# WikiNet: A Very Large Scale Multi-Lingual Concept Network

**Vivi Nastase**[1], **Michael Strube**[1], **Benjamin Börschinger**[1,2], **Cäcilia Zirn**[1,2], **Anas Elghafari**[1,3]

[1] HITS gGmbH, Heidelberg, Germany
[2] Department of Computational Linguistics, University of Heidelberg, Heidelberg, Germany
[3] Department of Linguistics, University of Tübingen, Tübingen, Germany

**Abstract**

This paper describes a multi-lingual concept network obtained automatically by mining for concepts and relations and exploiting a variety of sources of knowledge from Wikipedia. Concepts and their lexicalizations are extracted from Wikipedia pages. Relations are extracted from the category and page network, infoboxes and the body of the articles. The network consists of a central, language independent list of concepts (keeping track of their lexicalizations in various languages), interconnected with a variety of relations to form a very large scale multi-lingual concept network.

## 1. Introduction

Machine readable knowledge is crucial for realistic, robust applications such as translation, question answering, summarization. Acquiring such knowledge has, until recently, been done manually. Early expert systems focused on small and simplified domains, in the hope that the knowledge from such constrained worlds can be easily captured. Two projects were more ambitious: Cyc aimed at encoding (all) common sense knowledge (Lenat et al., 1990); WordNet (Fellbaum, 1998) organized the words in the English language according to their meanings, and using a selection of relations – such as synonymy, hypernymy/hyponymy, meronymy/holonymy, antonymy, derivational. WordNet was extremely successful, in that it was widely adopted to support numerous language processing tasks. The Open Mind Common Sense (Singh, 2002) and MindPixel projects had the ground-breaking idea of using distributed human computation – in the form of collaborative endeavours – to gather large repositories of common sense and general knowledge from willing contributors over the web. Splitting the enormous task of gathering knowledge over a large base of contributors had the potential to break the knowledge acquisition bottleneck. These approaches were not widely adopted, and ended short of their stated goals. Acquiring knowledge from unrestricted texts is currently a very active research area. The problem is that while millions of facts can be fast extracted from texts, there are issues related to noise, and – probably most importantly – the large collections built will consist of isolated/loosely-connected facts.

The answer to our knowledge acquisition desires is considered nowadays Wikipedia. Shortly after its launch in January 2001, it has become apparent to researchers in our field that it has huge potential as a vast, multi-lingual source of knowledge. The collaborative nature of the Wikipedia project ensures a wide base of contributors. Its simple editing process and guidelines have led to its steady growth and semi-structured, high quality content.

In this paper we introduce a very large scale, multi-lingual concept network, obtained by exploiting several facets of Wikipedia. The resource consists of a language-independent concept base extracted from Wikipedia articles and categories, and relations between them. The articles and categories stand for concepts (as commonly done, see (Medelyan et al., 2009)). The cross-language links give lexicalization options in various languages. Relations between concepts are extracted from the article bodies, infoboxes, categories and the category network. This organization of the resource – an index of concepts and their lexicalizations plus a large repository of relations – mirrors WordNet, which has been intensely used for almost two decades, and with which the NLP community is very much accustomed. We intend WikiNet to supplement WordNet, and provide from the start a multi-lingual resource, with millions of named entities (which were outside the scope of WordNet) and numerous relations.

An advantage of such a resource is that it can be used for multiple languages even when it is obtained from only one language version (English), through the multi-lingual index of concepts. The algorithms to build this resource can be applied to the latest Wikipedia versions, and thus obtain an up-to-date resource.

## 2. Related Work

The most obvious part of Wikipedia consists of the articles. The majority of articles contain text structured in some way (into sections and paragraphs) and include hyperlinks to other articles. Auer et al. (2007), Suchanek et al. (2007), Kasneci et al. (2008) have exploited these links and extracted large knowledge bases. Many articles contain infoboxes, and they have also been (separately) used to provide simple facts, and also as a basis for learning how to identify and retrieve the type of relations captured from unstructured text (Wu and Weld, 2007; Nguyen et al., 2007). Articles also contain cross-language links. Wentland et al. (2008) used these to compile a large database of multi-lingual named entities. Adafre and de Rijke (2006) used them to obtain parallel corpora and to bootstrap dictionaries for languages with few electronic resources.

