Automatic Creation of Multilingual Semantic Networks from Wikipedia

Océane Chabrol, David Norrestam, Pierre Nugues

Department of computer science
Lund University, Lund
chabrol@ensta.fr, david.norrestam.289@student.lu.se, pierre.nugues@cs.lth.se

Abstract

This paper describes the automatic creation of semantic networks from Wikipedia. Following Lipczak et al. (2014), we constructed the graphs corresponding to the semantic networks by merging across languages the categories manually assigned by the users. This results in a network of related concepts for each entity of Wikipedia. We used these networks as a component of an entity linking system. the networks showed they could improve the results by 1% over an already strong baseline.

1. Introduction

Semantic networks are a way to represent relations between concepts and entities. They consist of graphs, where the nodes are concepts and the arcs, the relations. WordNet (Fellbaum, 1998) is a well-known example of semantic network that has found its way in a large number of applications.

In the case of entity linking, where a mention in a text can refer to two or more entities, semantic networks can provide a context for each of the candidates and help the disambiguation of the mention. *Dublin*, for example, is mainly known as the capital of Ireland, but there exist cities called Dublin in Canada, Belarus, Australia, and ten in the United States. Semantic networks for each of these entities (cities) would link them to different nodes such as Europe, Georgia, or Ohio, that would match (or not) the words surrounding a mention of *Dublin*.

In the rest of the paper, we describe the automatic creation of multilingual semantic networks from categories we gathered from Wikipedia. This enabled us to model and access the context of a word.

2. Previous Work

Peirce (1909) mentioned the possibility to represent knowledge in the form of graphs at the beginning of the 20^{th} century. His idea was to represent logical relations between ideas with links. Figure 1 shows the representation of the assertion

> If a farmer owns a donkey, then he beats it

from Sowa (1997) as an *existential graph*.

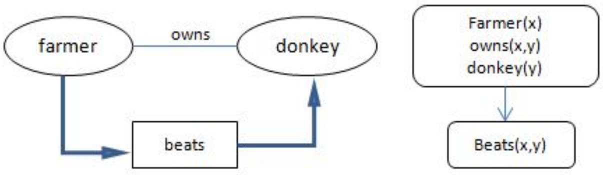

Figure 1: Existential graph. After Sowa (1997)

Quillian (1967) created a graph representing associations between concepts in an human-like way. The concepts were nodes linked by associations. It is the beginning of semantic graphs. More recently, Navigli and Velardi (2005) used semantic graphs derived from WordNet for sense disambiguation and Han and Zhao (2010) applied them to entity disambiguation, where they built a semantic graph from sentences, scoring the relatedness between concepts.

Lipczak et al. (2014) described an entity disambiguation algorithm using a semantic graph, where they linked entities to their categories in Wikipedia. Our work is based on their idea.

3. Resources

We created the semantic graphs from information we gathered from Wikipedia, and more specifically from the relations between Wikipedia articles and their categories. This section gives a short introduction to Wikipedia, and the features we used, categories and Q-numbers. We also outline DBpedia and the structure of the Wikipedia dumps.

3.1 Wikipedia

Wikipedia is the largest online encyclopedia with millions of articles in more than 200 languages. As a rule, every article on an entity or a concept should have at least one category and up to twenty. Dublin, for instance, the capital of Ireland, belongs to 13 categories in the English version: Dublin (city), 841 establishments, Capitals in Europe, Cities in the Republic of Ireland, County towns in the Republic of Ireland, Leinster, Local administrative units of the Republic of Ireland, Populated coastal places in the Republic of Ireland, Port cities and towns of the Irish Sea, Staple ports, University towns in Ireland, Viking Age populated places, and Populated places established in the 9th century.

The categories of an article are different across the languages. In German, the article on Dublin (Ireland) only belongs to 7 categories. This means the different versions of Wikipedia can be used as independent data sources.

3.2 Categories and Subcategories

Wikipedia categories are organized as a hierarchical network. For example, the category *Dublin (city)* is a subcategory of five categories: County Dublin, Capitals in Europe, Cities in the Republic of Ireland, Populated coastal places in the Republic of Ireland.

