Decoding Wikipedia Categories for Knowledge Acquisition

Vivi Nastase and **Michael Strube**
EML Research gGmbH
Heidelberg, Germany
{nastase,strube}@eml-research.de

Abstract

This paper presents an approach to acquire knowledge from Wikipedia categories and the category network. Many Wikipedia categories have complex names which reflect human classification and organizing instances, and thus encode knowledge about class attributes, taxonomic and other semantic relations. We decode the names and refer back to the network to induce relations between concepts in Wikipedia represented through pages or categories. The category structure allows us to propagate a relation detected between constituents of a category name to numerous concept links. The results of the process are evaluated against ResearchCyc and a subset also by human judges. The results support the idea that Wikipedia category names are a rich source of useful and accurate knowledge.

Introduction

When people understand language, lexical, common sense and world knowledge all come into play. While acquiring lexical knowledge and making it available in machine readable form is a feasible task – examples include WordNet (Fellbaum, 1998), machine readable dictionaries and concordances from corpora (Kilgarriff et al., 2004) – common sense and world knowledge on a scale large enough to be useful for natural language processing (NLP) still eludes us. Current trends in knowledge acquisition aim at collecting "bite-sized" pieces – simple facts expressed as relations – in the quest of getting closer to gathering more complex representations (Schubert, 2006). The approaches range from human-based contributions to fully automatic methods to mine knowledge from texts. Most automatic approaches start with seed concepts or patterns, or mine for instances of a prespecified set of relations.

We join these efforts in between the two extremes, by mining for knowledge in Wikipedia categories and the category network. Contributors to a Wikipedia page are encouraged to link this page to the existing category network, and to create new categories as necessary. From this process has emerged a "folksonomy" – a collaborative organizing backbone. The category names reflect our intuitions about classification and organization: BOOKS BY GENRE covers CHILDREN'S BOOKS, REFERENCE WORKS, TEXTBOOKS, NOVELS (BY GENRE), and so on, NEWSPAPERS PUBLISHED BY NEWSQUEST covers EVENING TIMES, THE OXFORD TIMES and others[1]. We develop methods that automatically decode these strings and determine the relations, classes and attributes they encode. We induce numerous instances of the relations detected between constituents of a category name using the category network. The quality of the information extracted is assessed against ResearchCyc, the largest (manually created) knowledge resource available, and a subset through manual evaluation.

The novelty of our knowledge acquisition approach is that we focus on very small fragments that encode a variety of human knowledge about concepts and relations, and then propagate this knowledge by traversing the category network downward, toward the pages. We regard each category name as a repository of knowledge, that we can gain access to directly by decoding the patterns in this string. Using the category network we are able to propagate the information extracted from a category name, and multiply the relations extracted by finding their instances. The evaluation shows that this approach leads to the extraction of high quality relations, more numerous and more qualitative than what we would get by mining for the same relations in the Wikipedia articles using patterns, as it is commonly done in text-based relation extraction (Hearst, 1992; Berland & Charniak, 1999; Zhao & Grishman, 2005).

This research complements the type of work described in Ponzetto & Strube (2007), where the category network is transformed into a taxonomy. We now replace nodes in this taxonomy that have an organizational purpose with concepts and relations that reflect semantic and associative links.

Wikipedia Category Names and Network

To organize Wikipedia for easy access to pages, contributors are given guidelines for categorizing articles and naming new categories[2]. Many categories – ALBUMS BY ARTIST,

[1] We use Sans Serif for patterns and words, *italics* for relations, SMALL CAPS for Wikipedia categories and pages, and **BOLD SMALL CAPS** for concepts.

[2] http://en.wikipedia.org/wiki/Wikipedia:Categorization
http://en.wikipedia.org/wiki/Wikipedia:Naming_conventions_(categories)