The next facet of Wikipedia that comes to mind are the categories: each article is assigned to categories which the contributors consider relevant. Gabrilovich and Markovitch (2007) have shown how useful the article-category links are

to compute semantic relatedness of any words mentioned within articles.

Categories, however, are themselves interconnected. Strube and Ponzetto (2006) have noticed the similarity between Wikipedia's category network and WordNet's network, and have mapped word similarity/relatedness computation methods from WordNet to Wikipedia's category network. To compute similarity *isa* relations were needed, so they were gradually introduced (Ponzetto and Strube, 2007).

The current largest repositories of machine readable knowledge are YAGO and FreeBase. YAGO (Kasneci et al., 2008) also combines several Wikipedia aspects to build its knowledge base. Pages are linked to WordNet, and a number of categories explicitly providing relational information (e.g. *1975 births*, categories starting with *Countries in ..., Rivers of ...*) are processed to obtain relations (e.g. *bornInYear, establishedIn, locatedIn, politicianOf*). YAGO includes relations extracted from the infoboxes, and also links to DBpedia (a large scale repository of facts extracted from Wikipedia's underlying relational database), SUMO, and other resources.

Freebase[1] started with a large base of concepts and facts automatically extracted from Wikipedia, and merged them with other resources (MusicBrainz, the Notable Names Database). The knowledge base is freely available, and also editable, as Wikipedia is, by registered users. For a comprehensive overview of work processing Wikipedia see (Medelyan et al., 2009).

Apart from YAGO, other work concerned with extracting knowledge from Wikipedia has only considered part of its sources of knowledge – the articles separately, the categories separately, sometimes the articles with their categories. YAGO makes partial use of the categories (as a source for a number of relations), but is mostly focused on extracting specific relations.

The resource presented here exploits together several facets of Wikipedia: the articles – including hyperlinks, anchor texts, cross-language links, the infoboxes – the categories and the category network.

## 3. Building the Network

Similarly to WordNet, WikiNet consists of an index of concepts and relations between them. The index serves to separate the lexicalization of concepts from their relations. Within WikiNet, this separation allows us to have a multilingual index and a language independent relation network. The index covers the Wikipedia articles and categories. The lexicalizations of these concepts and the relations between them are extracted through various methods, detailed in the following subsections.

### 3.1. The index

Articles and categories in Wikipedia are customarily considered to correspond to concepts, and they constitute the backbone of the resource. These concepts can be lexicalized in various languages, or in various ways within one language. Cross-language links connect articles on the same topic/concept in various languages, and they provide the multi-lingual expressions of the same concept. Anchor texts – the "names" of the hyperlinks that connect to specific articles – provide lexicalization variants within a language. For example, the TV series *Seinfeld* is referred to as *Seinfeld*, *The show about nothing*, and even misspelled variations (e.g. *Sienfeld*). Exploiting these sources of information will result in the multi-lingual index.

The multi-language concept index is straightforward to obtain – cross language links are included for each Wikipedia article and partly for categories. There are lapses in this regularity. Wentland et al. (2008) address this problem through triangulation (linking two concepts in different languages through overlapping cross-language links). In our current endeavour this should not pose a problem – once a language is processed (in particular English, which has the most content), all existing lexicalizations are included. When processing the Wikipedia content in a new language $L$, if we encounter a concept that does not appear in the previously created index, its own cross-language links can help determine to which existing concept this lexicalization refers to, or indicate that a new entry is necessary. This is similar to the triangulation step from (Wentland et al., 2008).

The concept index includes both articles and categories. The inclusion of the category network ensures the connectivity of the resulting resource. Moreover, these connections are qualitative – users have added them consciously to reflect their structural intuitions, and Ponzetto and Strube (2007) have shown that they lead to relatedness and similarity scores that correlate highly with human judgements. Categories in various languages differ. The resource presented here includes the English category network. While we also process the German Wikipedia dump, we have not yet evaluated the impact of combining category systems from different languages, and as such they are not included. The index consists of a list of integer IDs representing concepts, and their lexicalizations. An article and its homonymous supercategory share an ID. The lexicalizations are collected from the article name, the cross-language links, anchor texts and disambiguation links.

### 3.2. The Relations

Relations connect the concepts in the extracted index. They are obtained from several sources, detailed below.