For some articles, these parent or ancestor categories (categories higher in the hierarchy) can be more relevant

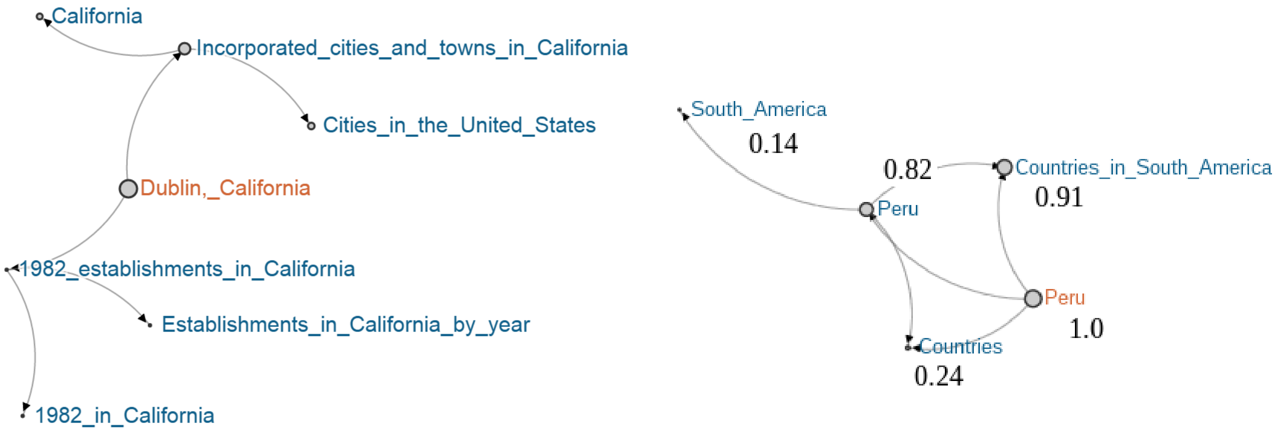
California
Incorporated_cities_and_towns_in_California
Cities_in_the_United_States
Dublin,_California
1982_establishments_in_California
Establishments_in_California_by_year
1982_in_California
South_America
0.14
0.82
Countries_in_South_America
0.91
Peru
Peru
1.0
Countries
0.24

Figure 2: Category graphs of Dublin, California (left) and Peru (right)

than the ones directly assigned by the Wikipedia editors, that can be too restrictive. In Fig. 2, Dublin belongs to *Cities in the Republic of Ireland*, but not to *Ireland*. A graph using direct categories only would then link Dublin to places in Ireland, but not to Irish movies, politicians, or museums.

Dublin, California shows the same limitations. It has only two direct categories: *Incorporate cities and towns in California* and *1984 establishments in California*. We can extend them with their parent categories, among others, *California* and *Cities in the United States*.

To broaden the categories, we added their parents in the creation of our semantic network; see Sect. 4.

3.3 Q-numbers

Almost all the articles and categories on Wikipedia are assigned a unique identifier called a Q-number. This number enables to retrieve all the language versions of an entity. Sweden, for example, has the number Q34, and links to the articles: Sweden, Sverige, Schweden, Suède, Suecia, etc.

3.4 DBpedia

DBpedia (Bizer et al., 2009) is a crowd-sourced community effort to extract structured information from Wikipedia and make it available on the Web.

Instead of parsing Wikipedia to obtain both the article/category tree and the category tree, we used dumps from DBpedia dated from December 2015. These dumps consist of two files for each language. The first one contains entries for each article listing all its categories; the second one contains entries for each category listing all its parent categories.

4. Implementation

We divided the implementation into three steps:

1. For each language, parse the data;
2. Merge the parsed data from different languages;
3. Create the graph.

The first two steps are a one-time process assuming the information we are considering (Wikipedia articles and their corresponding categories) are unchanged and that the process can produce a re-usable data-structure (semantic graph) modeling the result. In order to save space, the data structure is divided into smaller parts, one per article/category. Because of this, one last step is necessary. These need to be combined, in order to produce a semantic graph. These steps are explained in more detail below.

4.1 Parsing

We first parse the DBpedia files language per language. Their size ranges from 300 Mb to 5 Gb. We extracted the articles and their categories from each language dump and we translated the names to Q-numbers. This resulted in one JSON object for each article as shown in Table 1.

Article	Categories	Article	Categories
Q34	Q4884449	Q62132	Q8583058
	Q4366558		Q9472290
	Q4587626		Q6307954
	Q4368475		Q6963927
			Q6984119
			Q10211395
			Q9443016

Table 1: Categories extracted from one language (here Swedish) for Sweden (Q34) and Lapland (Q62132)

4.2 Merging

In the second step, we collected the categories of every language in one file and we created one universal JSON-object per article. Table 2 shows the results for Q34/Sweden, where we kept only the five best categories. A is the number of language versions for the article, and C is the number of language versions that have this category (at most A).