Category type	Category name	Pattern	Relations
explicit relation	QUEEN (BAND) MEMBERS	X members members of X	**FREDDY MERCURY** *member_of* **QUEEN (BAND)** **BRIAN MAY** *member_of* **QUEEN (BAND)** ...
explicit relation	MOVIES DIRECTED BY WOODY ALLEN	X [VBN IN] Y	**ANNIE HALL** *directed_by* **WOODY ALLEN** **ANNIE HALL** *isa* **MOVIE** **DECONSTRUCTING HARRY** *directed_by* **WOODY ALLEN** **DECONSTRUCTING HARRY** *isa* **MOVIE** ...
partly explicit relation	VILLAGES IN BRANDENBURG	X [IN] Y	**SIETHEN** *located_in* **BRANDENBURG** **SIETHEN** *isa* **VILLAGE** ...
implicit relation	MIXED MARTIAL ARTS TELEVISION PROGRAMS	X Y	**MIXED MARTIAL ARTS** $\mathcal{R}$ **TELEVISION PROGRAMS** **TAPOUT (TV SERIES)** $\mathcal{R}$ **MIXED MARTIAL ARTS** **TAPOUT (TV SERIES)** *isa* **TELEVISION PROGRAM** ...
class attribute	ALBUMS BY ARTIST	X by Y	**ARTIST** *attribute_of* **ALBUM** **MILES DAVIS** *isa* **ARTIST** **BIG FUN** *isa* **ALBUM** ...

Table 1: Examples of information encoded in category names and the knowledge we extract

VILLAGES IN BRANDENBURG, MEMBERS OF THE EUROPEAN PARLIAMENT – do not correspond to the type of lexical concepts we would expect to encounter in texts. Instead, they capture instances of human classification and relations that we can use as a source of information. We identify the following types of category names based on the type of information they encode:

explicit relation categories: This is the case for categories that overtly indicate a relation such as *member_of* – e.g. MEMBERS OF THE EUROPEAN PARLIAMENT – or *caused_by* – AIRPLANE CRASHES CAUSED BY PILOT ERROR. The second type of explicit relation can be identified by searching for a [VBN IN][3] pattern.

partly explicit relation categories: Prepositions, although sometimes ambiguous, are strong indicators of semantic relations (Lauer, 1995). in for example, may indicate a spatial relation – as in VILLAGES IN BRANDENBURG – or a temporal one – CONFLICTS IN 2000. Such ambiguous situations can be resolved using named entity type information, or if this in not available, Wikipedia's category network: supercategories of BRANDENBURG (GEOGRAPHY) and 2000 (CENTURIES, YEARS) indicate which type of relation the category encodes.

implicit relation categories: Categories whose names are complex noun compounds do capture relations, but do not give explicit indicators of what the relation is. The category MIXED MARTIAL ARTS TELEVISION PROGRAMS, for example, has two noun phrase components, **MIXED MARTIAL ARTS** and **TELEVISION PROGRAMS**. The relation encoded in this category is the relation between **MIXED MARTIAL ARTS** and **TELEVISION PROGRAMS** – for example, *topic*.

class attribute categories: Categories with the name following the pattern X by Y – e.g. ALBUMS BY ARTISTS – show a grouping of instances of class X by attribute Y. This indicates generalizations (all pages listed under this category can be generalized to X) and class attributes (Y).

Once we decode the information captured in a category name, we can use the category network to propagate this information. Categories, such as ALBUMS BY ARTIST, are further specified with more detailed subcategories (e.g. MILES DAVIS ALBUMS, U2 ALBUMS, QUEEN (BAND) ALBUMS), and are ultimately linked to pages corresponding to specific albums. Table 1 presents an overview of the knowledge we extract for each category type.

Extracting Knowledge from the Wikipedia Categories and Category Network

We process Wikipedia category names to obtain semantic relations and class attributes, as discussed in the previous section, and then propagate these relations based on the category network. In the following we describe in detail the phases of the knowledge extraction process.

1. Identify the dominant constituent. In category names that match specific patterns such as members of X, X [VBN IN] Y, X [IN] Y, the dominant constituent is identified as X. For complex noun compound categories we extract the dominant constituent from the phrase constituents of the category name with an algorithm similar to head identification.

Example: CHAIRMEN FOR THE COUNTY COUNCILS OF NORWAY has three constituents: chairmen, county councils, Norway, with the dominating constituent chairmen.

2. Extract relations. We first gather the pages $\{P_i\}$ categorized under current category C[4], and then add relations according to the category name type:

[3] VBN is the part of speech for participles and IN is the part of speech for prepositions in the Penn Treebank set (Santorini, 1990). We delimit POS patterns with square brackets.