#### 3.2.1. The category network

Categories in Wikipedia are added by users to structure the content. Multiple adjustments from numerous contributors have led this structure to converge to something that reflects some of our own conceptual preferences. Strube and Ponzetto (2006) and Ponzetto and Strube (2007) have shown that using this structure one can compute concept relatedness and similarity measures that correlate highly with human judgements.

Categories come in different varieties: some do represent concepts (e.g. ROME – added as a category because there is much to say about the city of Rome, and all these aspects can conveniently be linked to this category); some

[1] www.freebase.com

| 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]$^2$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

serve a purely organizational purpose: e.g. NOVELS BY PHILIP K. DICK, NOVELS BY AUTHOR which cluster together pages describing concepts with specific properties. Such categories can be deconstructed, to recover the knowledge encoded in them (Nastase and Strube, 2008). In brief, five types of categories can be deconstructed to obtain various relations, as presented in Table 1.

These relations are induced following the processing steps, for each category $Cat$ in the English version of Wikipedia[3]:

1. determine the constituents $C_i$ of the phrase $Cat$ from a syntactic parse of $Cat$ – they will correspond to the (non-overlapping) noun phrases in the parse;
2. determine the dominant constituent $C_D$ (it is the one that has the same syntactic head as $Cat$).
3. form pairs ($\mathbf{C}_i$, $\mathbf{C}_D$) for all constituents $C_i$ of $Cat$ ($C_i \neq C_D$ and determine the relation $\mathbf{C}_i\ \mathcal{R}\ \mathbf{C}_D$ based on matching $Cat$ to the patterns identified and shown in Table 1;
4. for each page $P_j$ subsumed by $Cat$, and all pairs ($\mathbf{C}_i$, $\mathbf{C}_D$):
   - add relations $\mathbf{P}_j\ \mathcal{R}\ \mathbf{C}_i$.
   - add relations $\mathbf{P}_j$ *isa* $\mathbf{C}_D$;

Propagating the relation $\mathcal{R}$ from the category constituents to the pages follows the rule:
*if* $\mathbf{P}_j$ *isa* $\mathbf{C}_i$ *and* $\mathbf{C}_i\ \mathcal{R}\ \mathbf{C}_D \Longrightarrow \mathbf{P}_j\ \mathcal{R}\ \mathbf{C}_D$,
in a way similar to propagating an explicit relation found within a category name, as illustrated in Table 1.
Finding the relation between one pair, ($\mathbf{C}_i$, $\mathbf{C}_D$) means automatically finding the relation between numerous ($\mathbf{P}_j$,$\mathbf{C}_D$) and ($\mathbf{P}_j$,$\mathbf{C}_i$) pairs.

[2]"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.

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

One processing problem that needed to be addressed at this step is finding the concepts corresponding to category constituents, such that the relations induced can be mapped onto relations between concepts. To solve this we use the category network, and choose the closest category or article for a constituent derived from a deconstructed category.

#### 3.2.2. Infobox relations

Infoboxes are another source of user structured knowledge. Based on the network built until this point – consisting of the existing category and article network, enhanced with the relations discovered in the previous step – we propagate infobox relations, based on the following observation: relations in infoboxes are often important enough and shared by enough entities that Wikipedia contributors use them for categorization. For example: the place of origin for artifacts is often included in the article's infobox, if it has one, and is also used in categorization. *Katyusha*, a weapon, has as place of origin *Soviet Union* (in its infobox), and is also categorized under *Military equipment of the Soviet Union*, together with several other entities, not all of which have an infobox. The category was used in the previous step (deconstructing categories) to determine that there is a relation between each article under this category and the concepts *Military equipment* and *Soviet Union*. The existing infoboxes help determine that the relation between the articles and *Soviet Union* is place-of-origin, as illustrated in Figure 1. This step of propagating information from infoboxes using the categories will determine (name) some of the anonymous or very general relations established in the previous step.

To establish the link between category names and relations in infoboxes and spread this information through the network, we apply the following steps for each category CAT:

**1. Gather the information about CAT**

- establish the constituents of CAT (as obtained during the category deconstruction process). To simplify the argument let us assume that CAT has two constituents, $C_X$ and $C_Y$, and $C_X$ is the dominant constituent.

