Artificial Intelligence 175 (2011) 1737–1756

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Taxonomy induction based on a collaboratively built knowledge repository

Simone Paolo Ponzetto [a,*], Michael Strube [b]

[a] Institut für Computerlinguistik, Heidelberg University, Heidelberg, Germany
[b] Heidelberg Institute for Theoretical Studies gGmbH, Natural Language Processing Group, Heidelberg, Germany

ARTICLE INFO

Article history:
Received 23 August 2010
Received in revised form 6 January 2011
Accepted 10 January 2011
Available online 12 January 2011

Keywords:
Natural language processing
Knowledge acquisition
Lexical semantics

ABSTRACT

The category system in Wikipedia can be taken as a conceptual network. We label the semantic relations between categories using methods based on connectivity in the network and lexico-syntactic matching. The result is a large scale taxonomy. For evaluation we propose a method which (1) manually determines the quality of our taxonomy, and (2) automatically compares its coverage with ResearchCyc, one of the largest manually created ontologies, and the lexical database WordNet. Additionally, we perform an extrinsic evaluation by computing semantic similarity between words in benchmarking datasets. The results show that the taxonomy compares favorably in quality and coverage with broad-coverage manually created resources.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Research in Artificial Intelligence (AI) has made tremendous progress in the last decades by employing data-driven techniques for solving tasks of ever increasing difficulty. However, working on knowledge-intensive applications such as e.g. semantic web technologies [8] and question answering engines [58] calls for complementing statistical methods with semantically rich representations based on world and encyclopedic knowledge, thus bringing the *knowledge acquisition bottleneck* problem into focus yet again.

While the need for wide-coverage knowledge bases when simulating human intelligence has been acknowledged since McCarthy's seminal work [59], the manual creation of resources such as Cyc [50] has been shown to scale poorly [99]. In addition, most of the existing knowledge resources are domain dependent or have limited and arbitrary coverage. The field of ontology learning deals with these problems by taking textual input and transforming it into a taxonomy or a proper ontology. However, such learned ontologies are small and mostly domain dependent, and evaluations have revealed rather poor performance of the methods (see [14] for an extensive overview).

We try to overcome these problems with a novel perspective[1] by utilizing Wikipedia, a wide coverage collaboratively generated encyclopedia. Our previous work on using the category network suggests that Wikipedia categories can be considered a semantic network in its own right [104,85]. Unfortunately, Wikipedia categories do not form a proper knowledge base with a full-fledged subsumption hierarchy, but only a thematically organized thesaurus. The lack of clear semantic relations between the categories poses a serious limitation to the amount of information provided. In this work we develop the idea of using the Wikipedia categorization system as a semantic network a step further and present methods for generating a large scale taxonomy by automatically assigning *isa* and *notisa* labels to the relations between categories. We use methods

* Corresponding author.
E-mail addresses: ponzetto@cl.uni-heidelberg.de (S.P. Ponzetto), michael.strube@h-its.org (M. Strube).

[1] This article builds upon and expands on [84]. The resource described in this paper is freely available under a GFDL license at http://www.h-its.org/nlp/download/wikitaxonomy.php (see Appendix C for details).

0004-3702/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.artint.2011.01.003

based on connectivity in the network and lexico-syntactic patterns to label the relations between categories. Based on these methods, we are able to derive a large scale taxonomy. The main contributions of this article are as follows:

1. We propose to derive a taxonomy from the system of categories in Wikipedia. This amounts to transforming the Wikipedia categorization system into a full-fledged subsumption hierarchy such as the one found in Cyc [50] and WordNet [28].
2. We develop a set of lightweight heuristics to automatically distinguish *isa* and *notisa* relations between the categories in Wikipedia. Our method works by capturing linguistic regularities in category labels (syntax-based methods, Section 2.3), exploiting naming conventions and connectivity in the graph (connectivity-based methods, Section 2.4), as well as mining large corpora for patterns expressing semantic relations (lexico-syntactic based methods, Section 2.5). The result is a large scale taxonomy including 335,128 semantic links.
3. We perform an evaluation which (1) determines the quality of our taxonomy based on human assessment, and (2) automatically compares its coverage with ResearchCyc and WordNet, arguably two of the largest manually annotated knowledge bases. For the manual evaluation we report an F_1 measure of up to 84%. For the automatic evaluation of coverage, we develop a taxonomy mapping method based on the syntactic structure of the Wikipedia category labels. This evaluation shows that there is little overlap in terms of concept relations between our taxonomy and ResearchCyc and WordNet, which indicates that Wikipedia complements those resources. Compared with ResearchCyc our taxonomy provides 28.2% extra coverage, compared with WordNet 211.6%.
4. We extrinsically evaluate the resource by computing the semantic similarity of word pairs on benchmarking datasets and improve our previous results from [104] by a large margin. The results obtained by using the taxonomy for computing semantic similarity are competitive with the best ones from the literature, i.e. up to a Pearson correlation coefficient r of 0.87, and lie near the estimated upper bound for performance for this task.

The remainder of this article is structured as follows: in Section 2 we present our methods for generating a subsumption hierarchy from the network of categories in Wikipedia. In Section 3 we evaluate the automatically generated taxonomy by comparing it with ResearchCyc and WordNet, as well as by computing semantic similarity between words in benchmarking datasets. We finally present related work in Section 4 and conclude with suggestions for future work in Section 5.

2. Methods

Since May 2004 Wikipedia has allowed for structured access by means of *categories*.[2] The categories form a graph which can be taken to represent a conceptual network with unspecified semantic relations [104,85]. In this section we present our methods to derive *isa* and *notisa* relations from these generic links. This allows us to generate a taxonomy from the Wikipedia category graph by performing the following task: for each pair of categories ⟨Subcat, Supercat⟩ where Subcat[3] is categorized into Supercat, decide whether Subcat *isa* Supercat or not. This aims at transforming a graph with unlabeled semantic relations into a semantic network where the links between categories are augmented with *isa* relations.