4.3 Creation of the Graph

The semantic graph is constructed incrementally, starting from the article, adding the categories it belongs to in the form of child nodes, and repeating this process for each child node (category) we add. When two categories have the same Q-number, we merge them into a single node.

Article	A	Categories	C	Ratio	Description
Q34	113	Q4368475	97	0.85	Sweden
		Q4366558	37	0.33	Member states of the European Union
		Q4587626	29	0.26	Countries in Europe
		Q7162174	19	0.17	Scandinavia
		Q7363642	19	0.17	Constitutional monarchies

Table 2: Data from 123 languages for the article on Sweden (Q34): It has versions 113 languages and it is tagged 97 times in the Sweden category and 37 times in Member states of the European Union

4.31 Width and Depth

To avoid very large graphs, we restrict the size with two parameters: the width and the depth.

The width defines the maximal number of new nodes (categories) to include for each parent node (article or category). The depth defines how far up in the category hierarchy one wants to expand the graph.

A graph with a width of 5 and a depth of 1, for example, will simply contain a parent node (the article) and the five most frequent categories for this article. If we extend the depth to two, the graph will also take into consideration the subgraph of width 5 (and depth 1) and, for each of the categories, their five most frequent categories.

4.32 Calculation of Ratios

Finally, we rank the category importance by calculating their ratios using the formulas:

$$ratio_1(a, c) = \frac{N_c(a)}{N_a} \qquad (1)$$

$$ratio_n(a, c) = \sum_{p \in c.parents} (ratio_n(a, p) \times ratio_1(c, p)), \quad (2)$$

where N_a is the number of languages the article exists in, and $N_c(a)$ is the number of languages where article a belongs to category c.

The first equation refers to the calculation of a direct relationship between article a and category c. This is the only one needed when the depth of the graph is set to one. When the depth is greater than one, the second equation is necessary to take into account categories with multiple links to the parent article. It is used to find categories of categories. When a category can be reach from different paths (such as Countries in Figure 2, its ratios is the sum of the ratios obtained from each part. Equation 2 basically says that for each parent of category c, multiply the ratio between category c and its parent with the ratio between the parent and article a (n refers to the depth we are looking at). A category with a higher ratio is a more relevant category.

Table 2 shows category ratios obtained with Sweden (only Eq. 1 is used because we are only looking to the direct categories of Sweden).

5. Results

We implemented the program so that it produces a unique graph from the DBpedia dump that we store as a JSON map. The processing time is about 45 minutes for 123 languages and the map has a size of 800 Mb.

Once created, the map loads in one minute, and from this map we can extract the semantic graph of an entity. The output is also a JSON object with the most relevant categories and their corresponding ratios. We can adjust its size with the width and depth parameters. Table 3 shows the 15 best categories we obtained with Mahatma Gandhi with width 5 and depth 5 (there are 297 categories in this case); the Universal Declaration of Human Rights using width of 5 and a depth of 2; and the Four Seasons with width 4 and depth 3 where we only show categories with ratios higher than 0.15.

6. Data Visualization

We developed an interactive service that enables the users to experiment with the graph. The user has first to select a language and then type the named entity with a depth and a width. The graph is generated with Q-numbers that are translated in the selected language.

While Lipczak et al. (2014) used a similar interface, a user can only retrieve graphs for English. In contrast, our system is multilingual. Since some entities only exist in certain languages, the translation from a Q-number to the chosen language might fail. The interface will simply display the Q-number then. Figure 3 shows an example of this with the words *Scania* and *Skåne*, where Q6056748, *Administrativa indelningar av länder* 'Country subdivisions by country' has no page in English.

7. Applications

We believe our multilingual semantic network useful to applications where it is needed to classify or gather entities. For example it could be used to classify politicians from a text by country, you would just have to look for the first country category in each politician network/graph. An other example could be to get an idea of the subject of an article. We could collect all entities and find the best common categories of all the entities from the text. It could also be used to keep only the thrillers from a list of movies. We could find numerous other examples where this graph would be useful, every time we need to select, gather, classify, create a context.

We integrated it as a component of an entity linker, where it showed it could improve the system by 1% (Södergren et al., 2016).

8. Future Work

The system can be improved in many ways:

- Instead of DBPedia, we could build the map using Wikipedia resources directly. This will enable us to use all the 200 languages of Wikipedia.

