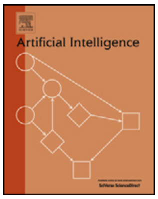
Artificial Intelligence ••• (••••) •••–•••

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Transforming Wikipedia into a large scale multilingual concept network

Vivi Nastase *, Michael Strube

HITS gGmbH, Heidelberg, Germany

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Knowledge base
Multilinguality
Knowledge acquisition

ABSTRACT

A knowledge base for real-world language processing applications should consist of a large base of facts and reasoning mechanisms that combine them to induce novel and more complex information. This paper describes an approach to deriving such a large scale and multilingual resource by exploiting several facets of the on-line encyclopedia Wikipedia. We show how we can build upon Wikipedia's existing network of categories and articles to automatically discover new relations and their instances. Working on top of this network allows for added information to influence the network and be propagated throughout it using inference mechanisms that connect different pieces of existing knowledge. We then exploit this gained information to discover new relations that refine some of those found in the previous step. The result is a network containing approximately 3.7 million concepts with lexicalizations in numerous languages and 49+ million relation instances. Intrinsic and extrinsic evaluations show that this is a high quality resource and beneficial to various NLP tasks.

1. Introduction

While the availability of large amounts of data has encouraged the development of successful statistical techniques for numerous natural language processing tasks, there is a concurrent quest for computer accessible knowledge. Knowledge allows a system to counter data sparsity (e.g. lexical semantic knowledge), as well as make connections between entities (e.g. **Barack Obama** *president_of* **United States of America**).

Shortly after its launch in January 2001, the potential of Wikipedia as a large scale source of knowledge for Artificial Intelligence and Natural Language Processing in particular became apparent to researchers in the field. The appeal of Wikipedia is that it strikes a middle ground between accurate, manually created, limited-coverage resources such as WordNet [9], Cyc [18], general purpose (SUMO, [33]) or domain-specific ontologies (Gene Ontology,[1] UMLS[2]), dictionaries and thesauri, and automatic, wide-coverage, but still noisy knowledge mined from the web [38].

Unlike resources prepared by trained linguists, Wikipedia's structures have arisen through the collaboration of contributors and, with the exception of the category structure which was encouraged by the contribution guidelines, without prior planning. This may bring the quality of a resource based on such underspecified criteria into question, but its usefulness in a variety of Natural Language Processing (NLP) tasks has already been shown [22]. The category structure was not intended to be an ontology-like structure, but what has emerged is a *folksonomy*, mirroring the shared categorization preferences of the contributors. The collaborative aspect has also led to the implicit encoding of much information that when made explicit, reveals millions of new bite-sized pieces of knowledge.

* Corresponding author.
E-mail addresses: vivi.nastase@h-its.org (V. Nastase), michael.strube@h-its.org (M. Strube).

[1] http://www.geneontology.org.

[2] http://www.nlm.nih.gov/research/umls/.

0004-3702/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artint.2012.06.008

Wikipedia contains a wealth of multi-faceted information: articles, links between articles, categories which group articles, infoboxes, a hierarchy that organizes the categories and articles into a large directed network, cross-language links, and more. These various types of information have been usually exploited independently from each other.

This paper presents WikiNet[3] – the result of jointly bootstrapping several information sources in Wikipedia to produce a large scale, multilingual and self-contained resource. The starting point is the category and article network. The most interesting feature of our approach is that it works completely automatically, in that it itself discovers relations in Wikipedia's category names for which it then finds numerous instances based on the category structure.

Building WikiNet involves three main steps. First, category names are deconstructed to retrieve the categorization criterion, which leads to the discovery of numerous binary relation instances. In the second step the relation instances discovered in the first step are refined based on information in the articles' infoboxes. In the last step the network obtained up to this point is formalized by merging nodes that refer to the same concept, and by adding lexicalizations for these concepts from redirect, disambiguation and cross-language links from Wikipedia versions in different languages. The resulting resource is a network consisting of 3 707 718 concepts and 49 931 266 relation instances (for 454 relations),[4] and covers multiple dimensions: multilinguality, world knowledge, lexical semantics, collocations, paraphrases, named entities. Because the processing does not rely on manual feedback, and both the relations and their instances in the network are automatically discovered in Wikipedia's categories and infoboxes, the algorithm can easily be applied to the latest Wikipedia versions to generate an updated resource.

Intrinsic evaluation of the knowledge extracted shows that combining different types of information leads to the derivation of accurate facts, not overtly expressed within articles or infoboxes, and as such not to be found by processing single aspects of Wikipedia. We contrast this approach with DBpedia [1] and YAGO [45] – the largest repositories of facts extracted from Wikipedia to date. We perform extrinsic evaluation through two tasks – semantic relatedness computation between pairs of terms, and metonymy resolution, i.e. finding the correct interpretation of terms which are not used in any of their literal senses (e.g. *White House* is often used to refer to the *President of the United States*). The extrinsic evaluation results show that the resource is of good quality – evidenced by high correlation results with manually assigned relatedness scores on disambiguated data – but it also has high ambiguity which cannot be solved for pairs of terms out of context. Applying WikiNet to the task of metonymy resolution shows consistent increase in precision and recall when using world knowledge to find the correct interpretation of potentially metonymic words, but due to the small size of the available data these increases are not statistically significant.

2. Building WikiNet

The starting point for building WikiNet is the category and article network from one language version of Wikipedia. This network is modified step by step as more types of information from Wikipedia are taken into account. In the final step the nodes in the network are considered to represent concepts. Concepts and their lexicalizations are separated, and each concept – now represented through a language independent ID – has associated numerous lexicalizations in a variety of languages. An overview of the processing is shown in Algorithm 1, and each step is presented in more detail in the remainder of the section.

Algorithm 1 Algorithm for building a large scale multilingual knowledge network.

```
Input:
    $W$ – the English Wikipedia dump
    $\mathcal{R}$ – a set of relational nouns
    $\{W_X\}$ – a set of additional Wikipedia dumps in different languages
Output:
    WikiNet – a graph with nodes as concepts, and edges as relations between them

$R_1 = \text{DeconstructWikipediaCategories}(W, \mathcal{R})$
$R_2 = \text{PropagateInfoboxRelations}(W, R_1)$
return $\mathit{WikiNet} = \text{BuildConceptNetwork}(R_2, W, \{W_X\})$
```

The Wikipedia dump W is the file containing all English Wikipedia articles in XML format,[5] and $\mathcal{R}$ is a set of relational nouns extracted from an existing resource (NOMLEX,[6] Meyers et al. [23]) used for detecting one of four classes of relations in Wikipedia category names. The result of the category deconstruction process – R_1 – is a set of relation instances, represented as tuples (x, r, y), where x, y are strings, some of which are Wikipedia article or category names, others are fragments of category names, and r is a relation, also derived from the category names. R_2, the result of infobox relation

[3] This article builds upon and expands [27,29]. It expands on this previous work by using a list of relational nouns extracted from NOMLEX. It presents a new method for computing semantic relatedness between a pair of terms, and its evaluation with standard data sets. It presents experiments on embedding the resource into an NLP task, in particular metonymy resolution. WikiNet can be downloaded from http://www.h-its.org/english/research/nlp/download/wikinet.php.

[4] The statistics reported in this paper refer to the WikiNet built starting from the English Wikipedia dump of 2011/01/15, and adding several other language versions. Details are in Section 3.1.

[5] Wikipedia dumps are available from http://download.wikimedia.org/. We use the pages-article.xml.bz2 file.

[6] http://nlp.cs.nyu.edu/meyers/nombank/nombank.1.0/NOMLEX-plus.1.0.

propagation, has the same structure as R_1, with the difference that some of the previously extracted relation instances are assigned new relations. *WikiNet*, derived from R_2 and additional information from W and $\{W_X\}$, is a graph. The nodes are concepts, which are identified through a unique numeric ID and have associated multiple lexicalizations in various languages. The edges are relation instances between concepts corresponding to the tuples in R_2, after mapping the arguments onto concepts and filtering out the tuples for which at least one argument could not be mapped onto a concept.

Each processing stage transforms the structure produced by the previous stage. The starting point is a Wikipedia dump, based on which we build a network whose nodes are the pages and categories and whose edges are the category–category and category–page links. The category deconstruction step (DeconstructWikipediaCategories) adds new nodes – substrings of category names – and (named) edges to the network, and it renames some of the existing edges. The infobox relation propagation step (PropagateInfoboxRelations) renames existing edges. After these two steps we build the actual network (BuildConceptNetwork), by collapsing together nodes (page/category nodes/category name substrings) that refer to the same entity, and by associating with each node numerous lexicalizations in multiple languages.

In the following discussion we will make use of the terminology included in Fig. 1.

part of speech (POS): The part of speech is the word class of a word – e.g. noun, verb, adjective, adverb, determiner, pronoun.

phrase: Phrases are (grammatical) elements of clauses. There are five types of phrases: verb phrases, noun phrases, adjective phrases, adverbial phrases and prepositional phrases.

head word: The head word defines the syntactic (and frequently also the semantic) properties of the phrase.

noun phrase (NP): A noun phrase is a syntactic phrase whose head word is a noun.

constituent: A constituent is a fragment of a larger grammatical construction that is itself a proper grammatical construction (word/phrase/clause). In this paper we refer to constituents of a noun phrase, and we only consider constituents which are noun phrases themselves.

dominant/head constituent: A dominant constituent is the constituent of a phrase that has the same head word as the parent phrase.

relational noun: Relational nouns are nouns that imply a relationship, e.g. *member, president*.

relation: A relation – such as *is_a*, *president*, *caused_by* – describes a type of connection. In our case, we assume binary relations, that require two arguments.

relation instance: A relation instance is a triple (x, r, y), where r is a relation, and x and y are specified concepts.

concept: In this paper we use the term *concept* to refer to what are called concepts (e.g. **mathematics**, **space shuttle**, etc.) as well as named entities (e.g. **Space shuttle Atlantis**), because from an algorithmic point of view they are all treated the same. The resource however includes named entity information, thus allowing the two to be distinguished.

category: By category we denote the Wikipedia category with all its implied structure (subsumed categories and pages).

category name: When we need to differentiate between the category as a structure and its name, we refer to the name explicitly.

Fig. 1. Glossary of relevant terminology.

2.1. Deconstructing Wikipedia categories

To organize Wikipedia for easy access to pages, contributors are given guidelines for categorizing articles and naming new categories. A quick inspection reveals that categories[7] are noun phrases, many of which – e.g. Albums by artist, People from Heidelberg, Members of the European Parliament – do not correspond to the type of lexical concepts we would expect to encounter in texts and are not needed for processing a text automatically. Instead, they capture examples of human classification and relations that can be used as a source of information [27]. Complex categories combine multiple classification criteria – People from Heidelberg contains pages about people who are from Heidelberg. From this perspective, deconstruction of categories can be interpreted as separating each classification criterion captured in the category.

Analysis of category names reveals several types, based on the type of information they encode. We present them succinctly in Table 1, and then in more detail in the subsections to follow.

2.1.1. Explicit relation categories

These categories overtly express a relation that is common to all articles in a category. The relation can be expressed through a relational noun – e.g. member, president – or through a verb–preposition combination – e.g. caused by, founded in – corresponding to two types of explicit relation categories:

Relational nouns Relational nouns are nouns that indicate a relation. Their meaning is complete when their arguments are also present: e.g. *prime-minister*. Such nouns are present in category names and give clues about specific properties of the subsumed articles. Let us take for example the category Members of the European Parliament: All articles in this category describe some $\mathbf{X}_i$, such that $\mathbf{X}_i$ *member_of* **European Parliament**.

[7] Sans Serif is used for patterns and words, *italics* for relations, Small Caps for Wikipedia categories and pages, **bold Small Caps** for concepts. Part of speech tags and grammatical categories are capitalized.