The Wikipedia category network contains categories which are used to refer either to an entity, e.g. the Microsoft category, or to a property of a set of entities, e.g. Multinational companies. Accordingly, the relation between a category and its super-categories can be either one of subsumption (i.e. a concept-to-concept strict IS-A relation) or instantiation (i.e. an entity-to-concept INSTANCE-OF relation). In this work we do not distinguish categories that are classes from categories that are entities: therefore we use a definition of *isa* which includes both the IS-A and INSTANCE-OF relations. This is similar to the semantics of the subsumption relation found in WordNet prior to version 2.1. Although this is not methodologically adequate [73], it represents a valid step toward generating a taxonomy from the category network. As in the case of WordNet [63], the distinction between classes and instances can be added to the generated taxonomy later [114]. These same considerations also apply when considering the *notisa* relation: although it does not carry any semantics *per-se*, i.e. it simply refers to 'what is not in an isa relation', it allows us to concentrate on generating a core subsumption hierarchy and does not rule out the generation of more specific relations, e.g. *part-of*, *located-in*, etc., at a later stage [69].

The pseudocode of our method is shown in Algorithm 1. We start with the unlabeled category graph found in Wikipedia and remove from it all nodes which refer to categories used for administration of the Wikipedia project (lines 1–6, Section 2.1). We then collect all remaining nodes and edges and build an initial taxonomy graph which assigns a default *notisa* relation to all category pairs (lines 7–12). Finally, given a set of processing components (described in Sections 2.2–2.6), we generate the *isa* relations by performing a cascade of tests on the category pairs which, at each step, have not yet been discovered as being in an *isa* relation (lines 13–18). For each processing component, we collect all edges in the taxonomy graph labeled with a *notisa* relation and test them for an *isa* semantic relation. As a result of the algorithm, the taxonomy graph is returned (line 19, category pairs for which no *isa* relation can be acquired retain the default *notisa* relation).

The order of the processing components is enforced by the size of Wikipedia. We start with lightweight heuristics to filter out the number of categories to be processed by subsequent modules: these generate *isa* relations by analyzing the syntactic

[2] Wikipedia can be downloaded at http://download.wikimedia.org. In our experiments we use the English Wikipedia database dump from March 12, 2008. This includes 2,276,274 articles, 99.1% of which are categorized into 337,741 categories.

[3] We use Sans Serif for words and queries, CAPITALS for Wikipedia pages and Small Caps for Wikipedia categories.

Algorithm 1: The taxonomy acquisition algorithm.

Input: an unlabeled category graph $G_{Wiki} = \langle V_{Wiki}, E_{Wiki} \rangle$
a category cleanup module *cleanup*
a list of processing components $ProcComp = \{ByMatcher, HeadMatcher, ModifierMatcher, InstanceCategorization, RedundantCategorization, PatternFinder, SisterPropagation, TransitivityPropagation\}$

Output: a labeled taxonomy graph $G_{WikiTax} = \langle V_{WikiTax}, E_{WikiTax} \rangle$, $V_{WikiTax} \subseteq V_{Wiki}$, $E_{WikiTax} \subseteq E_{Wiki}$

$G_{Clean} = \langle V_{Clean}, E_{Clean} \rangle \leftarrow \langle V_{Wiki}, E_{Wiki} \rangle$
for each $v \in V_{Clean}$
  if *cleanup*.REMOVE?(v) **then**
    $V_{Clean} \leftarrow V_{Clean} - \{v\}$
    for each $e = (v_1, v_2) \in E_{Clean}$ s.t. $v_1 = v$
      $E_{Clean} \leftarrow E_{Clean} - \{e\}$
$G_{WikiTax} = \langle V_{WikiTax}, E_{WikiTax} \rangle \leftarrow \langle \emptyset, \emptyset \rangle$
for each $v_i \in V_{Clean}$
  $V_{WikiTax} \leftarrow V_{WikiTax} \cup \{v_i\}$
  for each $v_j \in V_{Clean}$, $i \neq j$
    $V_{WikiTax} \leftarrow V_{WikiTax} \cup \{v_j\}$
    $E_{WikiTax} \leftarrow E_{WikiTax} \cup \{(v_i, notisa, v_j)\}$
for each $proc \in ProcComp$
  for each $e = (v_1, t, v_2) \in E_{WikiTax}$
    if t EQUALS? *notisa* **then**
      if *proc*.ISA?(v_1, v_2) **then**
        $E_{WikiTax} \leftarrow E_{WikiTax} - \{e\}$
        $E_{WikiTax} \leftarrow E_{WikiTax} \cup \{(v_1, isa, v_2)\}$
return $G_{WikiTax}$

structure of the category labels – *ByMatcher* (Section 2.2), *HeadMatcher* (Section 2.3.1) and *ModifierMatcher* (Section 2.3.2) – as well as the local connectivity that a category has with its neighbors in the category graph – *InstanceCategorization* (Section 2.4.1) and *RedundantCategorization* (Section 2.4.2). We then continue with more computing intensive methods[4] which aim to acquire taxonomic relations by mining large corpora for occurrences of lexico-syntactic patterns (*PatternFinder*, Section 2.5). Finally, a last set of components propagates the previously discovered *isa* relation based on multiple inheritance (*SisterPropagation*, Section 2.6.1) and transitivity constraints (*TransitivityPropagation*, Section 2.6.2).