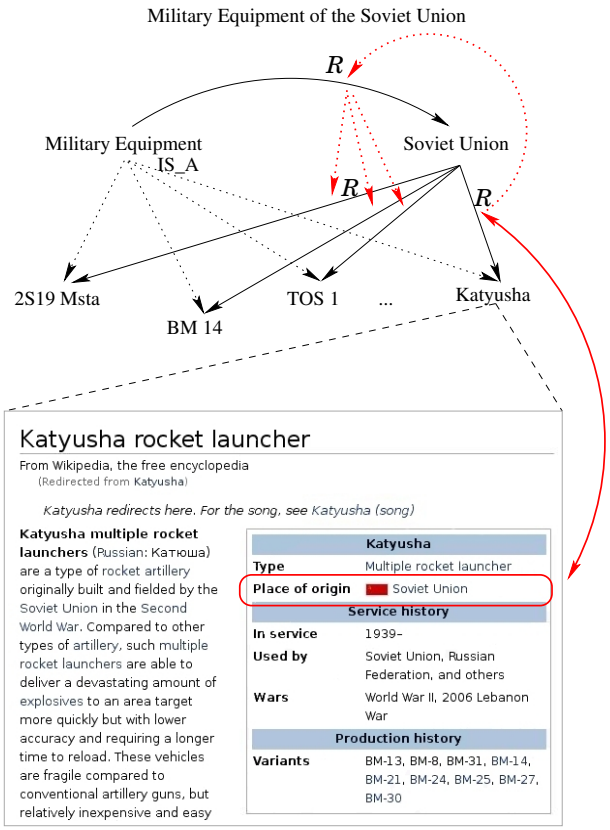


Figure 1: Some articles under a Wikipedia category contain infoboxes with hand-picked relations. From these we can determine the relation that holds between the corresponding concept and the concepts derived from its parent category, and then propagate these relations in the network.

$constituents(\text{CAT}, C_x)$.
$constituents(\text{CAT}, C_Y)$.

- gather the pages $P_i$ subsumed by CAT:
  $P_{Cat} = \{P_i | subsumes(\text{CAT}, P_i)\}$

**2. Extract the relations from the pages that have infoboxes** From the pages $P_i$ subsumed by CAT that contain an infobox we extract the tuples:
$infoboxRel(P_i, R_j, V_j)$.
where $R_j$ is an attribute with value $V_j$ in $P_i$'s infobox.

**3. Extract candidates for the relation between $C_X$ and $C_Y$** We say that a relation $R_j$ from the infobox for a page $P$ is a candidate relation for a category CAT with constituents ($C_X$,$C_Y$), if it is associated with the same value $V$ in all infoboxes in which it appears under $C$, and $V$ is compatible with $C_Y$. Two values are compatible if they are identical, paraphrases of each other, or are connected in a systematic way – $V$ is a specialized concept of $C_Y$, or for locations for example $V$ *part-of* $C_Y$ (a specific location in Europe is compatible with **EUROPE**). This predicate can become more specialized as more relations are added to the fact base. Formally:

$candidate(R_j, \text{CAT}) \leftarrow$
$constituent(\text{CAT}, C_Y)$
$subsumes(\text{CAT}, P)$
$infoboxRel(P, R_j, V)$
$\wedge\ (\nexists V_k,$
$subsumes(\text{CAT}, P_i)$
$\wedge\ infoboxRel(P_i, R_j, V_k)$
$\wedge\ V \neq V_k)$
$\wedge\ is\_compatible(V, C_Y))$

To simplify the process we use $V = C_Y$.
The set of candidate relations is:
$R_{Cat} = \{R_j | candidate(R_j, \text{CAT})\}$

**4. Choose the relation and propagate it in the network**
If there is only one candidate relation ($|R_{Cat}| = 1$), we propose that this relation (let us call it $R_0$) replaces the (possibly underspecified) relation $\mathcal{R}$ determined in the category deconstruction process to hold between $C_X$ and $C_Y$. The triples ($\mathbf{P}_i\ \mathcal{R}\ \mathbf{C}_Y$) added during the category deconstruction phase will be replaced by ($\mathbf{P}_i\ R_0\ \mathbf{C}_Y$).
If there is more than one relation, but they are compatible (e.g. *birth place*, *place of birth*), we choose one to be $R_0$, and replace all extracted relations just as above. If the relations extracted are not compatible (or we cannot establish that they are compatible), no further processing takes place, and the originally proposed $\mathcal{R}$ remains. At this point compatibility is established through lexical overlap.
The higher up in the hierarchy, the more pages a category subsumes, and the more heterogeneous the category is. Because of this it is easy to introduce noise, by inducing relations that apply only to a small subset of the subsumed pages. Also, processing a heterogeneous category may prevent the method from finding one relation that applies to all pages, as there may be several possible. For example, the category PEOPLE FROM CHICAGO, ILLINOIS subsumes pages **PHILIP K. DICK**, whose relation with **CHICAGO, ILLINOIS** is *born-in*, and **SAUL BELLOW**, who *lived-in* **CHICAGO, ILLINOIS** but was not born there. To address this situation, for a category CAT we seek relation candidates and propagate them within smaller and more homogeneous sets of pages, as given by specialized subcategories of CAT, following the same processing steps. The process of propagation is illustrated in Figure 1.
At this point, not all relations from the infobox are included in the resource. The reason is that numerous values of the attributes/relations included in the infobox are not concepts themselves (e.g. surface area and population for a country). Such relations will be included in future versions of the resource, possibly as separate files.