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

Category name type	Pattern	Relation instances
explicit relation (Section 2.1.1)		
Queen (band) members	X members members of X	**Freddy Mercury** *member_of* **Queen (band)** **Brian May** *member_of* **Queen (band)** . . .
Movies directed by Woody Allen	X [VBN IN] Y	**Annie Hall** *directed_by* **Woody Allen** **Annie Hall** *is_a* **Movie** **Deconstructing Harry** *directed_by* **Woody Allen** **Deconstructing Harry** *is_a* **Movie**
partly explicit relation (Section 2.1.2)		
Villages in Brandenburg	X [IN] Y	**Siethen** *located_in* **Brandenburg** **Siethen** *is_a* **Village**
implicit relation (Section 2.1.3)		
Mixed martial arts television	X Y	**Mixed martial arts** $\mathcal{R}$ **television programs** **Tapout (tv series)** $\mathcal{R}$ **Mixed martial arts** **Tapout (tv series)** *is_a* **television program**
class attribute (Section 2.1.4)		
Albums by artist	X by Y	**artist** *attribute_ of* **Album** **Miles Davis** *is_a* **artist** **Big Fun** *is_a* **album**

Categories of this type can be identified if their name matches a pattern NP_1 NP_2 or NP_2 (of|of the) NP_1, where the head of the noun phrase NP_2 is a relational noun, such as member, president, prime-minister. To recognize explicit relations that involve a relational noun, we use a set $\mathcal{R}$ of singular and plural forms of relational nouns from NOMLEX [23] (699 word forms). *member* is an extreme example of a relational noun, in that it does not have meaning in the absence of both its arguments. *president*, on the other hand, is informative even when only one of its arguments is present – e.g. Barack Obama is a president. For this reason, the relations instances $\mathbf{P}_i$ *is_a* r_n are also added, where $r_n \in \mathcal{R}$ is the relational noun that matches (part of) the category name.

Verb–preposition combinations In Wikipedia categories, verb past-participle and preposition combinations – such as *directed by, built in* – indicate a relation. The category Airplane crashes caused by pilot error provides an example. All articles in this category describe airplane crashing events, *caused_by* pilot error. These categories match the part of speech pattern NP_1 VBN IN NP_2.[8] The relation consists of the verb–preposition combination in the category name. Recognizing this type of explicit relations relies on part of speech tags; to obtain them we use Stanford's POS tagger.[9]

2.1.2. Partly explicit relation categories

Prepositions, although sometimes ambiguous, are strong indicators of semantic relations [17]. The preposition of, for example, may indicate a spatial (Treasure troves of Europe) or a temporal (Treasure troves of the Iron Age) relation, and the same is the case for the preposition in: Villages in Brandenburg encodes a spatial relation, Conflicts in 2000 captures a temporal one.

Categories of this kind have the pattern NP_1 IN NP_2 (IN is the part of speech tag for (all) prepositions). Note that there is no overlap with explicit relation categories whose patterns are more restrictive (and will be attempted first). For partly explicit relation categories, we do not have the constraint that the noun phrase is headed by a relational noun, nor do we have a verb appearing in the category name.

To determine the relation R expressed in a category with the pattern NP_1 IN NP_2, we use the supercategories of NP_1 and NP_2 – S_{NP_1} and S_{NP_2}:

- if S_{NP_1} matches person or people, and S_{NP_2} is organization or group, the relation assigned is *member_of*;
- if S_{NP_2} matches 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 the preposition in, etc.). To facilitate the evaluation process, all spatial relations detected are named *spatial*, and their instances are labeled accordingly.
- if S_{NP_2} matches time, the relation assigned is *temporal*.

[8] VBN is the part of speech tag for participles, and IN is the part of speech tag for prepositions in the Penn Treebank POS tag set [43].

[9] http://nlp.stanford.edu/software/tagger.shtml.

2.1.3. Implicit relation categories

Some category names are complex noun compounds. These do capture relations, but do not give any (overt) indication of what the relation is: MIXED MARTIAL ARTS TELEVISION PROGRAMS has two noun phrase components – **MIXED MARTIAL ARTS** and **TELEVISION PROGRAMS** – and the relation between them, encoded in the category name, is *topic*. These categories have the pattern NP_1 NP_2.

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 Fig. 2.

Each embedded phrase is considered to be 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^h according to the syntactic structure of the category name (in our example, $C_2 = C_1^h$, i.e. C_2 dominates C_1). The constituent which corresponds to the phrase head is the dominant/head constituent of the category name and is denoted by C^h (C_2 is also C^h in the above example). We use only the noun phrase constituents of a category name, and denote the constituents accordingly as NP_j, NP^h.

Fig. 3 shows examples of relations and some of their instances induced for this type of category. The process is shown in detail below.

1. add relation instances $\mathbf{P}_i$ *is_a* $\mathbf{NP}^h$;
2. form pairs ($\mathbf{NP}_j$, $\mathbf{NP}^h$) for all NP_j for which $NP_j^h = NP^h$ – form constituent pairs in which the first constituent is dominated by the main dominant constituent. Determine the relation r that holds between ($\mathbf{NP}_j$, $\mathbf{NP}^h$) (detailed below);
3. add relation instances $\mathbf{P}_i$ r $\mathbf{NP}_j$.

   Propagating the relation r from the category constituents to the pages follows the rule captured in Fig. 4:
   if $\mathbf{P}_j$ *is_a* $\mathbf{NP}^h$ *and* $\mathbf{NP}^h$ r $\mathbf{NP}_x \Longrightarrow \mathbf{P}_j$ r $\mathbf{NP}_x$.

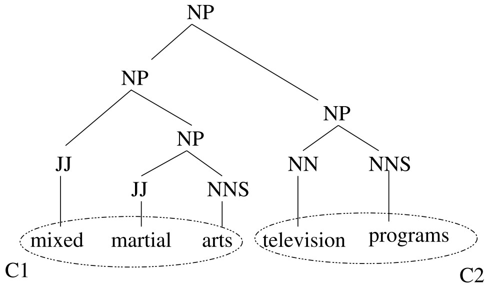


Fig. 2. Example of parse tree for a category name.

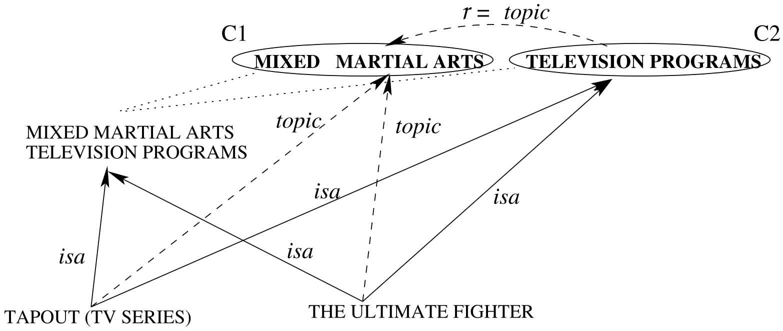


Fig. 3. Example of relations and some of their instances induced after extracting components of a category name.

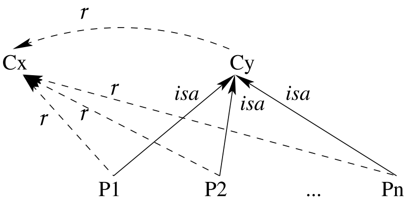


Fig. 4. Propagating the relation between category constituents to the subsumed pages.

Finding the relation between one pair ($\mathbf{NP}_x$, $\mathbf{NP}^h$) means automatically finding the relation between numerous ($\mathbf{P}_j$, $\mathbf{NP}_x$) pairs.

2.1.4. Class attribute categories

For categories with names that match the pattern [NP_1 by NP_2], we identify NP_1 as a class and NP_2 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 Fig. 5, NP_1 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 instance **MILES DAVIS** *is_a* **ARTIST**, as shown in Fig. 5.

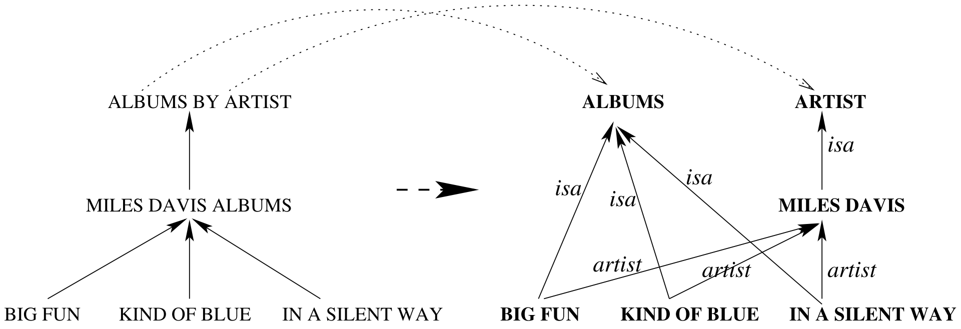


Fig. 5. Example of relations and some of their instances inferred from "by" categories.

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 case we start processing the attribute first (NP_2):

1. if the attribute is a category in Wikipedia, collect the pages it subsumes (P_{ai}) as possible attribute values;
2. if a P_{ai} appears in the subcategory name, it serves as confirmation that this is a possible attribute value and we add the links $\mathbf{P}_{ai}$ *is_a* $\mathbf{NP}_2$ and $\mathbf{P}_i$ *is_a* $\mathbf{NP}_1$ for each page P_i subsumed by the "by" category;
3. extract the remainder of the subcategory name as an instance of $\mathbf{NP}_1$.

In the example above, $NP_1 =$ *heads of government*, $NP_2 =$ *country*. We expand COUNTRY[10] to all its pages and test whether any of them appear in the name of the subcategory PRIME MINISTERS OF CANADA. We thus identify $P_{ai} =$ *Canada*, and add the links **CANADA** *is_a* **COUNTRY** (step 2) and **PRIME MINISTER** *is_a* **HEADS OF GOVERNMENT** (step 3). For all pages P_i under HEADS OF GOVERNMENT BY COUNTRY, the relation instances P_i *is_a* **HEADS OF GOVERNMENT** are added (step 2).

2.1.5. The algorithm

Using the information obtained from processing the category names as detailed above, we induce knowledge which is added to the original category and article network. This process is sketched out in Algorithm 2. In the preceding discussion we have presented category types based on the type of relation they encode. In the algorithm the ordering is in reverse order of the specificity of the matching pattern, meaning that more constraining patterns are applied first.

To apply the aforementioned rules, the category names are processed with the POS tagger and parser developed by the Stanford NLP group.[11]

The result of this first stage of the process is a network with a heterogeneous mixture of nodes: Some represent categories, some pages, some strings obtained after splitting category names. To obtain the WikiNet version whose statistics we included in this paper, at this point of processing we had a network with 70 540 640 edges and 3 885 940 nodes.

2.2. Propagating infobox relations

In the previous step the information encoded in Wikipedia's category names was used to induce relations and their corresponding instances. Some category names provide only very general clues and general relations, such as *temporal* or *spatial*. For practical reasons it may be useful to have more specific relations. To obtain them we look at links between category names and information in the infoboxes: The category name encodes the categorization/grouping criterion that is respected by all of the subsumed pages, while the infoboxes contain a summary of the most important information in the corresponding pages and the categorization criterion may be part of that.