2.1. Category network cleanup (1)

We start with the full categorization network consisting of 337,522 category nodes with 734,140 direct links between them. We first clean the network of meta-categories used for encyclopedia management, e.g. the categories under WIKIPEDIA ADMINISTRATION. However, these categories are connected to many content bearing categories, e.g. the content bearing category MICROSOFT is categorized under the meta-category CATEGORIES NAMED AFTER COMPANIES. Therefore, we cannot remove this portion of the graph entirely. Instead, we remove all those nodes whose labels contain any of the following strings: wiki, lists, disambiguation, template, user, portal, categories, articles, pages, redirect, navigational boxes or stub. This leaves 240,760 categories and 515,423 links to be processed.

2.2. Refinement link identification (ByMatcher – 2)

The next preprocessing step includes identifying so-called *refinement links*. Wikipedia users tend to organize many categories using patterns such as Y X and X BY Z (e.g. MILES DAVIS ALBUMS and ALBUMS BY ARTIST). We label these patterns as expressing *is-refined-by* semantic relations between categories. While these links could be assigned a full *isa* semantics, they represent meta-categorization relations, i.e., their sole purpose is to better structure the categorization network. We take all categories containing by in the name and label all links with their subcategories with an *is-refined-by* relation. This labels 126,920 category links and leaves 388,503 relations to be analyzed.

2.3. Syntax-based methods (3)

The first set of methods to label relations between categories as *isa* is based on the string matching of syntactic components of category labels (Fig. 1 contains a glossary of the computational linguistics terminology required to understand these methods).

[4] The entire taxonomy generation process requires approximately 1000 CPU hours on a 2 GHz Opteron four processor (dual-core) server with 2 GB memory where 99.7% of the overall runtime is required by the lexico-syntactic based methods. In practice, the absolute runtime can be drastically reduced by making use of parallelization optionally coupled with an appropriate text indexing strategy.

Lemma: The lemma is the canonical form of a word, e.g. the infinitive of inflected words or the singular of nouns. A **lemmatizer** automatically determines the lemma for a word.

Stem: The stem (or the root) is the part of the word which does not change when the word is inflected or the word class changes. E.g. contain- is the stem of the words contained and container. A **stemmer** automatically reduces a word to its stem.

Part of speech: The part of speech (POS) of a word is its word class, i.e., noun, verb, determiner, adjective, etc. The set of available parts of speech is generally language specific and is provided by reference corpora, e.g. the Penn treebank for English [56]. A **POS tagger** automatically labels words with their parts of speech.

Chunk: A chunk is the segment of a sentence that identifies a basic non-recursive phrase corresponding to one of the major parts of speech: noun phrases (NPs), verb phrases (VPs), adjective phrases (APs) or prepositional phrases (PPs) [106]. In contrast to traditional phrase structures, chunks build flat, i.e. non-hierarchical and non-overlapping, sequences. A **chunker** segments sequences of words into chunks and labels them as NP chunk, VP chunk, etc.

Parse: Syntactic structures of sentences are typically assumed to have a hierarchical representation in the form of a tree. A parse is the syntactic tree of a sentence. A **parser** determines this structure automatically.

Head, modifier: Syntactic phrases consist of one (lexical) head and possibly of modifiers. The head of a phrase is the word which is grammatically most important in the phrase, since it determines the nature of the overall phrase [89]. For instance, the head of a verb phrase is a verb, the head of a noun phrase is a noun. Modifiers are optional elements of phrases. They can be words, phrases and clauses.

Named entity: An entity for which one or many rigid designators [46] can be used to refer to it, e.g. *the software company created by Bill Gates in 1975* can be referred to as Microsoft or Microsoft Corporation. Following the terminology found in some ontological analysis studies, e.g. OntoClean [36], we sometimes also refer to them as *individuals*. The recognition of proper names of persons (Bill Gates), geographical entities (Redmond, Washington) and organizations (Microsoft) is an important task in computational linguistics. A **Named Entity Recognizer** performs this task automatically.

Word sense: Words can have different meanings depending on their context of occurrence. E.g. star can be used to refer to an astronomical object, an actor or the state of being prominent, etc. Typically, sense inventories are obtained from semantic lexica such as WordNet. Word senses from WordNet can be denoted with a superscript indicating the sense number (ordered by frequency of occurrence in the manually sense-tagged SemCor corpus [64]) and a subscript indicating the word class, e.g. star_n^1, star_n^4, star_v^1. The task of automatically determining word senses is called **Word Sense Disambiguation** [71].

Fig. 1. Glossary of relevant computational linguistics terminology.

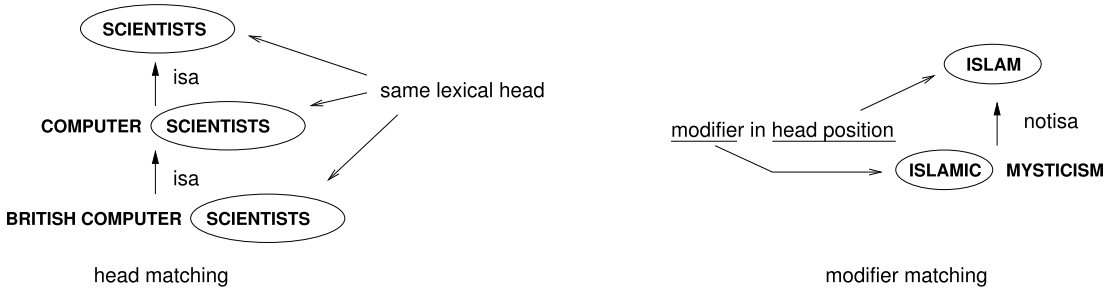

Fig. 2. Syntax-based methods. Two category labels are parsed and their relation set to *isa*, if they share the same lexical head word or lemma (head matching); the relation is set to *notisa*, if the stem of the lexical head of one category occurs in non-head position in the other (modifier matching).

