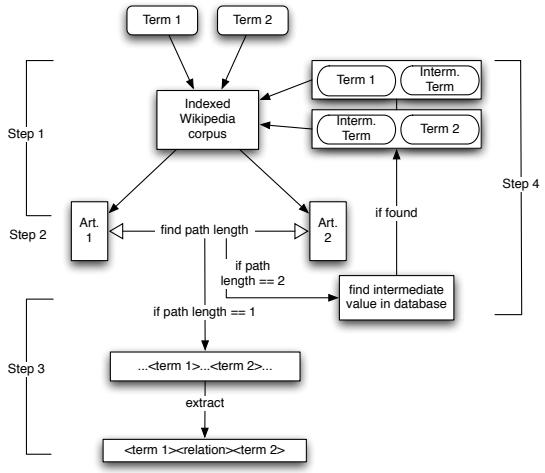

Figure 1: Schematic overview of the system

Step 1 We look for the most relevant Wikipedia page for each term, by looking up the term in titles of Wikipedia articles. As Wikipedia formatting requires the article title to be an informative and concise description of the article's main topic, we assume that querying only for article titles will yield reliable results.

Step 2 The system finds the shortest link path between the two selected Wikipedia articles. If the path distance is 1, this means that the two concepts are linked directly to each other via their Wikipedia articles. This is for instance the case for *Megophrys* from the genus column, and *Anura* from the order column. In the Wikipedia article on *Megophrys*, a link is found to the Wikipedia article on *Anura*. There is no reverse link from *Anura* to *Megophrys*; hierarchical relationships in the zoological taxonomy such as this one are often unidirectional in Wikipedia as to not overcrowd the parent article with links to its children.

Step 3 The sentence containing both target concepts as links is selected from the articles. From the *Megophrys* article this is for instance "*Megophrys is a genus of frogs, order [[Anura]], in the [[Megophryidae]] family*."

Step 4 If the shortest path length between two Wikipedia articles is 2, the two concepts are linked via one intermediate article. In that case the system checks whether the title of the intermediate article occurs as a value in a database column other than the two database columns in focus for the query. If this is indeed the case, the two additional relations between the first term and the intermediate article are also investigated, as well as the second term and that of the intermediate article. Such a bridging relation pair is found for instance for the query pair *Hylidae* from the taxonomic order column, and *Brazil* from the country column. Here, the initial path we find is *Hylidae* $\leftrightarrow$ *Sphaenorhynchys* $\rightarrow$ *Brazil*. We find that the article-in-the-middle value (*Sphaenorhynchys*) indeed occurs in our database, in the taxonomic genus column. We assume this link is evidence for co-occurrence. Thus, the relevant sentences from the Wikipedia articles on *Hylidae* and *Sphaenorhynchys*, and between articles on *Sphaenorhynchys* and *Brazil* are added to the possible relations between "order" – "genus" and "genus" – "country".

Subsequently, the selected sentences are POS-tagged and parsed using the Memory Based Shallow Parser (Daelemans et al., 1999). This parser provides tokenisation, POS-tagging, chunking, and grammatical relations such as subject and direct object between verbs and phrases, and is based on memory-based classification as implemented in TiMBL (Daelemans et al., 2004). The five most frequently recurring phrases that occur

between the column pairs, where the subject of the sentence is a value from one of the two columns, are presented to the human annotators. The cut-off of five was chosen to prevent the annotators from having to evaluate too many relations and to only present those that occur more often, and are hence less likely to be misses. Misses can for instance be induced by ambiguous person names that also accidentally match location names (e.g., *Dakota*). In Section 7 we discuss methods to remedy this in future work.

5 Evaluating Relations from Wikipedia

Four human judges evaluated the relations between the ontological class pairs that were extracted from Wikipedia. Evaluating semantic relations automatically is hard, if not impossible, since the same relation can be expressed in many ways, and would require a gold standard of some sort, which for this domain (as well as for many cultural heritage domains) is not available.

The judges were presented with the five highest-ranked candidate labels per column pair, as well a longer snippet of text containing the candidate label, to resolve possible ambiguity. The items in each list were scored according to the total reciprocal rank (TRR) (Radev et al., 2002). For every correct answer 1/*n* points are given, where *n* denotes the position of the answer in the ranked list. If there is more than 1 correct answer the points will be added up. For example, if in a list of five, two correct answers occur on positions 2 and 4, the TRR would be calculated as $(1/2 + 1/4) = .75$. The TRR scores were normalised for the number of relation candidates that were retrieved, as for some column pairs less than five relation candidates were retrieved.

As an example, for the column pair "Province" and "Genus", the judges were presented with the relations shown in Table 2. The direction arrow in the first column denotes that the "Genus" value occurred before the "Province" value.

The human judges were sufficiently familiar with the domain to evaluate the relations, and had the possibility to gain extra knowledge about the class pairs through access to the full Wikipedia articles from which the relations were extracted. Inter-annotator agreement was measured using Fleiss's Kappa coefficient (Fleiss, 1971).