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

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

Algorithm 2 Deconstructing Wikipedia categories.

```
Input:
    $W$ – a Wikipedia dump
    $\mathcal{R}$ – a set of relational nouns
Output:
    $R_1$ – a set of binary relation instances

$W_C$ = the set of categories from $W$
$R_1 = \{\}$
for $w \in W_C$ do
    $P$ = set of pages under $w$
    if $w$ matches "NP$_1$ by NP$_2$" then
        // Class attribute categories, Section 2.1.4
        $R_1 = R_1 \cup$ ProcessByCategory($w$)
        $R_1 = R_1 \cup$ ExtractClassAttributes($w$, W)
    else
        if $w$ matches "$r$ (of|of the) [NP]" or "[NP] $r$", where $r \in \mathcal{R}$ then
            // Explicit relation category, relational nouns, Section 2.1.1
            $R_1 = R_1 \cup \{P_i\, r\, [NP] \mid \forall P_i \in P\}$
        else
            if $w$ matches "NP$_1$ VBN IN NP$_2$" then
                // Explicit relation category, verb–prep combination, Section 2.1.1
                $R_1 = R_1 \cup \{(P_i\ \text{[VBN IN]}\ NP_2), (P_i\ \mathit{is\_a}\ NP_1) \mid \forall P_i \in P\}$
            else
                if $w$ matches "NP$_1$ IN NP$_2$" then
                    // Partly explicit relation category, Section 2.1.2
                    $r$ = DetermineRelation(IN, NP$_1$, NP$_2$)
                    $R_1 = R_1 \cup \{(P_i\, r\, NP_2), (P_i\ \mathit{is\_a}\ NP_1) \mid \forall P_i \in P\}$
                else
                    if $w$ is a complex noun compound then
                        // Implicit relation category, Section 2.1.3
                        $C$ = ExtractConstituents($w$)
                        $C = \{NP_x, NP^h\}$, $NP^h$ is the dominant (head) constituent of $w$
                        $R_1 = R_1 \cup \{(P_i\ \mathit{related\_to}\ NP_x), (P_i\ \mathit{is\_a}\ NP^h) \mid \forall P_i \in P\}$
return $R_1$
```

We hypothesize a connection between the information encoded in category names and the information summarized in infoboxes, and use this to propagate relations from infoboxes through the category network, as shown in Fig. 6. Let us have a closer look at this example. First, the category name deconstruction step splits the string Military equipment of the Soviet Union into two parts: Military equipment and Soviet Union. It then proposes a rather generic *spatial* relation based on the preposition of and the fact that Soviet Union has a corresponding article whose ancestor is the Geography category. In the final step of the category deconstruction process, we have added the relation instances (P_i *spatial* Soviet Union) for all pages P_i subsumed by the category Military equipment of the Soviet Union. We would now like to find a more informative relation to replace the generic *spatial* in the relation instances mentioned before. Several of the articles under the category Military equipment of the Soviet Union contain infoboxes,[12] in which we find Soviet Union as the value of the attribute *place-of-origin*. This attribute becomes the relation, and the previously extracted relation instances are replaced by (P_i *place-of-origin* Soviet Union) for all pages P_i subsumed by the category Military equipment of the Soviet Union. This can be seen as propagating a relation first from the pages that contain infoboxes to the category which subsumes them, and then from this parent category to all the other page siblings that did not have an infobox.

Working "on top" of an existing network introduces an important difference and advantage relative to work that extracts knowledge from open texts: A piece of extracted knowledge – a predicate–arguments tuple – is immediately connected to other such facts, and it has a context. Because of this, every piece of information we add to this network has an impact on its neighbours, and, depending on its type, its influence can reach far. This happens when establishing the connection between the information in infoboxes for some pages and their parent categories – it impacts their siblings.

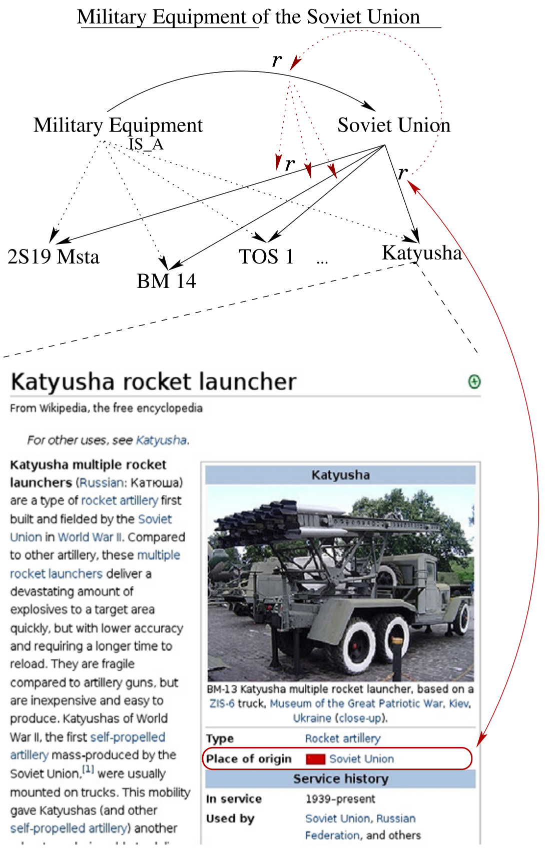
In other words, r is a candidate relation for the constituents pair (NP^h, NP_i) corresponding to category C, if it is associated with the same value V in all infoboxes in which it appears under C, and if V is compatible with NP_i. Two values are compatible if they are identical, paraphrases of each other, or are connected in a systematic way – V is an instance or a concept more specific than the one corresponding to NP_i, or for locations for example V *part-of* NP_i (a specific location in Europe is compatible with **Europe**). This predicate can become more specialized as more relation instances are added to the fact base.

Algorithm 3 shows the processing steps for linking a category name with information in the infoboxes of its subsumed pages and for propagating the induced relations to the page–category links.

The result of this processing step is a network consisting of the same nodes and edges as in the previous processing step, with the difference that the names of certain edges are now changed, according to the information in the infoboxes.

[12] When processing a category, we consider all subsumed articles, including those subsumed by subcategories.

Please cite this article in press as: V. Nastase, M. Strube, Transforming Wikipedia into a large scale multilingual concept network, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.06.008



Katyusha rocket launcher

From Wikipedia, the free encyclopedia

For other uses, see Katyusha.

Katyusha multiple rocket launchers (Russian: Катюша) are a type of rocket artillery first built and fielded by the Soviet Union in World War II. Compared to other artillery, these multiple rocket launchers deliver a devastating amount of explosives to a target area quickly, but with lower accuracy and requiring a longer time to reload. They are fragile compared to artillery guns, but are inexpensive and easy to produce. Katyushas of World War II, the first self-propelled artillery mass-produced by the Soviet Union,[1] were usually mounted on trucks. This mobility gave Katyushas (and other self-propelled artillery) another

Katyusha	
BM-13 Katyusha multiple rocket launcher, based on a ZIS-6 truck, Museum of the Great Patriotic War, Kiev, Ukraine (close-up).	
Type	Rocket artillery
Place of origin	🟥 Soviet Union
Service history	
In service	1939–present
Used by	Soviet Union, Russian Federation, and others

Fig. 6. Some articles under a Wikipedia category contain infoboxes with hand-picked relations, from which 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.

Algorithm 3 Linking to infoboxes and propagating relations.

```
Input:
    $W$ – a Wikipedia dump
    $R_1$ – the set of relations extracted in the category name processing step
Output:
    $R_2$ – a set of binary relations

$W_C$ = the set of categories from $W$
$R_2 = R_1$
for $w \in W_C$, $w \mapsto \{NP_1, \ldots, NP_i\}$ according to the previous deconstruction step do
    $P$ is the set of pages subsumed by $w$
    build set of candidates $R = \{r_i \mid (r_i, NP_x) \in P_j\}$ – the set of attributes which are associated with one of the category's constituents $NP_x$ extracted from infoboxes from pages $P_j \in P$
    if $|R| = 1$ and matching constituent is $NP_x$ then
        $r \in R$
        for $P_j \in P$ do
            replace relations $(P_j\, r_x\, NP_x)$ in $R_2$ with $(P_j\, r\, NP_x)$
    else
        if $|R| > 1$ and matching constituent is $NP_x$ then
            if all relations in $R$ are compatible then
                for $P_j \in P$ do
                    replace relations $(P_j\, r_x\, NP_x)$ in $R_2$ with $(P_j\, r\, NP_x)$
return $R_2$
```

2.3. Building a concept network

2.3.1. Mapping relation arguments to concepts

Up until this point we have built a heterogeneous network based on various types of information in Wikipedia. Now this network will be transformed into a concept network. In works using Wikipedia as a source of knowledge, categories and articles are considered to represent concepts [36,25]. We start from this premise as well, and constrain it to eliminate redundancy – there should only be one node representing a concept. Mapping onto one node articles and categories that refer to the same concept (e.g. Cities and City) or are homonyms (there is a Rome article and a category) is trivial. A more complex issue arises when mapping the fragments resulting from deconstructing the category names onto the original Wikipedia nodes – a prerequisite for the final node mapping step.

The category Chemistry albums, for example, is split into two parts: *Chemistry* and *albums*, both of which are ambiguous with respect to the nodes corresponding to articles or categories in the network. *Chemistry*, for example, can refer to the science, the band, various albums and songs. To determine the correct corresponding article or category we compute a connectivity score that counts the number of times each of these potential referents is linked to from pages subsumed by the category Chemistry albums. Formally, if the fragment NP_X of a category C is ambiguous, we collect the possible nodes (articles/categories) N_i that NP_X can refer to, and all pages P_j subsumed by C. We choose as an interpretation for NP_X the node N_i towards which there are the most hyperlinks (outlinks) in the pages P_j:

$$NP_X \mapsto N = \operatorname{argmax}_{N_i} \sum_{P_j \text{ subsumed\_by } C} outlink(P_j, N_i)$$

$$outlink(P_j, N_i) = \begin{cases} 1 & \exists \text{ a link from page } P_i \text{ to } N_i; \\ 0 & \text{otherwise} \end{cases}$$

Note that only the strings obtained by splitting the category names are ambiguous, whereas the article and category nodes are not ambiguous in the network (i.e. each node has a unique title and ID extracted from the Wikipedia dump).

This disambiguation step impacts the number of relation instances extracted in the previous step, because some of the arguments of the relation instances could not be linked to a node in the network. This step and duplicate filtering reduces the set to 49 931 266 unique instances.

2.3.2. Extracting alternative lexicalizations

Following WordNet's example, which separates synsets and their lexicalizations, in the newly created network nodes are replaced with numeric IDs, and then each ID is associated with the corresponding node's name. More lexicalizations are added to each node through the redirect, the disambiguation and the cross-language links.

Redirect links pair a redirect page with its target. Redirect pages can be viewed as containing name variations for Wikipedia articles. They may contain morphological variants (actor, actors and actress redirect to the article Actor), proper synonyms (adrenaline redirects to Epinephrine), paraphrases (Seinfeld, The show about nothing for the article Seinfeld (TV series)) or even misspellings (Sienfeld for Seinfeld). The names of redirects are added to the ID corresponding to the article they point to.

Disambiguation links map a disambiguation page onto its possible targets. Disambiguation pages encode polysemy – the name of the source of this link applies to all possible targets. King points to the pages corresponding to the monarch, the chess piece, and many others. The name of the source of the disambiguation link is added to the IDs of all of the possible targets.

Cross-language links link an article to its variants in other languages. The corresponding names can be used as translations. For the concept **Actor** we find the following language variations through the cross-language links Schauspieler (German), Attore (Italian), etc.

The network presented here is built from the information extracted from the English version of Wikipedia (W_{en}). There were several reasons for choosing the English Wikipedia: It is the most comprehensive version, there are numerous language processing tools that allow for deeper processing of the article texts if necessary (at this point only POS tagging and head extraction were used for processing the category names), capitalization rules allow for easy and accurate extraction of named entity information. More information can be added from Wikipedia versions for other languages. We use page, redirect, disambiguation and cross-language information from other language versions (W_l) to add lexicalizations of concepts in different languages, following Algorithm 4. Essentially we map together pages in different languages either through existing cross-language links, or by checking for overlap in their respective cross-language links, following the "triangulation" algorithm from Wentland et al. [48].

Note that we do not aim to build a lexicon of synonyms for each concept based on cross-language links, as was done by de Melo and Weikum [7], who filter out inaccurate cross-language links. As shown before with redirect and disambiguation links, the index contains more than synonyms, reflecting the various ways a concept is lexicalized in different languages.

Table 2 shows how the coverage of concepts and number of included lexicalizations changes when adding information from the German Wikipedia. Because the supplemental information comes from the German Wikipedia, it is not surprising that the highest increase is for German, but we see substantial differences for other languages as well, indicating that merging lexicalizations from different Wikipedia language versions is useful. The number of covered concepts has increased, meaning that we have established a mapping between concepts that were not linked through cross-language links; furthermore the number of lexicalizations has increased, meaning that for the existing entries we now have more lexicalization variations. The increase is much higher with respect to lexicalizations (5.13 times) than concept coverage (1.17 times), which is to be expected as there are several sources for lexicalizations (redirect, disambiguation, cross-language links), while concept coverage relies on mapping entries to each other. Other language versions are added in a similar fashion.