2.3.1. Head matching (HeadMatcher)

We first label pairs of categories sharing the same lexical head, e.g. British computer scientists *isa* Computer scientists. We parse the category labels using the Stanford parser, based on the factored model from [45],[5] and use the head rules from [23, Appendix A]. Most Wikipedia category labels are noun phrase (NP) fragments rather than full sentences which required modifying Collins' head finding rules (see Appendix A).

Given the lexical heads for a pair of categories, we label a category link as *isa* if the two categories share the same head lemma, as determined by a finite-state morphological analyzer [67].

2.3.2. Modifier matching (ModifierMatcher)

We next label category pairs as *notisa* if the stem of the lexical head of one of the categories (as output by the Porter stemmer [87]) occurs in modifier, i.e. non-head, position in the other category. This is to avoid interpreting thematic categorization links as *isa* – such as the one between Crime comics and Crime or the one between Islamic mysticism and Islam.

Examples of head and modifier matching are presented in Fig. 2. These methods achieve good coverage by identifying 141,728 *isa* relations by head matching and 67,437 *notisa* relations by modifier matching, respectively. Both methods are high-precision heuristics, i.e. they achieve high precision *except* in the following cases.

[5] Although chunkers perform more accurately than full syntactic parsers, in order to find the head of the category labels we need phrase structures since these are not necessarily base noun phrases (NPs), e.g. Ice hockey players by club in Canada.

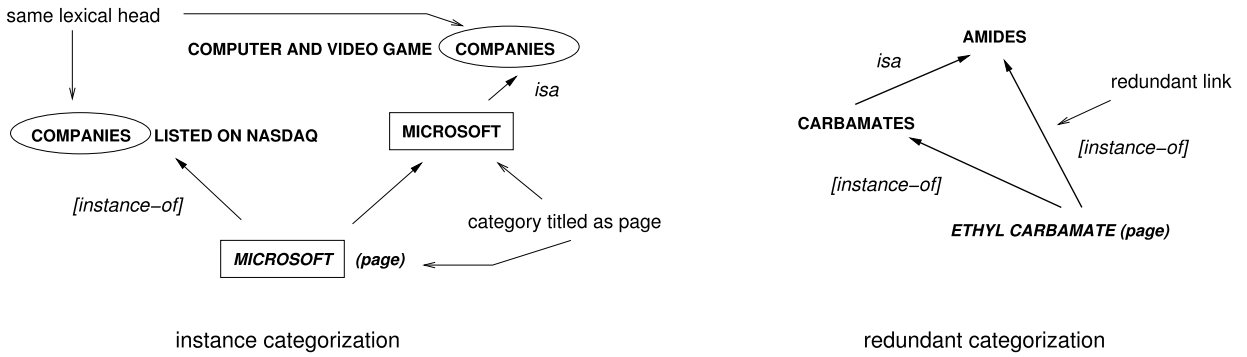

Fig. 3. Connectivity-based methods. The relation between two categories is set to *isa* if the subcategory refers to an individual which is an instance of the super-category (instance categorization) or the categories share at least one page (redundant categorization).

- Head matching will erroneously succeed, if the modifiers select different senses for the respective heads, e.g. CAUCUS CHAIR and CHAIR, or the relation expressed is not an *isa* relation, e.g. meronymy as in WEST JAVA and JAVA.
- Modifier matching will erroneously succeed in those cases in which the head of a category label modifies the head of another to select a compatible sense, e.g. ELECTRONIC MUSIC and MUSIC GENRES.

These methods are lightweight and high-coverage. Sample errors show that they rely merely on the string matching of syntactic constituents and therefore do not take advantage of any notion of *semantics*.

2.4. Connectivity-based methods (4)

The next set of methods utilizes the structure and connectivity of the categorization network.

2.4.1. Instance categorization (InstanceCategorization)

Previous work from [105] shows that *instance-of* relations in Wikipedia between individuals (denoted by pages) and classes of individuals (denoted by categories) can with high accuracy be found heuristically by determining whether the head of the page's category is plural, e.g. ALBERT EINSTEIN belongs to the NATURALIZED CITIZENS OF THE UNITED STATES category (where the syntactic head CITIZENS is plural). Since our definition of *isa* relations includes instantiation we apply this method for identifying *isa* relations between categories as follows.

1. Find page: given a category CAT, find the page P titled as the category or its lemma.
2. Collect candidate *isa* relations: collect all lexical heads which are plural nouns from the list of categories P is categorized into: $HP = \{head_1, head_2, \ldots, head_n\}$. Plural nouns are found using the output of the morphological analyzer.[6]
3. Propagate candidate relations: for each super-category SUPERCAT of CAT, we label the relation between CAT and SUPERCAT as *isa* if the head lemma of SUPERCAT matches the head lemma of at least one candidate in *HP*.

The idea is to collect evidence from the pages describing the categories and propagate such evidence to the categorization network. As an example (illustrated in Fig. 3), given the pair ⟨MICROSOFT, MULTINATIONAL COMPANIES⟩:

1. We first find the page MICROSOFT for the category MICROSOFT;
2. From the page MICROSOFT being categorized into COMPANIES LISTED ON NASDAQ, COMPANIES IN THE NASDAQ-100 INDEX and COMPANIES BASED IN REDMOND, WASHINGTON we collect {companies} as a candidate extraction for an *isa* relation;
3. We find that MICROSOFT *isa* MULTINATIONAL COMPANIES (as well as *isa* VIDEO GAME COMPANIES, *isa* COMPUTER COMPANIES OF THE UNITED STATES, etc.).

Manual inspection of the output reveals that, when applied to the categorization network, this method, originally developed to populate WordNet with instances from Wikipedia [105], indeed mostly works with candidate subconcepts which refer to individuals (companies, cities, organizations). Nevertheless it is also able to generate strict subsumption relations, i.e. examples include ELECTRIC MOTORS *isa* ENGINES, LOGIC *isa* BRANCHES OF PHILOSOPHY, ECONOMICS *isa* SOCIAL SCIENCES, etc.