6 Results and Evaluation

As expected, between certain columns there are more relations than between others. In total 140 relation candidates were retrieved directly, and 303 relation label candidates were retrieved via an intermediate Wikipedia article. We work with the assumption that these columns have a stronger ontological relation than others. For some database columns we could not retrieve any relations, such as the "collection date" field. This is not surprising, as even though Wikipedia contains pages about dates ('what happened on this day'), it is unlikely that it would link to such a domain specific event such as an animal specimen collection. Relations between instances denoting persons and other concepts in our domain are also not discovered through this approach. This is due to the fact that many of the biologists named in the database do not have a Wikipedia page dedicated to them, indicating the boundaries of Wikipedia's domain specific content. Although not ideal, a named-entity recognition filter could be applied to the database after which person names can be retrieved from other resources.

Occasionally we retrieve a Wikipedia article for a value from a person name column, but in most cases this mistakenly matches with a Wikipedia article on a location, as last names in Dutch are often derived from place names. Another problem induced by incorrect data is the incorrect match of Wikipedia pages on certain values from the "Town" and "Province" columns. Incorrect relation candidates are retrieved because for instance the value 'China' occurs in both the "Town" and the "Province" columns. A data cleaning step would solve these two problems.

From each column pair the highest rated relation was selected with which we constructed the ontology displayed in Figure 2. As the figure shows, the relations that are discovered are not only 'is a'-relations one would find in strictly hierarchical resources such as a zoological taxonomy or geographical resource.

The numbers in the relation labels in Figure 2 denote the average TRR scores given by the four judges on all relation label candidates that the judges were presented with for that column pair. The scores for the relations between the taxonomic classes in our domain were particularly high, meaning that in many cases all relation candidates presented to the judges were assessed as

Direction	Label	Snippet
→	is found in	is a genus of venomous pitvipers found in Asia from Pakistan, through India,
→	is endemic to	Cross Frogs) is a genus of microhylid frogs endemic to Southern Philippine,
→	are native to	are native to only two countries: the United States and
→	is known as	is a genus of pond turtles also known as Cooter Turtles, especially in the state of

Table 2: Relation candidates for Province and Genus column pair

correct. The inter-annotator agreement was $\kappa = 0.63$, which is not perfect, but reasonable. Most disagreement is due to vague relation labels such as 'may refer to' as found between "Province" and "Country". If a relation that occurred fewer than 5 times was judged incorrect by the majority of the judges the relation was not included in Figure 2.

Manual fine-tuning and post-processing of the results could filter out synonyms such as those found for relations between "Town" and other classes in the domain. This would for instance define one particular relation label for the relations 'is a town in' and 'is a municipality in' that the system discovered between "Town" and "Province" and "Town" and "Country", respectively.

7 Conclusion and Future Work

In this work we have shown that it is possible to extract ontological relation labels for domain-specific data from Wikipedia. The main contribution that makes our work different from other work on relation extraction from Wikipedia is that the link structure is used as a strong indication of the presence of a meaningful relation. The presence of a link is incorporated in our system by only using sentences from Wikipedia articles that contain links to other Wikipedia articles. Only those sentences are parsed that contain the two terms we aim to find a relation between, after which the verb phrase and possibly the article or preposition following it are selected for evaluation by four human judges.

The advantage of the pre-selection of content that may contain a meaningful relation makes our approach fast, as it is not necessary to parse the whole corpus. By adding the constraint that at least one of the query terms should be the subject of a sentence, and by ranking results by frequency, our system succeeds in extracting correct and informative relations labels. However, there is clearly some room for improvement, for instance in the coverage of more general types of information such as dates and person names. For this we intend to incorporate more domain specific resources, such as research papers from the domain that may mention persons from our database. We are also looking into sending queries to the web, whilst keeping the constraint of hyperlink presence.

Another factor that may help back up the relations already discovered is more evidence for every relation. Currently we only include sentences in our Wikipedia corpus that contain the literal words from the title of the article, to ensure we have content that is actually about the article and not a related topic. This causes many sentences in which the topic is referred to via anaphoric expressions to be missed. (Nguyen et al., 2007) take the most frequently used pronoun in the article as referring to the topic. This still leaves the problem of cases in which a person is first mentioned by his/her full name and subsequently only by last name. Coreference resolution may help to solve this, although accuracies of current systems for encyclopaedic text are often not much higher than baselines such as those adopted by (Nguyen et al., 2007).