### 3.2.3. Relations from the article bodies

One of the most obvious sources of additional information from the article bodies are the hyperlinks. They (seem to) highlight concepts that are relevant or related to the concept being described. Milne and Witten (2008a) extracted these concept relations and have shown they can be successfully used for computing semantic relatedness.
A closer inspection shows that this is not necessarily the case, or rather, relations tend to be more distant. For example, the following sentence is from the article on *Chocolate*:

*Cacao pods are harvested by cutting the pods*

*from the tree using a machete, or by knocking them off the tree using a stick.*

While *machete* is relevant to *Cacao pods*, as they are used in harvesting, they are not directly relevant to *chocolate*. We aim for relations that reflect more direct connections between concepts – that indicate direct relevance of the two concepts to each other. Because of this we extract relations from the article bodies by identifying pairs of concepts that appear together in a sentence. The concepts considered are the one corresponding to the article that is being processed, and those corresponding to the hyperlinks in this article. All occurrences of these concepts in the article text will be identified (not all occurrences of a term anchor a hyperlink to the corresponding article), and we extract as relations the pairs of concepts that co-occur in a sentence. Currently the resource contains (in a separate file) approximately 163 million such cooccurrences, not filtered based on frequency.

## 4. Multi-linguality

The multi-lingual nature of the resource comes from several sources. Even when processing only one language version (English in our case, since it has the most content), the end result contains a multi-lingual index because of the cross-language links. The network can be accessed through concepts in any of the languages represented, and thus offers a medium for computing cross-language or "single language" semantic relatedness for a variety of languages. When processing the Wikipedia content in another language, the relations induced can be added to the existing resource. These new relations may introduce new ways of connecting and organizing the existing concepts, and have an impact on relatedness measures.

## 5. The Resource

The resource is obtained by processing the article dump for the desired language. The output is in an easy to understand text format. The script extends the extractWikipediaData.pl script distributed with the Wikipedia Miner system (Milne and Witten, 2008b). The data (and, in the near future, also the scripts) can be downloaded from our group's website[4]. Included is a simple script to obtain the relations for a user-specified concept. A toolkit for exploiting the resource – obtaining relations for a concept, computing relatedness and similarity and wikifying documents is under works. It will also be available as open source.

Table 2 shows the statistics of the currently posted resource. It was obtained from the 2009/07/13 version of the English Wikipedia.

The steps of the relation extraction process have been separately evaluated. Strube and Ponzetto (2006) and Ponzetto and Strube (2007) show the usefulness of the category network, as it can be used as a base for computing concept relatedness and similarity that correlate highly with human judgement.

[4] `http://www.h-its.org/english/research/nlp/download/wikinet.php`

| | | | |
|---|---|---|---|
| Concepts | 3,347,712 | articles | 2,857,497 |
| | | categories | 490,215 |
| Relations | 36,246,913 | *isa* | 10,063,364 |
| | | *isa* (for categ.) | 455,799 |
| | | *spatial* | 4,077,647 |
| | | *nationality* | 568,828 |
| | | *topic* | 337,564 |
| | | *genre* | 331,540 |
| Co-occurrences | 162,800,112 | | |

Table 2: Statistics on the resource generated, with some example of the relations extracted

Nastase and Strube (2008) evaluated the results of the category deconstruction process on the English Wikipedia version of 2007/08/02. We reproduce here the evaluation results obtained. The current resource was obtained running the same process on a more recent version of Wikipedia. Table 3 shows the number of unique extracted relations and evaluation results for the category deconstructing step. *isa*, *spatial* and *member_of* relations were evaluated against ResearchCyc. We report the precision $P$[5], 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).