[4] A category is not further expanded if it has a homonymous page. The reason is that a category can cover a wide variety of aspects related to the concept it represents, whereas the page is very specific. E.g., ROME has as subcategories ANCIENT ROME, CULTURE OF ROME, EDUCATION IN ROME, HISTORY OF ROME, etc.

explicit relation categories 1: Certain words imply a relationship, such as member, president, CEO. When such a word is encountered in a category name it indicates that the pages linked to this category correspond to concepts that can be linked through this relation to the organization/group/... mentioned in the category name. For now we focus on the *member* relation, and for members of X or X members categories, add relations $\mathbf{P}_i$ *member_of* $\mathbf{X}$.

explicit relation categories 2: For X [VBN IN] Y categories, add relations $\mathbf{P}_i$ *[VBN IN]* $\mathbf{Y}$ and $\mathbf{P}_i$ *isa* $\mathbf{X}$.

partly explicit relation categories: For X [IN] Y categories, determine the relation $\mathcal{R}$ between $\mathbf{X}$ and $\mathbf{Y}$ based on the preposition [IN] and supercategories of X and Y. We use rules that rely on Gen_x, Gen_y – the named entity type (if it is informative) or the generalizations of $\mathbf{X}$ and $\mathbf{Y}$ in the category network –, such as:

- if Gen_x is person or people, and Gen_y is organization or group, the relation assigned is *member_of*;
- if Gen_y is location or geography, the relation assigned is *spatial*. Once a spatial relation is detected, specifications can be made based on the connecting preposition (e.g. *located_in* for preposition in, etc.). To facilitate the evaluation process, all spatial relations detected are labeled *spatial*.

Add the relations $\mathbf{P}_i$ *isa* $\mathbf{X}$ and $\mathbf{P}_i$ $\mathcal{R}$ $\mathbf{Y}$.

implicit relation categories: For category names that are complex noun compounds, we use the parse tree to extract all embedded phrases (NP, PP, VP, ADJP, ADVP). An example is presented in Figure 1.

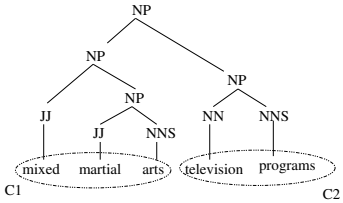


Figure 1: Example of parse tree for a category name.

Each embedded phrase is considered as a constituent C_j of the category name (C_1 = mixed martial arts, C_2 = television programs). Each C_j is dominated by another constituent C_j^D, according to the syntactic structure of the category name (in our example, $C_2 = C_1^D$). The constituent which corresponds to the phrase head is the dominant constituent of the category name, and is denoted by C^D (C_2 is also C^D in the above example).

Figure 2 shows the relations induced for this type of categories. The process is detailed below.

1. add relations $\mathbf{P}_i$ *isa* $\mathbf{C}^D$;
2. form pairs ($\mathbf{C}_j$, $\mathbf{C}^D$) for all C_j for which $C_j^D = C^D$ – form constituent pairs in which the first constituent is dominated by the main dominant constituent. Determine the relation $\mathbf{C}_j$ $\mathcal{R}$ $\mathbf{C}^D$ (detailed below);

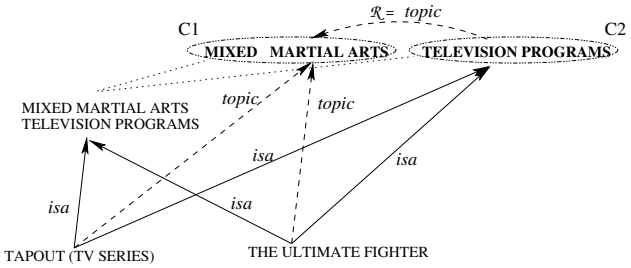
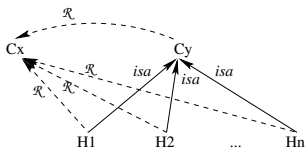


Figure 2: Example of relations induced after extracting components of category name.

3. add relations $\mathbf{P}_i$ $\mathcal{R}$ $\mathbf{C}_j$.
   Propagating the relation $\mathcal{R}$ from the category constituents to the pages follows the rule: *if* $\mathbf{H}_j$ *isa* $\mathbf{C}_y$



and $\mathbf{C}_y$ $\mathcal{R}$ $\mathbf{C}_x \Longrightarrow \mathbf{H}_j$ $\mathcal{R}$ $\mathbf{C}_x$,

Finding the relation between one pair, ($\mathbf{C}_x$, $\mathbf{C}_y$) means automatically finding the relation between numerous ($\mathbf{H}_j$,$\mathbf{C}_x$) pairs.