Algorithm 4 Adding lexicalizations from a Wikipedia version.

```
Input:
  $WikiNet = \{ID, Lex, R\}$ – the current version of WikiNet consisting of:
        $ID$ a set of concept IDs
        $Lex = \{L_{ID_x} \mid \forall ID_x \in ID\}$ a set of concept lexicalizations
        $R$ the set of relation instances between (concept) IDs
  $W_{Lg}$ – the Wikipedia dump in a specific language $Lg$
  $t$ – a threshold for overlap in concept lexicalizations
Output:
  $WikiNet'$ – WikiNet enriched with additional lexicalizations

Lex$'$ = Lex
extract $ID_{Lg}$ – the set of page IDs from $W_{Lg}$
extract $Lex_{Lg}$ – the lexicalizations of pages from $W_{Lg}$ through mapping page titles, redirect and disambiguation lexicalizations onto the corresponding page IDs
map pages from different language versions:
for $ID_{Lg,i} \in ID_{Lg}$ do
  for $ID_j \in ID$ do
    compute overlap $o_{ij} = |L_{ID_{Lg,i}} \cap L_{ID_j}|$ – the size of the intersection of the lexicalizations of concepts corresponding to concepts $ID_{Lg,i}$ in language $Lg$ and $ID_j$ in the current version of WikiNet
  if $\max_{ID_j, ID_j \in ID} o_{ij} > t$ then
    $n = \operatorname{argmax}_{j, ID_j \in ID} o_{ij}$
    $ID_{Lg,i} \mapsto ID_n$
    $L'_{ID_n} = L_{ID_{Lg,i}} \cup L_{ID_n}$
    replace $L_{ID_n}$ in Lex' with $L'_{ID_n}$
return $WikiNet' = \{ID, Lex', R\}$
```

Table 2
Partial language statistics: Concept coverage and number of lexicalizations in WikiNet before and after adding information from the German Wikipedia.

Language	No. of lexicalizations (no. of entries)	
	Before	After
English	7 239 290 (2 996 357)	7 515 310 (2 996 357)
French	523 835 (509 051)	728 349 (589 448)
German	484 688 (474 455)	2 488 983 (556 977)
Italian	384 192 (378 896)	555 452 (457 158)
…		
Hungarian	89 413 (87 526)	127 487 (115 558)
Romanian	89 030 (84 694)	137 582 (114 284)
Turkish	86 543 (80 429)	117 662 (103 733)
…		
Russian	265 062 (254 740)	415 638 (334 106)
Japanese	264 547 (263 186)	383 387 (323 239)
Chinese	154 404 (151 957)	221 398 (186 938)

The current version of WikiNet merges lexicalizations from English, Chinese, Dutch, French, German, Italian, Japanese and Korean Wikipedia dumps.

Anchor texts – the text fragments associated with hyperlinks in a page – are an additional lexicalization source. While they were shown to be useful for detecting concepts in texts [26], they are too noisy to be added to the resource.

2.3.3. Named entity information

Named entity (NE) information is an important feature of a concept. To retrieve it we use an approach similar to Bunescu and Paşca [5] for single words. For multi-word terms we resort to syntactic analysis, as detailed below. This processing is applied to texts from the English Wikipedia, and it relies on capitalization conventions for this language. Because the named entity information once extracted is attached to concepts, it becomes available in all languages that have a lexicalization for that concept.

- For single word entries w, we compute $c_{cs}(w)$ as the number of occurrences of the word (case-sensitive) in the text of its corresponding article, and $c(w)$ is its total number of occurrences (case-insensitive) in the text:

$$NE(w) = \begin{cases} TRUE\colon & \frac{c_{cs}(w)}{c(w)} < \tau \\ FALSE\colon & \frac{c_{cs}(w)}{c(w)} \geqslant \tau \end{cases}$$

$$\tau = 0.2$$

Please cite this article in press as: V. Nastase, M. Strube, Transforming Wikipedia into a large scale multilingual concept network, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.06.008

- For multi-word entries ($\overline{w}$), we match the NE information with the NE information of its syntactic head: $NE(\overline{w}) = NE(w)$, where w is the syntactic head of $\overline{w}$. This replaces [5] heuristic of assigning NE information based on capitalization only, which fails for categories like History of Europe, or Economy of the United States of America.

Of the 3 707 718 concepts in the network, 2 507 052 are tagged as named entities according to this method, leaving 1 200 666 common nouns/phrases.

3. Evaluation

To verify the quality of the extracted network we carry out intrinsic and extrinsic evaluations of the resource. The intrinsic evaluation only looks at the resource itself; to do this we evaluate the result of each step of processing. The extrinsic evaluation looks at the resource in context, and its usefulness for specific tasks. We perform this through two tasks – semantic relatedness computation between pairs of terms and metonymy resolution.

3.1. WikiNet overview

Before proceeding with the evaluation, we provide a summary of information regarding the latest version of WikiNet, built on the following Wikipedia dumps: 2011/01/15 English, 2011/01/11 German, 2011/02/01 French, 2011/01/30 Italian, 2011/06/28 Japanese, 2011/06/21 Korean, 2011/01/26 Dutch and 2011/06/28 Chinese.

concepts	3 707 718	
named entities	2 507 052	67.62%
(general) concepts	1 200 666	32.38%
number of languages[13]	196	
lexicalizations	128 505 704	
average lexicalization per concept	34.66	
unique strings	14 376 806	
unique strings when taking language into account	19 582 972	
relations	494	
relation instances	49 931 266	
between two named entities	13 529 382	27.09%
between a named entity and (general) concepts	31 423 078	62.93%
between two (general) concepts	4 978 806	9.97%

Detailed information about instances per relation and concepts covered for each language can be found here: http://www.h-its.org/downloads/nlp/stats.rels, http://www.h-its.org/downloads/nlp/stats.lang.

3.2. Intrinsic evaluation

3.2.1. Deconstructing Wikipedia categories

Processing the Wikipedia categories starts with loading the category network – in the form of nodes and category links extracted from Wikipedia dumps – and filtering out administrative categories (identified using keywords, e.g. stubs, articles, wikipedia). After this preprocessing, there are 277 918 categories in the network. We obtain 70 540 640 relation instances, 66 184 610 of which are distributed among the category processing steps as follows:

explicit relations categories: (Section 2.1.1)

relational noun pattern: 42 823 category names had a relational noun head and express 185 relations (e.g. *alumnus_of, president_ of, member_of, player_of, collections_of*).

VBN IN pattern: 14 238 category names match this pattern, and express 192 relations (e.g. *caused_by*, *written_by*, *based_in*).

partly explicit categories: 172 196 category names encode partly explicit relations, most of which are assigned *temporal* or *spatial* relations in this step, to be refined in the infobox relation propagation step (Section 2.1.2).

implicit relations categories: 39 049 category names give no overt information about the relation encoded (Section 2.1.3). The relations in this case are generically named *related_to*, and part of these will be named in the infobox relation propagation step.

class attribute categories: 9612 categories. Processing the category names reveals 840 classes with an average of 2.27 attributes (Section 2.1.4). A sample is presented in Table 3.

[13] We include in the table only statistics for the 196 languages that have at least 1000 terms represented in the resource.

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

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 4
Extracted relations and instances for each category type and manual evaluation results of some of the most frequent relations.

Category type	Relation	# categories	# inst. extracted	Precision	
				∩	∪
explicit		57 061	6 268 036		
	alumnus_of, *member_of*, *president_of*, *player_of*, …	29 189	4 355 964	95.56%	97.17%
	caused_by, *written_by*, *based_in*, …	17 297	1 912 072	94.37%	96.38%
partly explicit		172 196	24 058 117		
implicit		39 049	8 144 597		
class attribute		9612	27 713 860		
	is_a		17 714 648	76.40%	84.00%
	spatial		13 087 052	87.09%	97.98%

> The difference of 4 356 030 instances from the total extracted is made up of category–category and category–article links. Some of these instances will be updated with more informative relations after mapping relation arguments to concepts (Section 2.3.1).

Table 4 shows the number of (unique) extracted relation instances and evaluation results in terms of precision. For each of the two types of explicit relations – based on the VBN IN pattern, and the relational noun pattern – we extracted a random sample of 250 instances (covering different relations), which was manually annotated by two human judges. We evaluated separately *is_a* and *spatial*, which were two of the most frequent relations, by comparing against two other random samples of 250 instances annotated by two judges. For each of these four evaluation sets, the table includes two scores – one that corresponds to evaluation against the intersection ∩ (instances that the annotators agree are correct) and against the union ∪ (instances that at least one annotator marks as correctly assigned).[14]

In addition to the fact that it is easier to analyze a short phrase to extract a relation rather than a sentence or even a document, analyzing category names and the category and page network for knowledge acquisition has other advantages as well. The category names express very concisely a relation which may also appear in the article, but is expressed there in a more complex manner. We took the 42 711 *member_of* relation instances 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, indicating that further processing would have been necessary to derive this particular type of information, while by analyzing category names this information is readily available.

3.2.2. Propagating infobox relations

At this point, the data consists of 70 540 640 relation instances obtained by deconstructing categories and keeping the category–category and category–article links. Some of the relations' arguments correspond to Wikipedia pages, 1 049 724 of which contain instances of one of 4459 infobox types. Of the 277 918 categories deconstructed, for 42 060 a link was established between the category name (specifically, a constituent of the category name) and a value in the infobox. The link was established through 130 123 pages that contained infoboxes and an entry in the infobox corresponding to a category

[14] Cohen's κ on the four manually annotated data sets are 0.82, 0.73, 0.76 and 0.25 respectively (following the order in Table 4). However, the datasets are generated based on the system's output (the instances are randomly selected from the system's output), and as such are biased, which may lead to higher random agreement between the judges than would otherwise be expected in a purely random collection of relation instances, that both the system and the judges would then annotate.

name constituent, for a total of 175 350 (P_j, NP_i) page-constituent links. The information was propagated to a further 544 702 pages and their 698 929 relation instances to the corresponding category name constituent, as shown graphically in Fig. 6 (on page 8). This set of 698 929 relation instances is the result to be evaluated. Table 5 shows some of the most frequent relations, and their number of occurrences in this set.

Table 5
The most frequent propagated relations.

Relation	Count	Example
country	67 724	Duchy of Parma–Italy
subdivision_name	58 463	Aylmer, Quebec–Gatineau
location	40 525	Valvelspitze–South Tyrol
battles	39 425	John Paton–WWI
birth place	24 925	Franklin D. Roosevelt–New York
continent	24 228	1928 British Home Championship–Europe
region	18 590	1927 Crimean earthquakes–Crimea
genre	10 528	Cinepaint–graphics
industry	2056	Google–Internet

The propagation step is evaluated through two methods: (i) manual evaluation of two sets of relations carried out by two judges, (ii) manual and automatic evaluation of the overlap with YAGO's fact base[15] [45].

Manual evaluation Due to the large number of relations extracted and due to variation in their frequency, we split the set of relation instances obtained through propagation into two roughly equally large subsets – one corresponding to high frequency relations (they have more than 5000 instances), and one to low frequency ones. We extracted two samples of 250 relation instances – one from each of these two subsets – which contain the same distribution of relations as the subset they represent. This allows us to analyze a wider spectrum of relations, as low frequency relations would not appear in a small random sample that maintains the distribution of relations. The high frequency sample contains the following relations[16]:

battles (8), *birth_place* (8), *country* (22), *date* (7), *founded* (6), *genre* (26), *group* (5), *headquarters* (6), *industry* (10), *language* (6), *location* (34), *nationality* (14), *occupation* (5), *place* (5), *pushpin_map* (6), *region* (5), *ship_ country* (6), *sport* (6), *subdivision_name* (24), *subdivision_type* (4), *type* (18), *work* (6), *year* (6), *years_active* (7)

and the low frequency sample (we give a partial list):

address (2), *airdate* (2), *alma_mater* (3), *area_served* (2), *artist* (2), *associated_acts* (2), *author* (4), *awards* (4), *basin_countries* (3), *birthdate* (2), *birthplace* (5), *body* (2), *born* (2), *branch* (4), *bundesland* (3), etc.