[6] During system prototyping we noticed that this provides for a safer strategy than checking whether nouns have been POS-tagged as `NNS` or `NNPS` based on the parser's output. This is due to the parser producing erroneous output when analyzing small NP fragments.

I. *isa* patterns

1. NP$_2$, ? (such as|like|, especially) NP* NP$_1$
   a stimulant such as caffeine
2. such NP$_2$ as NP* NP$_1$
   such stimulants as caffeine
3. NP$_1$ NP* (and|or|,like) other NP$_2$
   caffeine and other stimulants
4. NP$_1$, one of det_pl NP$_2$
   caffeine, one of the stimulants
5. NP$_1$, det_sg NP$_2$ rel_pron
   caffeine, a stimulant which
6. NP$_2$ like NP* NP$_1$
   stimulants like caffeine
7. NP$_1$ (is|are) NP* NP$_2$
   caffeine is a stimulant
8. NP$_2$ includ(e|es|ing) NP* NP$_1$
   stimulants including caffeine

II. *notisa* patterns

1. NP$_2$'s NP$_1$
   car's engine
2. NP$_1$ in NP$_2$
   engine in the car
3. NP$_2$ with NP$_1$
   a car with an engine
4. NP$_2$ contain(s|ed|ing) NP$_1$
   a car containing an engine
5. NP$_1$ of NP$_2$
   the engine of the car
6. NP$_1$ are? used in NP$_2$
   engines used in cars
7. NP$_2$ ha(s|ve|d) NP$_1$
   a car has an engine
8. NP$_1$ (is|are) part of NP$_2$
   engines are part of cars

Fig. 4. Patterns for *isa* and *notisa* detection. NP$_1$ represents the hyponym, NP$_2$ the hypernym, i.e., we aim to retrieve NP$_1$ *isa* NP$_2$; NP* represents zero or more coordinated NPs.

2.4.2. Redundant categorization (RedundantCategorization)

This method labels pairs of categories which have at least one page in common (for an illustration see Fig. 3). If users redundantly categorize by assigning two directly connected categories to the same page, they often mark the page by implicature as being an instance of two different category concepts with different granularities, e.g. ETHYL CARBAMATE is both a CARBAMATE(S) and an AMIDE(S). Assuming that the page is an instance of both conceptual categories, we can by transitivity conclude that one category is subsumed by the other, i.e. CARBAMATES *isa* AMIDES. In order to identify instantiation we again use the shallow method from [105], i.e. *instance-of* relations between pages and their categories are found by determining whether the head of the page category is plural. Thus redundant categorization tags the relations between two directly connected categories as *isa* if (1) there is at least one page categorized in both categories and (2) the category labels both have a plural head.

Using instance categorization and redundant categorization we find 14,886 and 16,523 *isa* relations, respectively. Both methods provide positive *isa* links in cases where relations are unlikely to be found in free text, e.g. we find that ALAN TURING *isa* ENGLISH MATHEMATICIANS and AMERICAN COUNCIL ON SCIENCE AND HEALTH *isa* SCIENTIFIC ORGANIZATIONS (instance categorization), as well as that ALKALOIDS *isa* BIOMOLECULES and GASTROPODS *isa* MOLLUSCS (redundant categorization), although we do not find any evidence in text corpora using the pattern-based approach we describe in Section 2.5.

These methods suggest that it is possible to generate semantic relations by capturing patterns employed by Wikipedia's user base. We achieve this by analyzing the structure and connectivity of the categorization network. Nevertheless, these knowledge-poor heuristics are unconstrained and produce many errors, e.g. IMMANUEL KANT *isa* ETHICS (instance categorization) or ATOMIC PHYSICS *isa* QUANTUM MECHANICS (redundant categorization). In order to improve their precision while still taking advantage of their coverage we use a filter based on lexico-syntactic patterns from the literature on semantic relation extraction.

2.5. Lexico-syntactic based methods (PatternFinder – 5)

After applying methods (1–4) we are left with 147,929 unclassified relations. We next apply lexico-syntactic patterns (see Fig. 4) to sentences in large text corpora to identify *isa* relations [39,17]. In order to reduce the number of unclassified relations and to increase the precision of the *isa* patterns we also apply patterns to identify *notisa* relations. We assume that patterns used for identifying meronymic relations [7,35] indicate that the relation is not an *isa* relation. The text corpora used for this step are the articles from the English Wikipedia itself (8×10^8 words) and the Tipster corpus (2.5×10^8 words; [37]). We employ a majority voting strategy for providing evidence for semantic relations: we label a category pair with *isa* if the number of matches of *isa* patterns is greater than the number of matches of *notisa* ones.

In addition, we use the patterns to filter the *isa* relations created by the connectivity-based methods (Section 2.4). This is because instance categorization and redundant categorization return results which are not always reliable, e.g. we incorrectly find that CONSONANTS *isa* PHONETICS. We use the same majority voting scheme, except that this time we mark as *notisa* those pairs with a number of *notisa* matches greater than the number of *isa* ones. This acts as a post-processing filter for the connectivity-based methods and ensures better precision.