3. Extract class attributes and attribute values. For categories with names that match *X by Y*, we identify X as a class and Y as an attribute.

Categories with this pattern usually have subcategories that further group the pages, according to values of the class attribute. For example, ALBUMS BY ARTIST has subcategories MILES DAVIS ALBUMS, THE BEATLES ALBUMS, …. We then identify the value of the attribute in the subcategory names. In many cases, like the example presented in Figure 3, X appears in the subcategory name – **albums** *by artist* → *Miles Davis* **albums**. It is then easy to identify the attribute value (Miles Davis for artist), and we add the relation **MILES DAVIS** *isa* **ARTIST**, as shown in Figure 3.

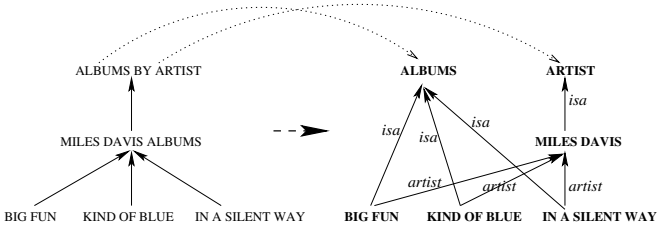


Figure 3: Relations inferred from "by" categories

Not all situations follow the patterns described above: the category HEADS OF GOVERNMENT BY COUNTRY is an example. Subcategories of this category include PRIME MINISTERS OF CANADA, CHANCELLORS OF GERMANY. In this situation we start processing the attribute first (Y):

Category type	# categories	# relations extracted	Evaluation		
			P	manual $\cap$	manual $\cup$
explicit relations	3,450	86,649			
caused_by, *based_in*, *written_by*, ...	2,152	43,938	-	94.37%	96.38%
member_of	1,298	42,711	24% (25)	95.56%	97.17%
partly explicit and implicit relation categories	98,855	9,751,748			
isa		3,400,243	44.57% (6,250)	76.4%	84%
spatial		3,201,125	39.69% (1,325)	87.09%	97.98%

Table 2: Extracted relations and evaluation results

- if the attribute is a category in Wikipedia, collect the pages it subsumes (P_i) as possible attribute values;
- if a P_i appears in the subcategory name, it serves as confirmation that this is a possible attribute value and we add the link $\mathbf{P}_i$ *isa* $\mathbf{Y}$,
- extract the remainder of the subcategory name as an instance of $\mathbf{X}$.

In the example above, *X = heads of government*, *Y = country*. We expand COUNTRY[5] to all its pages, and test whether any of them appear in the name of the subcategory PRIME MINISTERS OF CANADA. We identify *P = Canada*, and add the links **CANADA** *isa* **COUNTRY** and **PRIME MINISTER** *isa* **HEADS OF GOVERNMENT**.

Results

Processing the Wikipedia categories starts with loading the category network – in the form of nodes and category links extracted from Wikipedia dumps[6] – and filtering out administrative categories (identified using keywords, e.g. stubs, articles, wikipedia). After this preprocessing, there are 197,667 categories in the network. The category names are processed with the POS tagger, parser and named entity recognizer developed by the Stanford NLP group[7].

explicit relations categories:

VBN IN pattern: 2,152 category names match this pattern, and encode 101 relations (e.g. *caused_by*, *based_in*, *written_by*).

member pattern: 1,298 member categories.

partly explicit and implicit relations categories: 98,855 categories of these two types are processed similarly to each other. If none of the rules to determine the nature of the relation encoded in the category name applies, the processing continues as if no relation indicators were available.

class attribute categories: 7,564 categories. Processing the category names reveals 840 classes with an average of 2.27 attributes. A sample is presented in Table 3.

[5]Wikipedia categories are usually in plural. Before extracting the pages we transform Y to its plural form.

[6]We work with the English Wikipedia dump from 2007/08/02.