These two samples were manually annotated by two human judges. The guidelines instructed the annotators to assign a `true/false/not_relation` tag to each instance. The instances in the annotation set were grouped by relation, and before each batch corresponding to one relation was included a positive example from an infobox to help the annotators. During annotation a few issues became apparent: There are "attributes" in infoboxes which are not really attributes (e.g. *pushpin map*, *caption*) which link to the included map or image (we call these "false" relations). There were 15 instances of such relations in each of the two sample files, 14 on which the judges had agreed, and 1 on which they did not. The *pushpin map* replaced the rather general *spatial* relation assigned in the category deconstruction step. Another issue arose from wrongly categorized articles (at least in the opinion of the annotators). While the propagation process may have been correct, the relation instance was tagged "false". For example **Tricia Leigh Fisher** was assigned the *occupation* relation to **child actors**, because it was categorized under American child actors. She started her acting career when she was 16 or 17, and one of the annotators considered the assigned category American child actors – and consequently the induced **Tricia Leigh Fisher** *occupation* **child actors** relation instance – "false".

The results in terms of precision are presented in Table 6, relative to `true` tags assigned by both judges ($\cap$) or by at least 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.[17]

Comparison with YAGO The versions of WikiNet and YAGO that we compare were not generated from the same Wikipedia download. To maximize the matching between the extracted relation instances and YAGO, we processed the arguments of our relations and those in YAGO by removing information in parentheses (e.g. Time (Unix) → Time), by lowercasing and by removing all blanks and all non-letter/digit characters. Additionally, we considered that a WikiNet relation instance matches a YAGO one if their arguments match, as the resources do not share the same relations. The overlap of the set of 698 929

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

[16] The list formatting is *relation* (*frequency*).

[17] Again, these numbers should be taken with the proverbial grain of salt. The samples are biased because they are extracted after the infobox propagation step, and as such the agreement between judges may be higher than random. See also footnote 14.

Please cite this article in press as: V. Nastase, M. Strube, Transforming Wikipedia into a large scale multilingual concept network, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.06.008

Table 6
Manual annotation results and evaluation, on the sample without "false" relations (filtered)/on the full sample.

Sample	Instances	Evaluation	
	filtered/all	∩	∪
high freq	235/250	78.3%/73.6%	86.8%/81.6%
low freq	235/250	75.7%/71.2%	77.9%/73.2%

relation instances renamed through relation propagation with YAGO's fact base is 7143 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. We look in a bit more detail at the relations connecting these pairs in our approach and in YAGO. 306 YAGO relations are represented within the 7143 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 this paper:

artist (126), *writer* (89), *developer* (58), *director* (37), *manufacturer* (14), *producer* (12), *composer* (8).[18]

The relations assigned by propagating relations from the infoboxes are more specific than the relations in YAGO for the overlapping pairs. In the manually annotated sample we have 14 instances annotated with relations from this set. Their precision is 87.5% (both ∩ and ∪). 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), *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) achieve a precision of 86.25% (∪)/83.75% (∩) (Table 7).

Table 7
Evaluation relative to the overlap with YAGO.

YAGO relation	Overlap	Precision
full evaluation		
wrote	972	99.07%
directed	757	98.94%
estimation based on manually annotated sample		
located in	3163	86.25% (∪)/83.75% (∩)
created	390	87.5%

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 the relation *directed* have the 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% (∪)/70% (∩).

3.2.3. *Full comparison with YAGO*

In the previous subsection we evaluated the results of the infobox propagation method through a comparison with YAGO. Here we perform a full evaluation of the network at this stage as compared with YAGO's core set of facts extracted from Wikipedia. Matching our relation instances and YAGO's is done as described in the previous section. The discussion on the mapping between the two lists of relations (types) was partly done in the previous section. After this processing we compute three measures of the overlap between our relation instances (R_W) and YAGO's (R_Y), following the method of Ponzetto and Strube [37] (itself derived from Navigli and Ponzetto [30]):

Coverage is the ratio between the number of relation instances shared by the two resources, and the relation instances in the reference resource (here, YAGO)[19]:

$$Coverage(R_W, R_Y) = \frac{|R_W \cap R_Y|}{|R_Y|} = \frac{598\,782}{10\,329\,767} = 5.8\%$$

It is clear from the low overlap between the resources that each contains information that the other one does not. *Novelty* and *ExtraCoverage* quantify this:

Novelty quantifies the novelty rate of our relation instances, as the ratio between relation instances that appear only in our set, and the full size of the extracted set:

$$Novelty(R_W, R_Y) = \frac{|R_W \setminus (R_W \cap R_Y)|}{|R_W|} = \frac{49\,332\,484}{49\,931\,266} = 98.8\%$$

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

[19] For our version of YAGO (downloaded 11/12/2009), we have counted 10 329 767 relation instances between Wikipedia concepts.

The Novelty of YAGO's core relations relative to our relations is:

$$Novelty(R_Y, R_W) = \frac{|R_Y \setminus (R_Y \cap R_W)|}{|R_Y|} = \frac{9\,730\,985}{10\,329\,767} = 94.2\%$$

ExtraCoverage shows the 'gain' in knowledge provided by our set of relation instances with respect to YAGO's set of facts, as the ratio between the number of relation instances found only in our set and the number of relations in YAGO:

$$ExtraCoverage(R_W, R_Y) = \frac{|R_W \setminus (R_W \cap R_Y)|}{|R_Y|} = \frac{49\,332\,484}{10\,329\,767} = 477.5\%$$

The ExtraCoverage of YAGO's core with respect to our resource is:

$$ExtraCoverage(R_Y, R_W) = \frac{|R_Y \setminus (R_Y \cap R_W)|}{|R_W|} = \frac{9\,730\,985}{49\,931\,266} = 19.5\%$$

These measures show that the resources are not redundant, they introduce additional knowledge relative to each other.

A particular surprise in the comparison with YAGO was the low overlap in relation instances (598 782). Our analysis shows that this is due to several factors: (i) the use of a different Wikipedia version, leading to missed matches due to changes in the article names; (ii) the *instanceOf* relation in YAGO link a Wikipedia entity to a WordNet synset, which will not match WikiNet relations; (iii) the overlap between WikiNet and YAGO consists mostly of category–category links, and there is a small number of relations that link two named entities common to the two resources.

3.2.4. Partial comparison with DBpedia

DBpedia [3] is a very large repository that transforms all the structured information that Wikipedia provides into a large database. It also includes links to other resources and a classification of entities into a manually defined ontology based on infobox types. While DBpedia is based on the automatic reformatting of already existing (and structured) information in Wikipedia, in building our resource the purpose was to uncover information that is not explicitly given. Because DBpedia contains the explicit information in Wikipedia – including category–category, category–article links and relations extracted from infoboxes – we can perform a partial comparison between WikiNet and DBpedia by evaluating the differences in structure between WikiNet and Wikipedia. This way we can also avoid an imperfect mapping between the two resources. Because we started with Wikipedia's category–category and category–article links, we first evaluate how many relation instances we add to this initial structure by deconstructing categories. 69.81% of the relation instances obtained after processing category names are novel. We further test how many of these novel relation instances appear in an infobox – 2.91% (2.03% of the total number of relation instances) appear in an infobox (917 999 instances). This means that 66.65% of the relations in WikiNet are novel both with respect to the category structure and with respect to the information in the infoboxes. On the other hand, we do not explicitly add relations from the infoboxes, as these are available directly and can, as such, be added, as long as they connect two Wikipedia pages, and do not express a value (e.g. Germany *area* 357 021 km^2).

3.2.5. Relations

The version of WikiNet described here has 454 relations. A partial histogram is presented in Fig. 7.

Compared to other resources built from Wikipedia, WikiNet has more relations. DBpedia and YAGO are based on the explicit information in Wikipedia, and as such their relation sets consist of the attributes in infoboxes. YAGO also identifies a small number of specific categories that provide relational information, such as 1975 BIRTHS, categories starting with

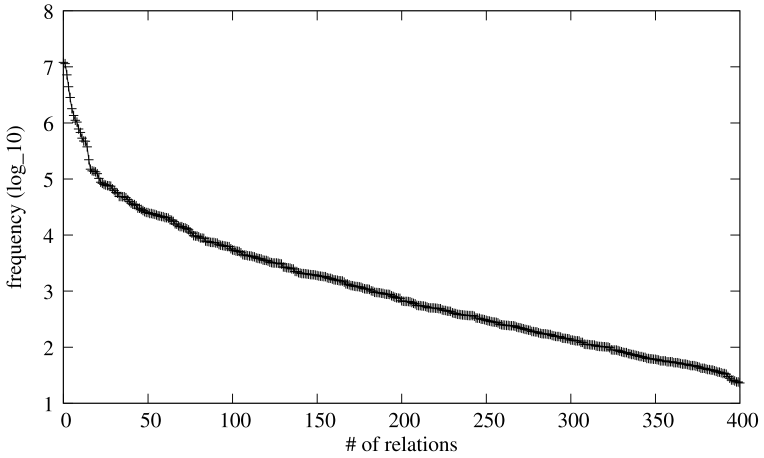


Fig. 7. Partial relations histogram on a $\log_{10}$ scale.

Countries in ..., Rivers of ..., Attractions in ..., and exploit them as a source of the following relations: *bornInYear, diedInYear, establishedIn, locatedIn, writtenInYear, politicianOf, hasWonPrize*. In the first stage of processing, the relation instances extracted by WikiNet are not explicit in Wikipedia, but rather encoded in the category names. They cover relational nouns, verb–preposition phrases, plus a small number of general and structural relations (*is_a*, *temporal*, *spatial*, *subcat_is_a*, *related_to*, *category*). Some of the relations extracted through this process are unique to WikiNet – e.g. *caused_by*, ... The comparison with YAGO shows that for the same pairs, the relations in WikiNet are more fine-grained.

3.2.6. Named entity information

We evaluate the results of the named entity (NE) identification process against the manually annotated data used to evaluate HeiNER,[20] which reimplements the Bunescu and Paşca [5] heuristics. The data and manual annotation statistics are presented in detail in Wentland et al. [48]. Two judges annotated a dataset containing 2000 NE candidates from Wikipedia, and three judges annotated a second set of 2000 candidates. In both cases a candidate is considered a true NE if at least two judges tag it as such. Comparison with our NE tags results in the following precision scores: 94.42% against set 1 and 96.16% against set 2, for an average of 95.29%.[21] Wentland et al. [48] report an average precision of 95% for HeiNER.

3.3. Extrinsic evaluation

The evaluation of the individual processing steps was shown as we progressed through the network building process. In this section we show the results of an extrinsic evaluation of the complete network, through semantic relatedness computation and metonymy resolution. These two tasks assess different aspects of the network. In relatedness computations the information provided by the structure is “compacted” in a numeric score. In metonymy resolution we use the structure of the network explicitly, by mining for specific relations.

3.3.1. Semantic relatedness

One of the basic tasks for which a lexical/knowledge resource is used in NLP is establishing the similarity or relatedness between terms. The suitability of a resource is assessed based on word pairs previously annotated by human judges, such as Rubenstein and Goodenough’s 65 noun pairs [42] (*R&G*), and Miller and Charles (*M&C*) 30 pairs subset of *R&G* [24]. High correlation between similarity and relatedness scores automatically computed on a resource and manually assigned scores is usually considered evidence of the high quality of a resource.

Previously used resources include WordNet and Roget’s Thesaurus, and most methods for computing similarity and relatedness have been designed by taking these resources’ specific attributes (structure, size) into account. Budanitsky and Hirst [4] present a comprehensive comparison of existing methods with their own for such computations. Recently Tsatsaronis et al. [46] proposed a new method for similarity computations in WordNet. Ponzetto and Strube [35,37], Milne and Witten [25], Gabrilovich and Markovitch [10,11] propose methods for semantic similarity/relatedness computation based on Wikipedia.

In comparison with WordNet and Roget’s Thesaurus and even with WikiTaxonomy [36,37], WikiNet is larger by an order of magnitude, and has many more relation instances (two orders of magnitude) and more varied relations. To deal with this structure and size, we implemented a novel metric. The structure is still a network, so the new methods still rely on connecting nodes in this network. From the methods implemented for WordNet we have learned that not all concepts are treated the same: For establishing similarity, nodes in the WordNet *is_a* hierarchy are considered more “important” the more specific they are, or the higher their information content relative to a corpus [41]. This observation is particularly relevant to our network, where nodes representing concepts such as **PEOPLE** or **CITIES** have hundreds of thousand of instances. When computing paths between concepts, expanding such a node causes a massive growth in the number of partial paths, which makes the computation very expensive both in terms of time and of CPU. On the other hand, such nodes do serve to connect many different concepts, and should be taken into account, as should the concepts they relate to through relations other than *is_a*.