Mahatma Gandhi		The Universal Declaration of Human Rights		The Four Seasons	
Categories	**Ratio**	**Categories**	**Ratio**	**Categories**	**Ratio**
People by occupation	0.68	Human Rights	0.55	Compositions by Antonio Vivaldi	0.63
Politics	0.65	United Nations	0.41	Musical compositions	0.6
India	0.62	International Organizations	0.25	Compositions by composer	0.52
People	0.64	Humans	0.21	Violin concertos	0.5
Indian politicians	0.53	Rights	0.18	Concertos	0.41
Indian people	0.50	International Law	0.17	Compositions for violin	0.37
Politics by country	0.47	United Nations General Assembly resolutions	0.13	Music	0.31
1948 deaths	0.45	United Nations General Assembly	0.08	Classical compositions	0.27
1869 births	0.45	United Nations resolutions	0.07	Composition by instrumentation	0.21
Politicians by nationality	0.44	Peace	0.06	Antonio Vivaldi	0.20
Politicians	0.41	Intergovernmental organizations	0.06	Violins	0.20
19^{th} century	0.40			Composition by musical form	0.20
People by nationality	0.35			Composition for symphony concertos	0.18
20^{th} century	0.35			Classical music	0.16
Politics of India	0.34				

Table 3: The categories found for Mahatma Gandhi (left), the Universal Declaration of Human Rights (middle), and The Four Seasons by Vivaldi (right)

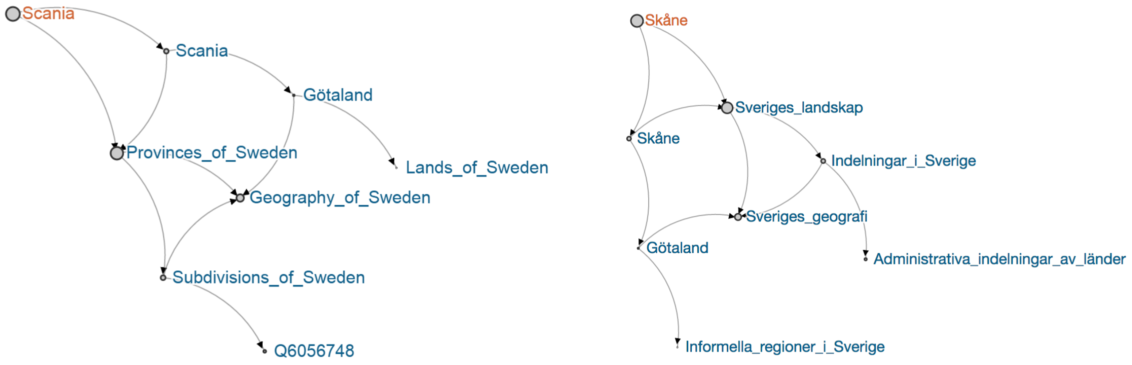

Figure 3: Scania (left) and Skåne (right) with a width of 2 and a depth of 3. Note the unresolved Q-number in the English version

- Wikipedia is pervaded with many automatically created categories such as *establishments created in ***** or *Birth in *****. Most of the time, they show high ratios because they are created automatically in every language. They are however too generic and do not provide specifically interesting information. Depending on the use we consider, those categories could be ignored.

Acknowledgements

This research was supported by Vetenskapsrådet, the Swedish research council, under the *Det digitaliserade samhället* program.

References

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia—a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech and Communication)*. MIT Press, Cambridge, Massachusetts.

Xianpei Han and Jun Zhao. 2010. Structural semantic relatedness: A knowledge-based method to named entity disambiguation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 50–59.

Marek Lipczak, Arash Koushkestani, and Evangelos Milios. 2014. Tulip: Lightweight entity recognition and disambiguation using wikipedia-based topic centroids. In *Proceedings of ERD 14*.

Roberto Navigli and Paola Velardi. 2005. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *EEE transactions on pattern analysis and machine intelligence*, 27(7):1075–1086.

Charles Sanders Peirce. 1909. Existential graphs, manuscript 514. Available at `http://www.jfsowa.com/peirce/ms514.htm`.

M. Ross Quillian. 1967. Word concepts: a theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12(5):410–430.

John F. Sowa. 1997. Peircean foundations for a theory of context. In *Conceptual Structures: Fulfilling Peirce's Dream*, pages 41–64. Springer-Verlag, Berlin.

Anton Södergren, Marcus Klang, and Pierre Nugues. 2016. A multilingual entity linker using pagerank and semantic graphs. In *Submitted*.