To improve the recall of applying these patterns, we use only the lexical heads of the categories which were not identified as named entities: if the lexical head of a category is identified by a Named Entity Recognizer [29] as belonging to a

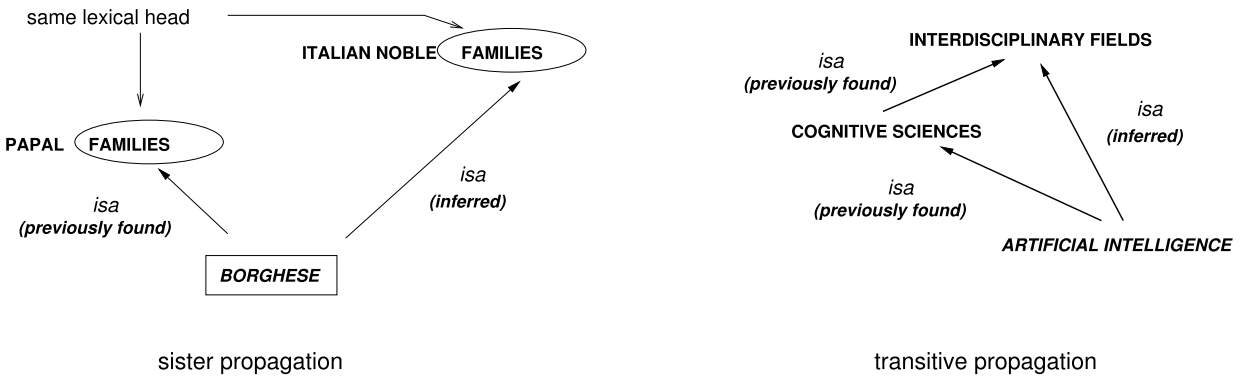

Fig. 5. Inference-based methods. The relation between two categories is set to *isa* if the super-category has the same head as a previously identified *isa* super-category (i.e. sister propagation) or if there is a path connecting the two categories along the previously discovered *isa* hierarchy (transitive propagation).

named entity, e.g. Brands in YUM! BRANDS, we use the full category name, otherwise we simply use the head, e.g. albums in MILES DAVIS ALBUMS. In order to ensure precision in applying the patterns, both the Wikipedia and Tipster corpora were preprocessed by a pipeline consisting of a trigram-based statistical POS tagger [11] and a SVM-based chunker [47] to identify noun phrases (NPs). POS tagging helped us disambiguate many ambiguous expressions, e.g. *isa* pattern 1 from Fig. 4 did not match when like was tagged as a verb. Chunking allowed us to identify phrase boundaries and hence to match phrases instead of simple strings.

These methods achieve large coverage by creating 49,054 and 37,188 *isa* relations when applied respectively to the output of the syntax-based methods (Section 2.3) and the connectivity-based methods (Section 2.4). In addition, when applied in order to filter the *isa* relations generated by the connectivity-based methods, they are able to filter out 3226 erroneously identified positive links.

2.6. Inference-based methods (6)

The last set of methods propagates the previously found relations by means of multiple inheritance and transitivity (Fig. 5).

2.6.1. Multiple inheritance propagation (SisterPropagation)

We first propagate all *isa* relations to those super-categories whose head lemmas match the head lemma of a previously identified *isa* super-category. This amounts to propagating the *isa* relation to all sisters of the previously identified *isa* super-categories which share the same head lemma. E.g., once we have found that BORGHESE *isa* PAPAL FAMILIES (via the 'redundant categorization' method from Section 2.4.2) we can infer also that BORGHESE *isa* POLITICAL FAMILIES OF ITALY, ITALIAN NOBLE FAMILIES and FAMILIES OF SIENA.

2.6.2. Transitivity propagation (TransitivityPropagation)

We then propagate all *isa* links to those super-categories which are connected through a path found along the previously discovered subsumption hierarchy, thus taking the transitive closure of the *isa* relation. E.g., given that ARTIFICIAL INTELLIGENCE *isa* COGNITIVE SCIENCE (using the pattern-based methods from Section 2.5) and COGNITIVE SCIENCE *isa* INTERDISCIPLINARY FIELDS (found via the 'instance categorization' method from Section 2.4.1), we can infer that ARTIFICIAL INTELLIGENCE *isa* INTERDISCIPLINARY FIELDS.

3. Evaluation

The methods presented in the previous section generate a very large taxonomy as output – i.e. using all methods we generate 208,208 *isa* semantic links between 169,009 categories. Although there is no consensus on how to evaluate such ontological and taxonomic resources [10], the large size of our taxonomy makes a comprehensive qualitative evaluation via manual inspection impractical.[7] We therefore opt for *manual evaluation* on a representative sample of category pairs (Section 3.1) following standard experimental procedures (see e.g. the setting proposed by [102] for evaluating taxonomy-based ontologization). In addition, we develop an *automatic evaluation* setting for quantifying the coverage and novelty of the semantic relations extracted (Section 3.2). Both evaluations can be seen as gold standard based evaluations, since we evaluate the automatically generated taxonomy by comparing it with gold standard resources, i.e. a manually annotated dataset as well as ResearchCyc and WordNet.

[7] In other words, manual assessment of how well the resource meets the criteria and requirements of a methodology such as e.g. OntoClean [36].

Table 1
Manual evaluation. Recall (R), precision (P), F_1 and accuracy (A) figures against the manually annotated gold standard (we report percentages).

	R	P	F_1	A
baseline	49.5	64.8	56.1	49.1
syntax (1–3)	56.2	92.4	69.9	68.1
connectivity (1–4, 6)	66.9	92.3	77.6	74.7
pattern-based (1–3, 5–6)	72.4	90.4	80.4	76.7
all (1–6)	78.5	90.3	84.0	80.3

Finally, in order to provide additional evidence for the quality of the resource, we perform an extrinsic, task based evaluation in Section 3.3. We evaluate the resource by using it in a Natural Language Processing (NLP) task. We compute *semantic similarity* on benchmarking datasets by coupling the WikiRelate! method [104] with the generated taxonomy. We evaluate the output by comparing it with similarity scores computed using the most widely used resource for such task, i.e. the semantic lexicon provided by WordNet.