To connect two nodes in the network while keeping the computation within a reasonable amount of time, the number of partial paths must be kept small. There are two ways to control this: have an upper bound on the number of steps taken from the starting node, and control the paths taken from each node reached. The upper bound for the path length between any two nodes is $2 * maxDepth$, where $maxDepth = 16$ is the maximum depth of the network[22] computed with respect to the original category hierarchy. This translates to an upper bound of 32 in the number of expansion steps.

To control the multiplication of partial paths, we also impose a bound on which nodes to expand, and how. For example, in searching for paths we would not want to expand a node with all of its instances, especially for nodes that have thousands or even hundreds of thousands of children. But we would want to follow links that lead to their supercategories, or superconcepts, for instance. Because of this, the bound imposed on expanding concepts is as follows: For each node N_x with

[20] http://heiner.cl.uni-heidelberg.de.

[21] For set 1 there were 8 instances for which we could not identify a corresponding entry in WikiNet’s index of concepts, and for set 2 there were 16 such cases. This situations arises from the fact that a non-ambiguous name in the version used by Wentland et al. [48] has become ambiguous in our version.

[22] For the 2009/07/13 English Wikipedia version.

relations r_{xi}, and relation counts $c_{r_{xi}}$, relations of the type r_{xi} are expanded if and only if $c_{r_{xi}} \leqslant \tau_r$ where τ_r is a threshold determined empirically. In our experiments we set $\tau_r = 10$.

3.3.2. Computing relatedness

To compute the similarity between two terms we proceed as usually done, by determining all possible concepts the terms may refer to, and then choosing the pair with the highest similarity score. The similarity score is computed by taking all paths between the corresponding starting nodes into account. To develop the similarity measure we used the 30 pairs in the Miller and Charles ($M\&C$) set. We experimented with various formulas, starting from those developed for WordNet. The formula that gave best results is presented, justified and evaluated in the remainder of the section. We make no claim that this is the best measure; it should rather be considered a lower bound. To compute the similarity, nodes closest to the lower level of the hierarchy (and thus more specific) and those closest to the edges of the path have a higher value. The value decreases as we go higher up in the hierarchy, and further from the starting nodes. A path's value is lower the longer it is and the higher up in the hierarchy its nodes are.

$$sim(n_x, n_y) = \frac{\sum_{path_{(n_x,n_y)} \in Paths(n_x,n_y)} value(path_{(n_x,n_y)})}{|Paths(n_x, n_y)|}$$

$Paths(n_x, n_y) = \{path(n_x, n_y)\}$ is the set of paths from n_x to n_y, and $value(path_{(n_x,n_y)})$ is the value of the path, computed as follows:

$$value(path_{(n_x,n_y)}) = \frac{1}{1 + \sum_{n_i \in path_{(n_x,n_y)}} value(n_i, path_{(n_x,n_y)})}$$

where $value(n_i, path_{(n_x,n_y)})$ is the value of node n_i on path $path_{(n_x,n_y)}$:

$$value(n_i, path_{(n_x,n_y)}) = \frac{ic(n_i)}{D - depth(n_i) + level(n_i, path_{(n_x,n_y)}) + 1}$$

D is the maximum depth of the category hierarchy, $depth(n_i)$ is the depth of the node n_i in this hierarchy, $level(n_i, path_{(n_x,n_y)})$ is the distance from n_i to the closest of n_x or n_y, and $ic(n_i)$ is an approximation of the information content for node n_i, computed in the same manner as done by Ponzetto and Strube [36]:

$$ic(n_i) = 1 - \frac{\log(hypo(n_i) + 1)}{\log(|Wkn|)}$$

$|Wkn|$ is the number of nodes in WikiNet, and $hypo(n_i)$ is the number of hyponyms of node n_i, computed as the number of pages subsumed by n_i in Wikipedia's category network.

In using similarity/relatedness measures between word pairs it is assumed that choosing the highest score leads to the correct disambiguation of the words in the pair (e.g. [4]). In previous work on evaluating WordNet, Wikipedia or other resources relative to similarity or relatedness computations, the issue of disambiguation has been disregarded, and the focus has been on obtaining good correlation scores with the human judges. In our case the assumption that words in a pair disambiguate each other to the correct senses does not hold. WikiNet is highly interconnected and contains numerous lexicalization variations for the included concepts. Because of this, all measures overestimate similarity/relatedness, because for most term pairs there are senses that are close or closely connected in the network. To evaluate the properties of the network built we separate the two tasks – disambiguating the concepts, and evaluating the similarity/relatedness measure against the manually assigned scores. Two judges have manually assigned concept IDs to the words in the Miller and Charles ($M\&C$) and the Rubenstein and Goodenough ($R\&G$) lists. More than one ID was allowed, and only IDs that had associated the particular term as a lexicalization were allowed as options. This caused some problems for the word journey, for example, whose best sense (as trip or voyage) cannot be retrieved from the information extracted from the Wikipedia dump: *Journey* is a disambiguation page, and while the first sentence is very helpful for the human reader (A **journey** is a trip or voyage), it does not contain a hyperlink to either of these concepts.

The disambiguated versions produced by the two human judges for each of $M\&C$ and $R\&G$ were intersected. The fact that for both data sets each word had at least one sense (ID) upon which the annotators agreed allows us to directly use the intersective annotations, without further need for adjudication. The results on the gold standards thus obtained and on the non-annotated data are presented in Table 8.

One of the strengths of the resource is its multilinguality. To test this feature we used the German version of the $R\&G$ lists – G [12]. We had fewer German lexicalizations of concepts than English ones; this reduced the number of possible ambiguities, which is reflected in the relatively high scores for this data set even without disambiguating annotations. Eight words from the G data did not appear in the multilingual index: Grinsen and Mittagsstunde (do not appear in the German Wikipedia), Irrenhaus had no overlap through cross-language links, and Schnur, Bursche, Fahrt do not have an appropriate sense. This leaves 45 pairs (out of the original 65) for which both terms had concept correspondents in WikiNet, and 51 pairs when manually disambiguated and the judges assigned a sense to Irrenhaus.

Table 8
Evaluation of similarity on the Miller and Charles, Rubenstein and Goodenough and Gurevych sets.

Resource	Dataset		Pearson	Spearman
WikiNet	*M&C*	disambig	0.86	0.83
		raw	0.59	0.58
WikiTaxonomy		raw	0.87	0.79
WordNet		raw	0.86	0.86
WikiNet	*R&G*	disambig	0.70	0.67
		raw	0.66	0.64
WikiTaxonomy		raw	0.78	0.75
WordNet		raw	0.88	0.86
WikiNet	*G*	disambig	0.72	0.72
		raw	0.62	0.63
WikiTaxonomy		raw	0.69	–
GermaNet		raw	0.76	–

The best previous results on semantic similarity/relatedness based on Wikipedia were reported on WikiTaxonomy by Ponzetto and Strube [37]. Using a taxonomy built based on the category network, they report highest results for Resnik's information content-based measure – 0.87 (Pearson) and 0.79 (Spearman) for *M&C* and 0.78 (Pearson) and 0.75 (Spearman) for *R&G*, computed without manual disambiguation on a 2008 version of Wikipedia with 337 741 categories and 2 276 274 articles. Since then the resource has grown, and it is now bigger, more connected and has more ambiguities. A direct comparison based on the 2009/07/13 English Wikipedia dump was not possible. The best results based on WordNet for *R&G* are 0.8614 (Spearman) and 0.876 (Pearson), and for *M&C* 0.856 (Spearman) and 0.864 (Pearson) [46].

We tested WikiNet on the German version of the Rubenstein and Goodenough word pairs. On this data set Ponzetto and Strube [35] report 0.69 Pearson correlation on WikiTaxonomy, while the best score on GermaNet – the German version of WordNet – was 0.76 [12]. The reported scores were computed on the pairs for which both terms had an entry in Wikipedia.

WikiNet's high scores on the manually disambiguated data is additional (to the intrinsic evaluation) proof of the quality of its underlying structure. On the other hand, the relatively low scores on non-disambiguated data – the more realistic setting in which WordNet is the clear winner – shows that more information is needed to disambiguate terms. The semantic similarity task between two words out of context is artificial, as it assumes that the two words disambiguate each other, and the correct senses are the ones closest according to the similarity/relatedness metric. This, however, does not hold in a resource like WikiNet, with highly interconnected concepts. To better assess WikiNet we use it in an NLP task, which we describe in the following section.

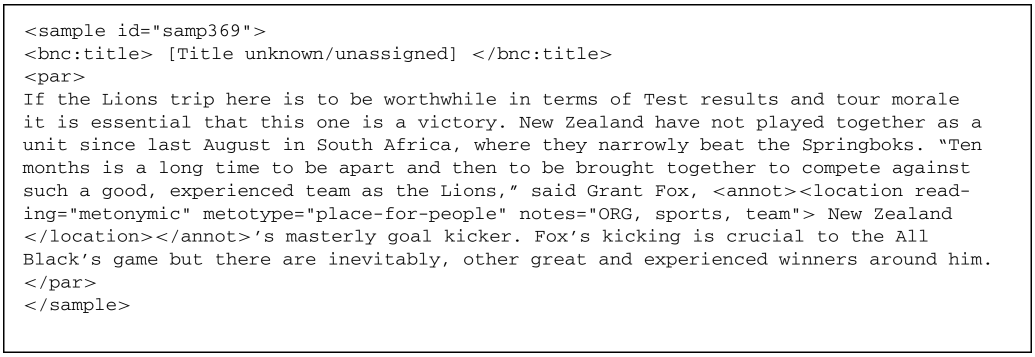
3.4. *Metonymy resolution*

Metonymies are figures of speech whereby the speaker is "using one entity to refer to another that is related to it." [15] – e.g. in the news article sentence 'Buckingham Palace announced at 8am on Friday that the Queen had bestowed the title of Duke of Cambridge on her grandson.', Buckingham Palace stands for the representative of the Queen whose official residence is the palace.

The task of metonymy resolution implies identifying the correct interpretation of a term in context. For example, the interpretation of the term New Zealand in the text fragment shown in Fig. 8 is not literal – as the country New Zealand – rather it stands in for a sports team representing the country in a sporting event.

The most common view of metonymies is that they violate semantic constraints in their immediate context. To resolve metonymies one must detect violated constraints, usually from those imposed by the verbs on their arguments [8,13,40]. Most of the recent work on metonymy resolution (for an overview see [21]) relies on syntactic clues from the local sentential context of a potentially metonymic word (PMW) to determine which interpretation (reading) is most appropriate. In the example from Fig. 8, the noun kicker suggests a `place-for-people` reading, because sports teams have kickers, rather than countries. Previously, Markert and Hahn [19] have shown that global context is useful in detecting metonymies that do not violate selectional restrictions. In this case one can use referential cohesion relations. At the time comprehensive knowledge bases were not available to help establish such relations. We test the usefulness of WikiNet for this task.

We use the data for the metonymy task at SemEval 2007 [21], in which the PMW are names of countries and companies. The data comes partitioned into training and testing (for country PMWs 925 training and 908 testing instances, for company PMWs 1090/842). As shown in Fig. 8, there is a larger context surrounding the PMW than just the corresponding sentence. We explore this for global constraints on the interpretation of the target word. The starting point is the system described in [28], which uses, as other work does, selectional preferences imposed by the grammatical context to decide on the correct interpretation.



```
<sample id="samp369">
<bnc:title> [Title unknown/unassigned] </bnc:title>
<par>
If the Lions trip here is to be worthwhile in terms of Test results and tour morale
it is essential that this one is a victory. New Zealand have not played together as a
unit since last August in South Africa, where they narrowly beat the Springboks. "Ten
months is a long time to be apart and then to be brought together to compete against
such a good, experienced team as the Lions," said Grant Fox, <annot><location read-
ing="metonymic" metotype="place-for-people" notes="ORG, sports, team"> New Zealand
</location></annot>'s masterly goal kicker. Fox's kicking is crucial to the All
Black's game but there are inevitably, other great and experienced winners around him.
</par>
</sample>
```

Fig. 8. Sample data from the metonymy resolution task at SemEval 2007.