The infobox relation propagation also leads to high quality relations, as shown through manual evaluation, and overlap with YAGO's fact base[6] with manual/automatic evaluation of the overlap[7]. We have established a connection between the category name (specifically, a constituent of the category name) and a value in the infobox of a subsumed page for 42,060 categories. 130,123 pages contain such an explicit connection, for a total of 175,350 $P_j, C_i$ (page-constituent) links. The information was propagated to a further 544,702 pages and their 698,929 relations to the corresponding category name constituent.

Manual evaluation was performed by two judges on two samples of 250 instances – one for high frequency relations, one for low frequency ones – which contain the same distribution of relations as the subset they repre-

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

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.

[6] `http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html`

[7] This evaluation is done on the data obtained using the 2009/07/13 English Wikipedia dump.

| 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 3: Extracted relations and evaluation results

sent. The guidelines instructed the annotators to assign a `true/false/not_relation` tag to each instance. The results in terms of precision are presented in Table 4, relative to `true` tags assigned by both judges ($\cap$) or at least by one judge ($\cup$). The agreement between judges in terms of Cohen's kappa is 0.62 for the high frequency sample, and 0.81 for the low frequency one.

| Sample | Instances | Evaluation | |
|---|---|---|---|
| | filtered/all | $\cap$ | $\cup$ |
| high freq | 235/250 | 78.3% / 73.6% | 86.8% / 81.6% |
| low freq | 235/250 | 75.7% / 71.2% | 77.9% / 73.2% |

Table 4: Manual annotation results and evaluation, on the sample with only valid relations/on the full sample

The overlap of the set of 698,929 relation instances with YAGO's fact base is 7,143 concept pairs. This small overlap shows that categories, the category structure and infoboxes are the combined source of novel information, not easily or directly accessible through the article texts or categories alone. 306 YAGO relations are represented within the 7,143 pairs. We consider the top 5, which cover 5854 of the pairs: *locatedIn* (3163), *wrote* (972), *directed* (757), *politicianOf* (572) and *created* (390).

To the pairs assigned *created* in YAGO correspond the following relations assigned through the method presented in the paper:

*artist* (126), *writer* (89), *developer* (58), *director* (37), *manufacturer* (14), *producer* (12), *composer* (8)[8].

It is clear that these relations assigned by propagating relations from the infoboxes are more specific instances than the *created* relation in YAGO. In the manually annotated sample we have 14 instances annotated with relations from this set. Their precision is 87.5% (both $\cap$ and $\cup$). The same phenomenon occurs for *locatedIn* – it is a rather general relation, and it corresponds to a variety of more specific spatial relations in our assignment: *subdivision_name* (1288), *prefecture* (660), *location* (257), *neighbouring_municipalities*, *district* (142), *country* (46), *basin_countries* (37), *bundesland* (30), *county* (18). Of these, the relations *subdivision_name*, *location*, *country* also appear in the manually annotated data (80 instances), and (together) have precision 86.25% ($\cup$) / 83.75% ($\cap$) (Table 5).

The *wrote* and *directed* YAGO relations are easily mapped onto the propagated relations: for 963 of the instances with relation *wrote* in YAGO, the inference process assigned the relation *author* (99.07%), and 749 instances of relation *directed* have relation *director* after propagation (98.94%). The relation *politicianOf* is harder to evaluate. None of the relations assigned through relation propagation expresses the same relation, however they are not erroneous: *birth_place* (350), *death_place* (93), *residence* (16), *nationality* (15). These relations were represented in the manually annotated data (40 instances), and their precision was 72.5% ($\cup$) / 70% ($\cap$).

| YAGO relation | Overlap | Precision |
|---|---|---|
| full evaluation | | |
| wrote | 972 | 99.07% |
| directed | 757 | 98.94% |
| estimation based on manually annotated sample | | |
| located in | 3,163 | 83.75% ($\cap$) / 86.25% ($\cup$) |
| created | 390 | 87.5% |

Table 5: Evaluation relative to the overlap with YAGO

## 6. Conclusion

We have presented a multi-lingual resource, to be used to complement WordNet with knowledge about numerous named entities as well as general concepts. It captures a wide variety of relations, reflecting the encyclopedic nature of the data. We build from the start a multi-lingual resource, to be used for cross-language tasks. To build it we exploit several sources of knowledge from Wikipedia – some explicit (articles, categories and their links, infoboxes), some implicit (category names). This coverage of multiple information differentiates the resource presented here from similar endeavours and resources extracted from Wikipedia.