3.1. Manual evaluation

In our first set of experiments we aim to assess the quality of our resource by manually evaluating it against a human-annotated gold standard. We manually annotated the *isa*/*notisa* relation for a random sample of 3500 category pairs from Wikipedia by asking three annotators to provide the ground truth for 1000 different pairs of categories each, following the guidelines given in Appendix B. In addition, in order to assess how reliable the annotations were as well as the difficulty of the task, all three annotators were asked to label a common separate dataset of 500 category pairs. We computed the degree of inter-annotator agreement among annotators using the kappa coefficient κ [18].[8] κ measures pairwise agreement among a set of annotators making category judgments, correcting for expected chance agreement. Our annotators achieved an agreement coefficient κ of 0.78, indicating substantial but not perfect agreement. This score is compatible with the only application of a reliability measurement for a taxonomy annotation task we are aware of, namely [62] who achieved a κ of 0.75. In the case of disagreement between annotators we selected the relation annotated by the majority (there were no three-way ties in the annotations). Each category pair from Wikipedia is assigned an *isa* or a *notisa* relation by both the annotators and the system. Accordingly, we are able to build a confusion matrix for all category pairs in our dataset and evaluate using standard metrics of precision (P, the ratio of correct *isa* relations to total *isa* labels output by the system), recall (R, the ratio of correct *isa* relations to total *isa* labels in the gold standard) and F_1 measure ($\frac{2PR}{P+R}$). In addition we calculate accuracy, which also takes into account the assignment of *notisa* relations.

Evaluation of the automatically generated taxonomy with the manually annotated category pairs is presented in Table 1. We perform an incremental evaluation by starting with the syntax-based methods and augmenting them with the connectivity and pattern based methods. As a baseline we use a random classification scheme, i.e. a category pair is randomly categorized as *isa* or *notisa*. All differences in performance are statistically significant at $p < 0.001$ with a McNemar test.

The results provide a first glance of the impact of our methods. First, the random baseline achieves an F_1 measure well above 50%: this is because the *isa* and *notisa* relations are not uniformly distributed in our dataset (and thus in the Wikipedia categorization itself, since our data come from a random sample), which contains 2302 *isa* and 1198 *notisa* relations. Syntax-based methods achieve very high precision but their impact is limited by their low recall. Augmenting them with the connectivity-based methods improves the recall (+10.7%) with practically no decrease in precision (−0.1%). We observe a similar trend by applying the pattern-based methods together with the syntax-based ones. They improve recall even more considerably (+16.2%), but also have lower precision (−2.0%).

The best results are obtained by combining all methods: in this way we achieve a 22.3% improvement in recall and a 2.1% decrease in precision, resulting in an overall improvement of 14.1% F_1 above the simple syntax-based methods. The resulting taxonomy achieves high precision and somewhat satisfying recall with an overall improvement of 27.9% F_1 with respect to the random baseline. We observe the same trend with accuracy: by making use of all our methods we are able to achieve an accuracy of more than 80%, with an improvement of 31.2 points with respect to the baseline.

3.2. Comparison with ResearchCyc and WordNet

Manual evaluation quantifies how good our taxonomy is when compared with human judgments. The results show that we are able to generate a very large taxonomy with high precision. However, while this evaluation tells us about the *quality* of the resource, it still does not say much on how well the taxonomy compares with other existing resources: while results thus far have shown that we can generate a high-quality taxonomic resource by using straightforward heuristics, it could still be the case that the information we extract from the Wikipedia category system can already be found in other semantic networks.

[8] There are several formulations of the kappa coefficient in the literature. We use one variant, Fleiss' kappa [30], which is a generalization of Scott's π for more than two annotators.

Table 2
Statistics for resources used for evaluation. For both ResearchCyc and WordNet we solely report the number of concepts and relations included in their taxonomic structure (i.e. those concept pairs which are annotated as being in an *isa* relation). The output of our system (Wikipedia taxonomy) consists of all categories and links from Wikipedia *minus* those removed by the 'category network cleanup' (Section 2.1), i.e. it includes the *is-refined-by* and *isa* relations. The size of the taxonomy favorably compares to that of both ResearchCyc and WordNet, as it ranges in the middle between these two.

		ResearchCyc	WordNet	Wikipedia	Wikipedia (taxonomy)
# nodes	# concepts	357,790			
	# synsets		95,322		
	# categories			337,522	209,919
# edges	# assertions	733,865			
	# semantic pointers		97,666		
	# category links			743,140	335,128

In order to investigate the *kind* of structured knowledge contained in the Wikipedia taxonomy, we designed a second set of experiments aimed at quantifying the novelty of our resource when compared with human annotated resources taken as gold standard. For this purpose, we use ResearchCyc,[9] the research version of the Cyc knowledge base [50] and WordNet[10] [28]. Figures on the size of the resources used in the gold standard evaluation are given in Table 2. When comparing against an existing gold standard taxonomy, we aim to answer two fundamental questions to characterize our Wikipedia-based resource:

1. How much information does our taxonomy contain that can already be found in existing resources? In other words, *how much do we cover existing resources*, thus only providing duplicated information from yet another knowledge source?
2. How much *novel information* does our resource contain when compared with other resources? In other words, what is the 'added value' of our taxonomy in comparison with existing knowledge resources?

When answering these two questions, we would ideally like our resource to be as novel as possible or otherwise to contain as little pre-existing information from other knowledge sources as possible. This is due to the fact that, while being highly disjoint, different resources can always be merged together in a second stage, for instance by means of an automatic mapping procedure [82]. Moreover, in order to quantify these two aspects on a large scale, namely for a very large sample of concepts and relations, we propose an automatic evaluation method based on a simple, yet effective, way of mapping Wikipedia categories to concepts in the gold standard resource. We then define a set of metrics to compute both coverage and novelty based on these mappings.

3.2.1. Taxonomy mapping

In order to check whether an *isa* relation between a pair of categories is in fact novel, we first need to find the concepts to which these can be mapped in the gold standard. In practice, we aim at finding a mapping between automatically generated and gold standard resources at the *concept level*, i.e. an instance of the general problem of *ontology matching* [27].