Errors in the database lead to some noise in the selection of the correct Wikipedia article. The queries we used are mostly single-word and two-word terms, which makes disambiguation hard. Fortunately, we have access to the class label (i.e., the database column name) which may be added to the query to prevent retrieval of an article about a country when a value from a person name column is queried. We would also like to investigate whether querying terms from a particular database column to Wikipedia can identify inconsistencies in the database and hence perform a database cleanup. Potentially, extraction of relation labels from Wikipedia articles can also be used to assign types to links in Wikipedia.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments. This research

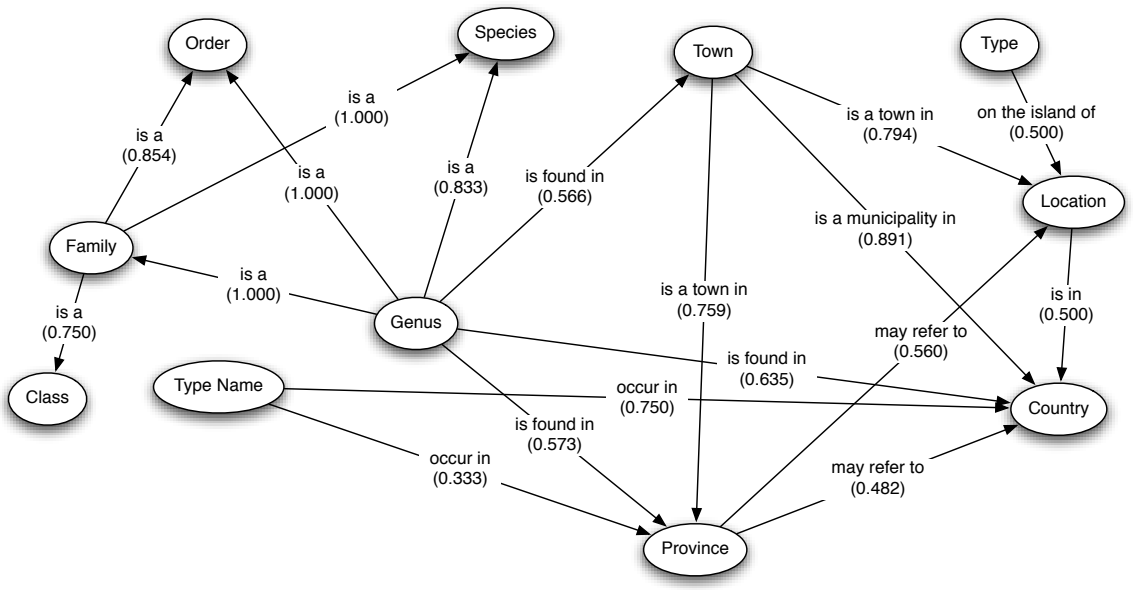

Figure 2: Graph of relations between columns, with TRR scores in parentheses

was funded as part of the Continuous Access to Cultural Heritage (CATCH) programme of the Netherlands Organisation for Scientific Research (NWO).

References

Sören Auer and Jens Lehmann. 2007. What have innsbruck and leipzig in common? extracting semantics from wiki content. In Franconi et al., editor, *Proceedings of European Semantic Web Conference (ESWC'07)*, volume 4519 of *Lecture Notes in Computer Science*, pages 503–517, Innsbruck, Austria, June 3 - 7. Springer.

A. L. Barabasi and R. Albert. 1999. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October.

Sebastian Blohm and Philipp Cimiano. 2007. Using the web to reduce data sparseness in pattern-based information extraction. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Warsaw, Poland, September. Springer.

A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Sergey Chernov, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou. 2006. Extracting semantic relationships between wikipedia categories. In *Proceedings of the First Workshop on Semantic Wikis - From Wiki to Semantics [SemWiki2006] - at ESWC 2006*, pages 153 – 163, Karlsruhe, Germany, May 15.

Walter Daelemans, Sabine Buchholz, and Jorn Veenstra. 1999. Memory-based shallow parsing. In *Proceedings of CoNLL'99*, pages 53–60, Bergen, Norway, June 12.

Walter Daelemans, Jakub Zavrel, Ko Van der Sloot, and Antal Van den Bosch. 2004. Timbl: Tilburg memory based learner, version 5.1, reference guide. Technical Report 04-02, ILK/Tilburg University.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature*, 438:900–901.

Jaap Kamps and Marijn Koolen. 2008. The importance of link evidence in wikipedia. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Rutven, and Ryen W. White, editors, *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008)*, volume 4956 of *Lecture Notes in Computer Science*, pages 270–282, Glasgow, Scotland, March 30 - April 3. Springer Verlag.

Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. Wikipedia link structure and text mining for semantic relation extraction towards a huge scale global web ontology. In *Proceedings of SemSearch 2008 CEUR Workshop*, pages 59–73, Tenerife, Spain, June 2.

Dat P. T. Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Exploiting syntactic and semantic information for relation extraction from wikipedia. In *Proceedings of Workshop on Text-Mining & Link-Analysis (TextLink 2007) at IJCAI 2007*, pages 1414–1420, Hyderabad, India, January 7.

Dragomir R. Radev, Hong Q, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *Demo section, LREC 2002*, Las Palmas, Spain, June.

F. M. Suchanek, G. Ifrim, and G. Wiekum. 2006. Leila: Learning to extract information by linguistic analysis. In *Proceedings of the ACL-06 Workshop on Ontology Learning and Population*, pages 18–25, Sydney, Australia, July.

Zareen Saba Syed, Tim Finin, and Anupam Joshi. 2008. Wikitology: Using wikipedia as an ontology. Technical report, University of Maryland, Baltimore County.

Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. 2006. Semantic wikipedia. In *WWW 2006*, pages 585–594, Edinburgh, Scotland.