To incorporate encyclopedic knowledge, we first identify concepts in the given paragraph surrounding a PMW using the existing lexicalizations in WikiNet. Solving ambiguities is done implicitly (albeit not necessarily very accurately) in the processing steps presented in Algorithm 5.

Algorithm 5 Extracting concepts for metonymy resolution.

```
Input:
    $T = \{t_i\}$ – the set of texts given for each instance in the data $WikiNet$
    $\nu$ – a threshold for related concepts frequency within a paragraph
    $\tau$ – a threshold for concept (feature) frequency in the data set
Output:
    $C$ – a list of concepts

$C = \{\}$
for $t_i \in T$ do
    $C_i = \{\}$
    for $c_k \in WikiNet$ do
        $C_{ik} = \{\}$
        if $c_k$ appears in $t_i$ then
            $C_{ik} = C_{ik} \cup \{c_k\} \cup \{c_j \mid (c_k, R, c_j) \in WikiNet\}$
        $C_i = C_i \cup \{c_j \mid c_j$ appears in at least $\nu C_{ik}\}$
    $C = C \cup \{c_j \mid c_j$ appears in at least $\tau C_i\}$
return $C$
```

The purpose of the ν threshold is to compensate for the rather simple disambiguation algorithm: If a concept is related to at least ν other (potential) concepts in the given context, it is more probable that it is an appropriate concept for this context. For the experiments described later in the paper $\nu = 3$ – it is the smallest value that counters possible noise in connectivity in WikiNet and cohesiveness of the context (in terms of concepts). The τ threshold is used to filter features for the machine learning stage. Features infrequent across instances would cause overfitting. We arbitrarily choose $\tau = 5$.

We consider the identified concepts as global context features, and added them to the features described in [28]: the syntactic features proposed in [20] and selectional preference features.

The features described above are used to represent each instance in the data. We used the Weka implementation of Support Vector Machine (SMO) [49] (default settings) to build a model for each metonymy type. The task had three settings: coarse – distinguish between literal and non-literal interpretations; medium – distinguish between literal, metonymic or mixed interpretations; fine – distinguish between all possible interpretations (a small number or prespecified possible interpretations) for the potentially metonymic word.

Table 9 shows the (non-zero) results obtained, in all three task settings, for the two types of PMWs – locations and organizations. *Base* shows the class distribution, the *Selectional preferences* columns show the results obtained using selectional preferences (described in detail in [28]), and the *+ WikiNet information* shows the results obtained by adding the concepts identified in the paragraph surrounding the PMW to the selectional preference features. Including the encyclopedic knowledge from WikiNet shows consistent improvement in terms of precision and recall, especially for the non-literal interpretations. Because of the relatively small number of instances with a non-literal interpretation, the improvement does not appear to be statistically significant.

The evaluation shows that global context is useful in interpreting PMWs despite the simple concept identification approach used.

4. WikiNet in context

The work described in this paper fits within the rather vast and intensively explored areas of knowledge acquisition and ontology induction. Unlike the majority of other such work which involves Wikipedia, our approach does not tie Wikipedia

Please cite this article in press as: V. Nastase, M. Strube, Transforming Wikipedia into a large scale multilingual concept network, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.06.008

Table 9
Accuracy (acc), precision (f), recall (r) and F-score (f) for detecting metonymic interpretations.

Task	Method									
	Base		Selectional preferences				+ WikiNet information			
	acc	f	acc	p	r	f	acc	p	r	f
LOCATION – coarse										
	79.4		85.6				86.2			
literal		79.4		87.9	94.9	91.3		88.1	95.6	91.7
non-literal		20.6		71.5	49.7	58.7		74.6	50.3	60.1
LOCATION – medium										
	79.4		85.2				85.9			
literal		79.4		87.9	94.9	91.3		88.1	95.6	91.7
metonymic		18.4		70.1	53.3	60.5		73.6	53.3	61.8
mixed		2.2		33.3	5.0	8.7		40.0	10.0	16.0
LOCATION – fine										
	79.4		84.4				85.1			
literal		79.4		87.9	94.9	91.3		88.1	95.6	91.7
place-for-people		15.5		64.3	57.4	60.7		67.5	57.4	62.1
mixed		2.2		33.3	5.0	8.7		40.0	10.0	16.0
ORGANIZATION – coarse										
	61.8		74.2				75.4			
literal		61.8		74.5	88.7	80.9		75.4	89.2	81.8
non-literal		38.2		73.5	50.9	60.2		75.3	53.1	62.3
ORGANIZATION – medium										
	61.8		71.7				73.3			
literal		61.8		74.5	88.7	80.9		75.4	89.2	81.8
metonymic		31.0		65.8	50.2	56.9		68.4	54.0	60.4
mixed		7.2		50.0	19.7	28.2		57.1	19.7	29.3
ORGANIZATION – fine										
	61.8		70.3				71.7			
literal		61.8		74.5	88.7	80.9		75.4	89.2	81.8
org-for-members		19.1		59.3	55.3	57.2		61.8	58.4	60.1
org-for-product		8.0		68.8	32.8	44.4		66.7	35.8	46.6
org-for-facility		2.0		60.0	18.8	28.6		83.3	31.3	45.5
org-for-name		0.7		45.5	83.3	58.8		50.0	83.3	62.5
mixed		7.2		50.0	20.0	28.6		57.1	20.0	29.6

to additional resources, but is instead focused on generating a self-contained, easily renewable resource. We review the most closely related work in decreasing order of the similarity of either the approach or resource generated.

Ponzetto and Strube [35,37] build on the category network from Wikipedia and induce *is_a* links based on several criteria: head matching, modifier matching, structural analysis of shared categories and pages between two linked categories, and patterns indicative of *is_a* relations and *not_is_a* relations. The result is WikiTaxonomy, with 208 208 *is_a* relations, evaluated at 84.0% F-score.

DBpedia[23] [1,3] converts Wikipedia's content into structured knowledge using information from Wikipedia's relational database tables and the structured information in infoboxes, and connects it to a variety of other resources, such as Freebase, WordNet, OpenCyc and more. The DBpedia dataset covers approximately 3.64 million entities, 1.83 million of which are classified into a shallow ontology consisting of 170 classes.[24] This ontology was built by manually organizing the most frequently used infobox templates from Wikipedia into a hierarchy. A distinguishing feature of DBpedia is live extraction: Its database is continuously updated whenever a Wikipedia article is changed. DBpedia offers a web-based interface to its database.

YAGO [47] also extracts information from Wikipedia, and links it to GeoNames and WordNet's hierarchies to take advantage of WordNet's manually produced taxonomy. In addition to structured information in Wikipedia, facts representing relations of 100 types are extracted from specific categories that provide relational information, such as 1975 BIRTHS, categories starting with Countries in ..., Rivers of ..., Attractions in .... These are used as a source of the following relations: *bornInYear, diedInYear, establishedIn, locatedIn, writtenInYear, politicianOf, hasWonPrize*. Accuracy is estimated based on a small sample of manually annotated relation instances out of the approximately 10 million extracted, and lies between 90.84 ± 4.28% and 98.72 ± 1.30%. YAGO's core knowledge base (derived from Wikipedia) contains approximately 11 million relations, while the full system covers 460 million. Distinguishing characteristics of YAGO are its accompanying search engine NAGA, and SOFIE, a logical reasoning mechanism for expanding YAGO with information from open texts and for checking internal consistency of the resource when adding external facts.

[23] http://dbpedia.org.

[24] December 2011.

MENTA [6] is a multilingual extension of the core YAGO knowledge base. It extracts entities from Wikipedia versions in all available languages, linking them through cross-language links, category–category and category–article links, and links them to WordNet. Evaluated on random samples between 104 and 322 instances, the accuracy of subclass and instance-of links to WordNet ranges between 83.38% and 92.30%. The resource covers 5.4 million entities.

BabelNet [30] is an automatically built resource that integrates WordNet and Wikipedia. It is obtained by disambiguating Wikipedia articles relative to WordNet senses, by adding different language information using the corresponding cross-language links, and by adding synonyms using redirect links and selected output from a machine translation system. The current version includes lexicalizations of existing WordNet synsets in 6 languages. In version 1.0.1, BabelNet contains 83 156 merged Wikipedia/WordNet entities, and provides 2 955 552 additional multilingual "babel-synsets". Because of the Wikipedia–WordNet mapping, BabelNet can take advantage of WordNet's clean hierarchy.

The largest ongoing project for acquiring knowledge from general texts is the Machine Reading project at the University of Washington [39]. One of the goals of the project is to induce a large scale ontology, which is now pursued through multiple interacting threads: rely on Wikipedia for annotated data from which to learn models for acquiring relations from open text [2,51,52], jointly perform knowledge extraction, ontology induction and population via recursive relational clustering [38]. Kylin [50] uses the existing infoboxes and their corresponding articles as sources of training data, to learn how to fill in infobox templates for articles which do not have such structured information. For four concepts, Wu and Weld [50] obtain precision between 73.9% and 97.3%, and recall between 60.5% and 95.9%. Read the Web[25] is a Never-Ending Language Learning (NELL) project at Carnegie Mellon University [16]. It has an iterative approach of adding more and more facts extracted from open text. As of yet, its results are not at the level of the Machine Reading project, with 978 383 facts extracted.

Nguyen et al. [32] 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.

Snow et al. [44] present an approach for automatically inducing semantic taxonomies, which relies on combining evidence from heterogeneous relations to derive the structure with the highest probability. This approach can be used to build a taxonomy from scratch, or to expand upon an existing one, such as WordNet. This approach is based on representing word pairs through lexico-syntactic patterns that capture how the pair is connected in sentences in the corpus. Because it deals with open texts, the method must address the sense disambiguation task as well. For 10 000 added hyponyms, precision was 84% (evaluated on a random sample of 100 pairs), for 20 000 precision was 68% (also on a sample of 100 pairs).

Navigli et al. [31] also investigate the automatic induction of semantic taxonomies from a given corpus. The process involves first obtaining a terminology from the corpus, which is then used iteratively to obtain taxonomic relations and new potential terms from the texts. In the final step, the graph is trimmed based on connectivity information and the restrictions imposed by taxonomic relations.

Freebase[26] is a large online collaborative knowledge base, originally populated with information from Wikipedia, MusicBrainz, and other online resources, open for editing by human contributors. Compared to DBpedia and YAGO, Freebase is structured as a graph, with entities as nodes and edges representing the links between them. It consists of approximately 22 million entities (called topics in Freebase),[27] which are grouped into types (e.g. people, places), and which can have associated attributes (e.g. birth date). While it does have multilingual content, there are no explicit links between the same concept in different languages.

ProBase[28] [53] is a probabilistic taxonomy automatically built from a corpus of 1.68 billion web pages and uses two years' worth of search log data for filtering concepts. Taxonomy induction starts with Hearst patterns for detecting *is_a* relations in the corpus, which are used to detect the concepts, distinguish their senses, and establish *is_a* links between them. ProBase contains 2 653 872 concepts, 16 218 369 distinct concept–instance pairs, and 4 539 176 distinct concept–subconcept pairs (20 757 545 *is_a* pairs in total).

Related to our work on deriving class attributes, Paşca [34] processes search engine queries to obtain similar information. 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 the class *drugs*, or *wing span* for the 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 labels on a sample of extracted attributes for the 40 classes considered range from 90% precision for 10 attributes to 76% for 50.

Compared to WikiTaxonomy, DBpedia, MENTA, and BabelNet, in building WikiNet we have revealed implicit knowledge in Wikipedia, and used it to derive novel connections between entities in Wikipedia. YAGO processes a limited number of specific name patterns to derive novel information, although not on the same scale as WikiNet. We also worked under the assumption that each manually generated resource (such as Wikipedia, WordNet, OpenCyc) was built according to its own principles and captures different aspects of linguistic or encyclopedic knowledge. Because of this, we aimed for WikiNet to be self-contained, and endowed it with the type of knowledge that research in the natural language processing field has shown to be useful and desirable (such as a taxonomy backbone).