## Acknowledgements

This work has been partially funded by the European Commission through the CoSyne project FP7-ICT-4-248531 and the Klaus Tschira Foundation.

[8]We show only the most frequent relations, which cover the majority of the pairs.

## 7. References


Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the EACL 2006 Workshop on New Text — Wikis and Blogs and Other Dynamic Text Sources,* Trento, Italy, 4 April 2006, pages 62–69.

Sören Auer, Christian Bizer, Jens Lehmann, Georgi Kobilarov, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a Web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference,* Busan, Korea, November 11-15, 2007, pages 722–735.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India, 6–12 January 2007, pages 1606–1611.

Gjergji Kasneci, Maya Ramnath, Fabian M. Suchanek, and Gerhard Weikum. 2008. The YAGO-NAGA approach to knowledge discovery. *SIGMOD Record*, 37(4):41–47.

Douglas B. Lenat, R.V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd. 1990. Cyc: Towards programs with common sense. *Communications of the ACM*, 33(8):30–49.

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Interaction*, 67(9):716–754.

David Milne and Ian H. Witten. 2008a. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08,* Chicago, Ill., 13 July 2008, pages 25–30.

David Milne and Ian H. Witten. 2008b. Learning to link with Wikipedia. In *Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM 2008),* Napa Valley, Cal., USA, 26–30 October 2008, pages 1046–1055.

Vivi Nastase and Michael Strube. 2008. Decoding Wikipedia category names for knowledge acquisition. In *Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence,* Chicago, Ill., 13–17 July 2008, pages 1219–1224.

Dat P.T. Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Relation extraction from Wikipedia using subtree mining. In *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence,* Vancouver, B.C., Canada, 22–26 July 2007, pages 1414–1420.

Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence,* Vancouver, B.C., Canada, 22–26 July 2007, pages 1440–1445.

Beatrice Santorini. 1990. Part of speech tagging guidelines for the Penn Treebank Project. http://www.cis.upenn.edu/∼treebank/home.html.

Push Singh. 2002. The Open Mind Common Sense project. http://www.kurzweilai.net/articles/art0371.html.

Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July 2006, pages 1419–1424.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge. unifying WordNet and Wikipedia. In *Proceedings of the 16th World Wide Web Conference,* Banff, Canada, 8–12 May, 2007, pages 697–706.

Wolodja Wentland, Johannes Knopp, Carina Silberer, and Matthias Hartung. 2008. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In *Proceedings of the 6th International Conference on Language Resources and Evaluation,* Marrakech, Morocco, 26 May – 1 June 2008.

Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying Wikipedia. In *Proceedings of the ACM 16th Conference on Information and Knowledge Management (CIKM 2007),* Lisbon, Portugal, 6–9 November 2007, pages 41–50.


## Appendix

The resource consists of four data files which we describe below.

“a day of nights” 13257772
“a day of renew (album)” 9947241
“a day to remember (1953 film)” 19680415
“a day to remember (1991 film)” 19498936
“a day to remember albums” 19473486
“a day with my son” 12253172
“a day with wilbur robinson (2006 film)” 2017208
“a day without art” 20386856
“a day” 3153583
“a day’s adventure” 13434748
“a day’s pleasure” 2553430
“a day’s reign, or the false stanislas” 1941557
“a day’s wait” 2415580
...

Figure 2: **The index file** contains an alphabetical listing of lexicalizations and their corresponding unique numeric ID. Each line consists of a term and its associated numeric ID. One term may have several possible ID – showing polysemy.