[7]http://www-nlp.stanford.edu/software/

Class	Attributes
ART	country, media, nationality, origin, period, region, type
BOOK	author, award, country, head of state or government, ideology, nationality, publisher, series, subject, university, writer, year
BUILDING	architect, area, city, community, county, country, function, grade, locality, province, region, shape, state, territory, town
MUSICIAN	band, community, ethnicity, genre, instrument, language, nationality, region, religion, state, territory
WORK	artist, author, genre, head of state or government, nationality, writer, year
WRITER	area, award, ethnicity, format, genre, language, movement, nationality, period, religion, state, territory

Table 3: Classes and attributes extracted from Wikipedia's "by" categories.

Table 2 shows the number of unique extracted relations and evaluation results. *isa* , *spatial* and *member_of* relations were evaluated against ResearchCyc. We report the precision P[8], and in parentheses the number of concept pairs for that particular relation that also appear in ResearchCyc. From the false positive instances we randomly select 250 for manual annotations. For relations extracted from X [VBN IN] Y and "member" categories we also randomly select 250 for manual annotation (because the overlap with ResearchCyc for *member_of* is only 25 instances). Each relation subset is independently annotated by 2 judges. We report two annotation scores – one that corresponds to the intersection $\cap$ (instances that the annotators agree are correct) and one to the union $\cup$ (instances that at least one annotator marks as correctly assigned).

Analysis and Discussion

Apart from the fact that it is easier to analyze a short phrase to extract a semantic relation rather than a sentence or even document, analyzing category names and the category and page network for knowledge acquisition has other advan-

[8]$P_{\mathcal{R}} = \frac{TP}{TP+FP}$

TP (true positives) is the number of instances that were tagged with relation $\mathcal{R}$ by both our method and ResearchCyc, FP (false positives) is the number of instances that were tagged with $\mathcal{R}$ by our method but not by ResearchCyc.

tages as well. The category names express very concisely a relation which may also appear in the article, but expressed there in a more complex manner. We took the 42,711 *member_of* relations discovered through category name analysis, and extracted from the Wikipedia article corpus the sentences in which the two elements of the pair appear together – 131,691 sentences. Of these, only 1985 sentences contained the word member. By determining accurately through category name analysis the semantic relation involved, we can contribute to paraphrase analysis: the joining phrases/terms from the corpus can be considered paraphrases expressing the discovered relation.

The high manual evaluation scores confirm our hypothesis that Wikipedia category names and structure are a rich and accurate source of knowledge. The low evaluation score with ResearchCyc has two causes: (i) one or both of the concepts do not appear with the intended sense in this resource, or (ii) the relation we look for is not contained in it. While useful for evaluating *isa* relations (Ponzetto & Strube, 2007), evaluating more diverse relation types using ResearchCyc is not meaningful anymore and should be replaced by evaluations within applications.

One of the advantages of the method presented is the fact that we can find numerous instances of a relation between two concepts using the category network and the category-page links. This can be also a problem if the relation detected was incorrect. This was especially true for "by" categories, which seemed to be a good source of *isa* relations as well. When for a X by Y category we added *isa* links between all the pages it covers and X, this did not lead to good results. The reason are categories such as ROMANESQUE ARCHITECTURE BY COUNTRY, under which are listed cities (e.g. page ANDERLECHT under subcategory ROMANESQUE SITES IN BELGIUM) or particular buildings (e.g. page MAINZ CATHEDRAL under subcategory ROMANESQUE SITES IN GERMANY).

To determine the relation between two concepts C_i and C_j when no indicators (such as prepositions) are present is similar to the problem of determining the semantic relation between nominals (see Girju et al. (2007) for an overview of research in this area). The difference between this and other approaches is that we do not have a list of relations with which to annotate the data, nor do we use labeled examples to learn. We plan to adopt a method closer to relation extraction (Yates & Etzioni, 2007), based on joining terms[9] found in the corpus (Turney & Littman, 2005).

Related Work

Approaches to knowledge acquisition can be grouped based on the input used: large, unstructured corpora; semi-structured data; human users.