[25] http://rtw.ml.cmu.edu/rtw/.

[26] http://www.freebase.com.

[27] December 2011.

[28] http://research.microsoft.com/en-us/projects/probase.

Deriving a resource like WikiNet has both advantages and disadvantages. The disadvantage is that, while it can serve as a starting point for further development, WikiNet is limited to the information included in Wikipedia. There are, however, multiple advantages: (i) we do not need to rely on external resources for the set of entities to be extracted or to validate the type of relations that are extracted; (ii) the structure arises through the building process, and combines the "folksonomy" that is the backbone of Wikipedia with a structure that arises from the knowledge decoded from the category names; (iii) it is clear how the extracted relations fit in the network and we do not have the problem of disambiguating terms extracted from text to their corresponding concepts; (iv) multilinguality. Thanks to the connection between languages in WikiNet, information easily obtained through one language can be ported to others. An example of this is the named entity information, which based on the article texts and the conventions of capitalization in English can be obtained with high accuracy, whereas for German this would not be as easy, because nouns in German are capitalized.

The work presented here was developed based on observations about the types of information that appear in Wikipedia specifically, and the way those types of information interact. However, some of the lessons learned are more general. We have learned how knowledge can be propagated throughout a network by connecting different pieces of individual information. This provides interesting avenues for further knowledge acquisition: A novel entity discovered in open text can first be classified under existing categories in Wikipedia, which then triggers the addition of links to the rest of the network through the relations implicit in the newly assigned parent categories. We have also learned that knowledge may come from unconventional, and probably "unintentional", sources. In Wikipedia's case, these sources are the category names (there may be others as well). Such situations may arise in processing other repositories of (manually built) information where knowledge is implicitly coded, for example, in directory names in a file hierarchy or in class names in manual annotations.

A question that we cannot yet answer is what is the impact of the structure – which is derived exclusively from the English Wikipedia – when applying the resource to NLP tasks in different languages, as we do not know to what extent ontologies are language specific. We can only speculate that encyclopedic knowledge, the kind reorganized into WikiNet, is less prone to cultural biases.

5. Conclusions

This paper described the construction of a multilingual, large scale resource, based on exploiting several facets – some obvious, some less so – of Wikipedia. Compared to related work in the domain of knowledge acquisition/ontology induction, the approach presented and the resource built has both advantages and disadvantages. Its main advantages are the rapid derivation of a large multilingual resource, easy to regenerate on new versions of Wikipedia. It can serve both English and languages poorer in resources, either as a stable resource or as a starting point for ontology population. Compared to WordNet, WikiNet has high coverage of named entities, and named entity lexicalizations in various languages are a useful resource for machine translation. Another advantage is the ability to port information – such as named entity – from one language to another. The main disadvantage is that WikiNet is not complete, and in the long run it is only a seed – although a large one – for further knowledge acquisition.

We plan to explore methods for enriching WikiNet with information extracted from open text, following an approach similar to the one for propagating information from infoboxes, by first linking (classifying) a novel entity to Wikipedia's categories, and then linking it to the rest of the network through the relations implied in its newly assigned parent categories. Future work plans also include further exploring WikiNet's use for different tasks, including coreference resolution and text alignment. Identifying common concepts in different language texts should arise naturally from WikiNet's structure. The resource is freely available for download,[29] together with a tool kit – WikiNetTK[30] – for visualization and various methods to facilitate embedding the world knowledge encoded in WikiNet into applications [14]. The scripts used to build WikiNet are also available at the WikiNet download site.

Acknowledgements

We thank the reviewers for their detailed and constructive comments, and the Klaus Tschira Foundation, Heidelberg for financial support. This work has also been partially supported by the EC-funded project CoSyne (FP7-ICT-4-24853).

References

[1] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, Z. Ives, 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, pp. 722–735.

[2] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, Open information extraction from the Web, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, 6–12 January 2007, pp. 2670–2676.

[3] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia – A crystallization point for the Web of data, Journal of Web Semantics 7 (2009) 154–165.

[4] A. Budanitsky, G. Hirst, Evaluating WordNet-based measures of semantic distance, Computational Linguistics 32 (1) (2006) 13–47.

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

[30] http://sourceforge.net/projects/wikinettk/.

[5] R. Bunescu, M. Paşca, Using encyclopedic knowledge for named entity disambiguation, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006, pp. 9–16.
[6] G. de Melo, G. Weikum, MENTA: Inducing multilingual taxonomies from Wikipedia, in: Proceedings of the ACM 19th Conference on Information and Knowledge Management (CIKM 2010), Toronto, Ont., Canada, 26–30 October 2010, pp. 1099–1108.
[7] G. de Melo, G. Weikum, Untangling the cross-lingual link structure of Wikipedia, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 844–853.
[8] D.C. Fass, met*: A method for discriminating metonomy and metaphor by computer, Computational Linguistics 17 (1) (1991) 49–90.
[9] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998.
[10] E. Gabrilovich, S. Markovitch, 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, pp. 1606–1611.
[11] E. Gabrilovich, S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, Journal of Artificial Intelligence Research 34 (2009) 443–498.
[12] I. Gurevych, H. Niederlich, Accessing GermaNet data and computing semantic relatedness, in: Companion Volume to the Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Mich., 25–30 June 2005, pp. 5–8.
[13] J.R. Hobbs, M.E. Stickel, D.E. Appelt, P. Martin, Interpretation as abduction, Artificial Intelligence 63 (1993) 69–142.
[14] A. Judea, V. Nastase, M. Strube, WikiNetTk – A tool kit for embedding world knowledge in NLP applications, in: Proceedings of the IJCNLP 2011 System Demonstrations, Chiang Mai, Thailand, 9 November 2011, pp. 1–4.
[15] G. Lakoff, M. Johnson, Metaphors We Live By, University of Chicago Press, Chicago, IL, 1980.
[16] N. Lao, T.M. Mitchell, W.W. Cohen, Random walk inference and learning in a large scale knowledge base, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–29 July 2011, pp. 529–539.
[17] M. Lauer, Designing statistical language learners: Experiments on noun compounds, Ph.D. thesis, Macquarie University, Dept. of Computing, Sydney, Australia, 1995.
[18] D.B. Lenat, R. Guha, K. Pittman, D. Pratt, M. Shepherd, Cyc: Towards programs with common sense, Communications of the ACM 33 (8) (1990) 30–49.
[19] K. Markert, U. Hahn, Metonymies in discourse, Artificial Intelligence 135 (1/2) (2002) 145–198.
[20] K. Markert, M. Nissim, Comparing knowledge sources for nominal anaphora resolution, Computational Linguistics 31 (3) (2005) 367–401.
[21] K. Markert, M. Nissim, SemEval-2007 Task 08: Metonymy resolution at SemEval-2007, in: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1), Prague, Czech Republic, 23–24 June 2007, pp. 36–41.
[22] O. Medelyan, D. Milne, C. Legg, I.H. Witten, Mining meaning from Wikipedia, International Journal of Human–Computer Interaction 67 (9) (2009) 716–754.
[23] A. Meyers, C. Macleod, R. Yangarber, R. Grishman, L. Barrett, R. Reeves, Using NOMLEX to produce nominalization patterns for information extraction, in: Proceedings of the COLING-ACL '98 Workshop on The Computational Treatment of Nominals, Montréal, Québec, Canada, 16 August 1998, pp. 25–32.
[24] G.A. Miller, W.G. Charles, Contextual correlates of semantic similarity, Language and Cognitive Processes 6 (1) (1991) 1–28.
[25] D. Milne, I.H. Witten, 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, pp. 25–30.
[26] D. Milne, I.H. Witten, 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, pp. 1046–1055.
[27] V. Nastase, M. Strube, Decoding Wikipedia category names for knowledge acquisition, in: Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence, Chicago, IL, 13–17 July 2008, pp. 219–1224.
[28] V. Nastase, M. Strube, Combining collocations, lexical and encyclopedic knowledge for metonymy resolution, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009, pp. 910–918.
[29] V. Nastase, M. Strube, B. Börschinger, C. Zirn, A. Elghafari, WikiNet: A very large scale multi-lingual concept network, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, La Valetta, Malta, 17–23 May 2010.
[30] R. Navigli, S.P. Ponzetto, BabelNet: Building a very large multilingual semantic network, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 216–225.
[31] R. Navigli, P. Velardi, S. Faralli, A graph-based algorithm for inducing lexical taxonomies from scratch, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, 19–22 July 2011, pp. 1872–1877.
[32] D.P. Nguyen, Y. Matsuo, M. Ishizuka, 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, pp. 1414–1420.
[33] I. Niles, A. Pease, Towards a standard upper ontology, in: Proceedings of the International Conference on Formal Ontology in Information Systems, Ogunquit, Maine, 17–19 October 2001, pp. 2–9.
[34] M. Paşca, Organizing and searching the World Wide Web of facts – Step two: Harnessing the wisdom of the crowds, in: Proceedings of the 16th World Wide Web Conference, Banff, Canada, 8–12 May 2007, pp. 101–110.
[35] S.P. Ponzetto, M. Strube, 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, pp. 1440–1445.
[36] S.P. Ponzetto, M. Strube, Knowledge derived from Wikipedia for computing semantic relatedness, Journal of Artificial Intelligence Research 30 (2007) 181–212.
[37] S.P. Ponzetto, M. Strube, Taxonomy induction based on a collaboratively built knowledge repository, Artificial Intelligence 175 (9/10) (2011) 1737–1756.
[38] H. Poon, J. Christensen, P. Domingos, O. Etzioni, R. Hoffmann, C. Kiddon, T. Lin, X. Ling, Mausam, A. Ritter, S. Schoenmackers, S. Soderland, D. Weld, F. Wu, C. Zhang, Machine reading at the University of Washington, in: Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, Los Angeles, CA, 6 June 2010, pp. 87–95.
[39] H. Poon, P. Domingos, Unsupervised ontology induction from text, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 296–305.
[40] J. Pustejovsky, The generative lexicon, Computational Linguistics 17 (4) (1991) 209–241.
[41] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montréal, Canada, 20–25 August 1995, vol. 1, 1995, pp. 448–453.
[42] H. Rubenstein, J.B. Goodenough, Contextual correlates of synonymy, Communications of the ACM 8 (10) (1965) 627–633.
[43] B. Santorini, Part of speech tagging guidelines for the Penn Treebank Project, http://www.cis.upenn.edu/~treebank/home.html, 1990.
[44] R. Snow, D. Jurafsky, A.Y. Ng, Semantic taxonomy induction from heterogeneous evidence, in: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–21 July 2006, pp. 801–808.
[45] F. Suchanek, G. Kasneci, G. Weikum, YAGO: A large ontology from Wikipedia and WordNet, Elsevier Journal of Web Semantics 6 (3) (2008) 203–217.
[46] G. Tsatsaronis, I. Varlamis, M. Vazirgiannis, Text relatedness based on a word thesaurus, Journal of Artificial Intelligence Research 37 (2010) 1–39.
[47] G. Weikum, G. Kasneci, M. Ramanath, F. Suchanek, Database and information-retrieval methods for knowledge discovery, Communications of the ACM 52 (4) (2009) 56–64.
[48] W. Wentland, J. Knopp, C. Silberer, M. Hartung, 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.

[49] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition, Morgan Kaufmann, San Francisco, CA, 2005.

[50] F. Wu, D. Weld, Automatically semantifying Wikipedia, in: Proceedings of the ACM 16th Conference on Information and Knowledge Management (CIKM 2007), Lisbon, Portugal, 6–9 November 2007, pp. 41–50.

[51] F. Wu, D. Weld, Automatically refining the Wikipedia infobox ontology, in: Proceedings of the 17th World Wide Web Conference, Beijing, China, 21–25 April 2008.

[52] F. Wu, D. Weld, Open information extraction using Wikipedia, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 118–127.

[53] W. Wu, H. Li, H. Wang, K.Q. Zhu, Towards a probabilistic taxonomy of many concepts, Tech. Rep. MSR-TR-2011-25, Microsoft Research Asia, 2011.

Please cite this article in press as: V. Nastase, M. Strube, Transforming Wikipedia into a large scale multilingual concept network, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.06.008