12 “gd”:“Ain-Riaghailteachd” “en”:“Anarchism” “fr”:“Anarchisme” “it”:“Anarchismo” ...
25 “en”:“Autism” “et”:“Autism” “ca”:“Autisme” “fi”:“Autismi” “es”:“Autismo” ...
39 “lt”:“Albedas” “en”:“Albedo” “ast”:“Albedu” “hu”:“Albed” “et”:“Albeedo” ...
290 “lb”:“A (Buschtaf)” “uz”:“A (harf)” “ku”:“A (herf)” “fr”:“A (lettre)” ...
303 “lb”:“Alabama (Bundesstaat)” “br”:“Alabama (stad)” “ro”:“Alabama (stat SUA)” ...
305 “lt”:“Achilas” “fr”:“Achille” “en”:“Achilles” “scn”:“Achilli” “sl”:“Ahil” ...
307 “en”:“Abraham Lincoln” “lv”:“Abrahams Linkolns” “la”:“Abrahamus Lincoln” ...
308 “ga”:“Arastotail” “uz”:“Arastu” “kab”:“Aristot” “fr”:“Aristote” ...
309 “pl”:“Amerykanin w Paryu (Gershwin)” “nl”:“An American in Paris (Gershwin)” ...
316 “en”:“Academy Award for Best Art Direction” “es”:“Anexo:scar a la mejor direccin de arte” ...
324 “en”:“Academy Award” “id”:“Academy Awards” “tr”:“Akademi dlleri” ...
...

Figure 3: **The reversed index file** contains an ordered listing of numeric IDs and their various lexicalizations. A line starts with a numeric concept ID, and its possible lexicalizations, including variants in other languages, as found in the cross-language links for articles and categories.

12 -FIELD_OF_STUDY 1072099 1324482 148725 153803 1749719 ...
25 CATEGORY 1267652 15335930 2687547 RELATED_TO 640668 ...
39 CATEGORY 1487579 5233412 IS_A 716907 7427968 RELATED_TO ...
290 -ARTIST 11487620 6309589 ALPHABET 17730 CATEGORY 1476950 ...
303 -EXECUTED_BY 16971198 1832115 1936741 2063265 2191317 ...
305 ASSOCIATED_WITH 691877 ASSOCIATION 33158 CATEGORY 11262809 ...
307 ASSOCIATED_WITH 691877 ASSOCIATION 33158 CATEGORY 1044730 ...
308 ASSOCIATION 24526 CAPITAL 1216 CATEGORY 1013656 10557882 ...
309 AUTHOR 13066 CATEGORY 15590942 1901286 COMPOSER 13066 COUNTRY ...
316 BASED_IN 1732034 692361 CATEGORY 6001393 773951 COUNTRY 3173217 ...
324 BASED_IN 1732034 692361 CATEGORY 14390148 14952319 773951 ...
330 CATEGORY 1052809 13583794 17176975 22912361 6014666 742885 CITY ...
332 AUTHOR 2511084 CATEGORY 13726390 14001347 4250978 7025604 GENRE ...
...

Figure 4: **The data file** contains a list of relations for each concept (ID), ordered by the ID. Relations are directed. The file contains the relations induced by processing the category structure, categories and infoboxes. The structure of each line is: *ID* $Rel_1$ $ID_{11}$ $ID_{12}$ ... $ID_{1n}$ $Rel_2$ $ID_{21}$ .... ID is a concept's ID, $\text{Rel}_i$ are relations from or towards (if the relation is prefixed by "-") this concept, and $\text{ID}_{ij}$ are the concepts connected to ID through relation $\text{Rel}_i$.

12 10001591 100052 10030 1003654 10060195 10072892 1007418 100758 1008 ...
25 10001591 10008586 10013 10037201 1004186 1004482 10048 10055 1005705 ...
39 1000165 10086584 10112744 10180397 1019817 1028264 1028265 103050 ...
290 100935 10306453 103358 10436364 10437467 10576525 1109441 11388236 ...
303 10001591 10003335 10003629 10003649 10006052 1000637 10006781 ...
305 100254 10069798 1009303 10095749 10140510 10141 101411 10150963 ...
308 1000660 100090 1000978 10010856 1001664 100224 10023307 10024702 ...
330 10145549 13169236 14851243 16765178 169568 18948337 1942277 21148681 ...
332 1176603 1210571 12833051 13393902 13591897 14149608 15092767 15550841 ...
333 1046699 2477285 4104030
334 1137594 11807783 1181 1209 1234 12993 1317 1327 1328262 13909226 ...
...

Figure 5: **The cooccurrence relations file** contains an ordered listing of numeric IDs and the concepts they co-occur with within a sentence (in some article). This file contains cooccurrence relations mined from article texts. Each line in the file has the structure: *ID* $ID_1$ ... $ID_n$ where (ID, $\text{ID}_i$) appear together in a sentence in an article.