Mining for knowledge in general texts is quite popular, and the redundancy can be used to filter some of the noise. TextRunner (Banko et al., 2007) first learns from a small corpus sample a model for classifying relations of interest, then extracts candidates from a larger corpus which are judged relevant or not relevant using the learned model. The goodness of the extracted relations is decided based on the support found in the corpus. From a 9 million Web page corpus, TextRunner extracts 11.3 million tuples, of which 1 million concrete tuples with arguments corresponding to real-world entities estimated to be correct in proportion of 88.1%, 6.8 million "abstract" tuples, with a correctness estimate of 79.2%. Zhao & Grishman (2005) focus on detecting specific relations (such as *Located-In* between entities – person, organization, facility – referred to in the text), by using a kernel method which combines lexical, grammatical and contextual data. In this supervised learning approach, the targets are the relations specified in the ACE corpus. 5-fold cross validations over the 4,400 relations in this corpus give a highest of 70.35% F-score. Davidov et al. (2007) extract both concepts and relations in an incremental approach. Processing starts with a small seed for a concept, which is expanded and used to extract contextual information, to generate a concept class and binary relations involving this class through iterative clustering. The method is evaluated by building three concept classes and their corresponding relations, with a precision ranging from 0.68 to 0.98 and recall from 0.51 to 0.90. The seed approach is also used for detecting relations of a specific type. Manually designed patterns have been used to find *isa* (Hearst, 1992) or meronymic relations (Berland & Charniak, 1999).

Current approaches to knowledge acquisition from human users are different from the efforts of the past when the burden of building the resource was placed on the shoulders of a small number of experts: the task is now distributed to numerous volunteers through collaborative projects supported by the Web – Cyc (Lenat & Guha, 1990), OpenMind Common Sense (Singh, 2002), Verbosity (von Ahn, 2006).

A way to avoid noise from unrestricted text is to exploit semi-structured data. Kylin (Wu & Weld, 2007) and the system presented by Nguyen et al. (2007) use Wikipedia infoboxes – snippets of structured information about a company, person, region, and so on – as training data, and learn how to fill in similar templates for pages that do not have such information. They process the page content and combine it with the filled in templates to learn how to find such information in Wikipedia articles. For four concepts, Wu & Weld (2007) obtain precision between 73.9% and 97.3%, and recall between 60.5% and 95.9%. Nguyen et al. (2007) filter article sentences, parse and analyze them for entity detection and keyword extraction. These elements are used to learn how to detect instances of previously seen relations, with 37.76% f-score. Yago (Suchanek et al., 2007) and DBpedia (Auer et al., 2007) extract information specifically from Wikipedia. Yago combines WordNet's hierarchy with Wikipedia pages to obtain a larger and more interconnected semantic network. Facts representing relations of 14 types are extracted from Wikipedia and are used to induce more links in this network. Accuracy is estimated based on a small sample of manually annotated relation instances out of the approximately 5 million extracted, and falls between 90.84$\pm$ 4.28% and 98.72 $\pm$ 1.30%. Yago also identifies specific categories that provide relational information, such as 1975 BIRTHS, categories starting with Countries in ..., Rivers

[9] A *joining term* for pair (X,Y) is a word sequence $WSeq$ that appears between X and Y in a corpus (X $WSeq$ Y).

of ..., Attractions in ...,, and exploit them as a source of the following relations: *bornInYear, diedInYear, establishedIn, locatedIn, writtenInYear, politicianOf, hasWonPrize*. In DBpedia the goal is to convert Wikipedia content into structured knowledge using information from Wikipedia's relational database tables, and the structured information in infoboxes. The information extracted – approximately 103 million RDF triples – is assumed to be accurate. In addition to this, they link the resulting database to other data sources on the Web, such as Geonames, MusicBrainz, US Census, WordNet, Cyc. Ponzetto & Strube (2007) build on the category network from Wikipedia and induce *isa* links based on several criteria: head matching, modifier matching, structural analysis of shared categories and pages between two linked categories, and patterns indicative of *isa* relations and *notisa* relations. The result are 105,418 *isa* relations, evaluated at 87.9% F-score compared with ResearchCyc (on the 85% pairs that overlap). Paşca (2007) processes search engine queries to obtain class attributes. The idea is that when writing a query, users have some elements of a relation on which they require further information – such as *side effects* for class *drugs*, or *wing span* for class *aircraft model*. From extensive logs of even noisy queries, a weakly supervised system can acquire large sets of relevant class attributes. Similarity between automatically ranked class attributes and manually assigned correctness label on a sample of extracted attributes for the 40 classes considered range between 90% precision for 10 attributes to 76% for 50.

Conclusions

We have explored category names and category structure in Wikipedia as sources of relations between concepts. The analysis and experiments performed show a wealth of information that can be induced from these elements: instances of relations, relation types and class attributes. We will refine this work by testing other methods for determining the semantic relation between concept pairs, and expand the category name analysis to even finer category name constituents.

In both statistical and semantic analysis tasks it is useful to be able to generalize a concept – to address the data sparseness issue, or to be able to cluster similar entities. Even when a taxonomy is available, finding the most appropriate level of generalization is not easy. We plan to explore in future work people's preferences for generalizations, as captured in the "by" categories.

This research has started from the observation that Wikipedia categories have complex names, which encode some form of human knowledge of organization and classification. Splitting category names into smaller strings, we retrieve concepts that are of interest in language processing, and salient relations between them. Our goal is to transform Wikipedia's category network into a network of concepts linked by a variety of semantic relations, ready to provide knowledge to higher end NLP applications such as coreference resolution, summarization and question answering.

Resource. The triples extracted with this method is available on our web page (http://www.eml-research.de/nlp/download/wikirelations.php).

Acknowledgements. We thank Simone Paolo Ponzetto for sharing and explaining his system for building the Wikipedia category network and Wikipedia taxonomy. We thank the Klaus Tschira Foundation for financial support.

References

Auer, S., C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak & Z. Ives (2007). DBpedia: A nucleus for a Web of open data. In *Proc. of ISWC 2007 + ASWC 2007*, pp. 722–735.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead & O. Etzioni (2007). Open information extraction from the Web. In *Proc. of IJCAI-07*, pp. 2670–2676.

Berland, M. & E. Charniak (1999). Finding parts in very large corpora. In *Proc. of ACL-99*, pp. 57–64.

Davidov, D., A. Rappoport & M. Koppel (2007). Fully unsupervised discovery of concept-specific relationships by Web mining. In *Proc. of ACL-07*, pp. 232–239.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database.* Cambridge, Mass.: MIT Press.

Girju, R., P. Nakov, V. Nastase, S. Szpakowicz, P. Turney & D. Yuret (2007). SemEval-2007 Task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1),* Prague, Czech Republic, 23–24 June 2007, pp. 13–18.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING-92*, pp. 539–545.

Kilgarriff, A., P. Rychly, P. Smrz & D. Tugwell (2004). The Sketch Engine. In *Proceedings of the 11th International Congress of the European Association for Lexicography,* Lorient, France, 6–10 July 2004, pp. 105–116.

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds.*, (Ph.D. thesis). Macquarie University, Sydney, Australia.

Lenat, D. B. & R. V. Guha (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project.* Reading, Mass.: Addison-Wesley.

Nguyen, D. P., Y. Matsuo & M. Ishizuka (2007). Relation extraction from Wikipedia using subtree mining. In *Proc. of AAAI-07*, pp. 1414–1420.

Paşca, M. (2007). Organizing and searching the World Wide Web of facts – Step two: Harnessing the wisdom of the crowds. In *Proc. of WWW-07*, pp. 101–110.

Ponzetto, S. P. & M. Strube (2007). Deriving a large scale taxonomy from Wikipedia. In *Proc. of AAAI-07*, pp. 1440–1445.

Santorini, B. (1990). *Part of Speech Tagging Guidelines for the Penn Treebank Project.* http://www.cis.upenn.edu/∼treebank.

Schubert, L. K. (2006). Turing's dream and the knowledge challenge. In *Proc. of AAAI-06*, pp. 1534–1538.

Singh, P. (2002). *The Open Mind Common Sense Project.* http://www.kurzweilai.net/articles/art0371.html.

Suchanek, F. M., G. Kasneci & G. Weikum (2007). YAGO: A core of semantic knowledge. In *Proc. of WWW-07*, pp. 697–706.

Turney, P. D. & M. L. Littman (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278.

von Ahn, L. (2006). Games with a purpose. *IEEE Computer*, 6(39):92–94.

Wu, F. & D. S. Weld (2007). Autonomously semantifying Wikipedia. In *Proc. of CIKM-07*, pp. 41–50.

Yates, A. & O. Etzioni (2007). Unsupervised resolution of objects and relations on the Web. In *NAACL-HLT-07*, pp. 121–130.

Zhao, S. & R. Grishman (2005). Extracting relations with integrated information using kernel methods. In *Proc. of ACL-05*, pp. 419–426.