Starting with a pair of categories from our taxonomy, for each category (a) we first define a set of *category descriptors*, namely a set of lexicalizations (i.e. words and phrases) of the category found in the gold standard resource. We then (b) generate a set of *candidate mappings* for the category descriptors and finally (c) *select the mappings* where the target concepts are also found in an *isa* relation in the gold standard.

Generating category descriptors. Given a category CAT, we first collect the set of concepts from the gold standard resource whose lexical realizations (namely, words and phrases) match the Wikipedia category labels. For instance, given the category ICE HOCKEY PLAYERS BY CLUB IN CANADA, we want to find those concept labels that can be found to describe it in Cyc or WordNet. We call these words and phrases *category descriptors* of a category CAT and denote them with $\phi(\text{CAT})$.

Our method starts by first looking for an exact match: if none can be found, we fall back to less and less specific matches by using the syntactic parse of the category label. Accordingly, our process to generate category descriptors involves three phases:

1. *Strict* match. We start by looking for a target concept where at least one lexical realization perfectly matches the category label. For instance, in the case of Cyc, we first search for a concept labeled as ice hockey players by club in Canada. Similarly, in the case of named entities such as ALAN TURING, we look for Alan Turing;
2. *Loose* match. If no category descriptor can be found by exact match, as in the case of ice hockey players by club in Canada, we start approximating our lookups as follows:

[9] http://research.cyc.com/. We use version 1.0 released in July 2006.

[10] http://wordnet.princeton.edu/. We use version 3.0 released in December 2006. We denote the i-th sense of a word w with part of speech p with w_p^i. We use word senses to unambiguously denote the corresponding synsets (e.g. plane_n^1 for {airplane_n^1, aeroplane_n^1, plane_n^1}). Hereafter, we use *word sense* and *synset* interchangeably.

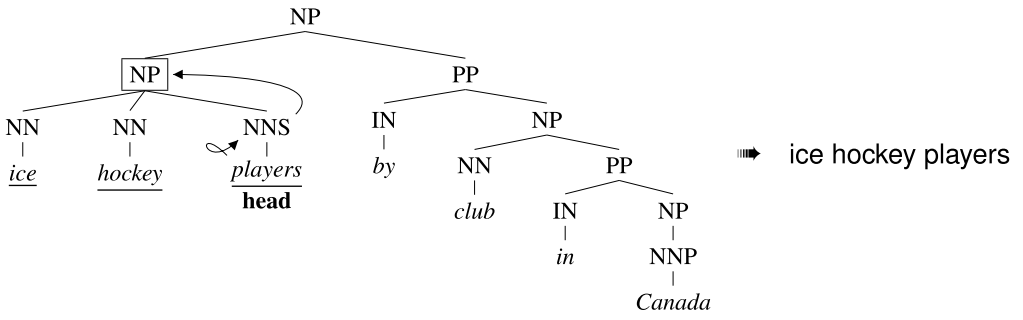

Fig. 6. Loose mapping of category labels to the gold standard lexicon. If the full category label cannot be found and it has not been recognized as a named entity, we take the output of the parser and find the minimal NP projection of the lexical head. Such an NP is found by starting from the head terminal and percolating up the tree until the first NP node is found.

(a) take the parse of the category label;
(b) find the lexical head of the tree;
(c) find the minimal NP projection of the head.
The latter is the lowest noun phrase in the tree that contains the head and noun modifiers (Fig. 6). In our example this amounts to looking for a concept labeled ice hockey players.[11]

3. *Head* match. Finally, if no loose match can be found, e.g. ice hockey players cannot be found in WordNet, we fall back to looking for a concept labeled as the category label's head, namely players.

As a result of the above three-tier procedure, the category descriptors for ICE HOCKEY PLAYERS BY CLUB IN CANADA are ice hockey player and player for Cyc and WordNet respectively.

Mapping category descriptors. Given a descriptor for a category CAT, in the next phase we aim at acquiring the 'correct' concept for the category descriptor in the gold standard. In practice, we want to acquire a mapping μ for each CAT such that:

$$\mu(\text{CAT}) = \begin{cases} c \in \mathit{Senses}_{\text{CYC/WN}}(\phi(\text{CAT})) & \text{if a link can be established,} \\ \epsilon & \text{otherwise,} \end{cases}$$

where $\mathit{Senses}_{\text{CYC/WN}}(\phi(\text{CAT}))$ is the set of concepts that $\phi(\text{CAT})$ can refer to in the gold standard (i.e. Cyc or WordNet). Note that, due to the polysemy of words and phrases, $\mathit{Senses}(\phi(\text{CAT})) \geqslant 1$, e.g. Toyota can refer to both `ToyotaCar` and `ToyotaCompany` in Cyc. Given a category CAT and its descriptor $\phi(\text{CAT})$, we find $\mathit{Senses}(\phi(\text{CAT}))$ by using an internal lexeme-to-concept denotational mapper in the case of Cyc, and by finding those synsets which contain the descriptor for WordNet. For instance, given $\phi(\text{ICE HOCKEY PLAYERS BY CLUB IN CANADA}) = \text{player}$, we find that $\mathit{Senses}_{\text{WN}}(\text{player}) = \{\text{player}_n^1 \ldots \text{player}_n^5\}$.

Given the category descriptors and their senses, we can view the mapping procedure as a disambiguation problem. In other words, the mapping algorithm must disambiguate $\phi(\text{CAT})$ based on some context. Given a category pair, our idea is to then jointly disambiguate the descriptors by letting them provide a context for each other. Formally, given a pair of (directly connected) categories $\langle \text{SUBCAT}, \text{SUPERCAT} \rangle$ in our Wikipedia taxonomy, we first create the set of all concept pairs in the gold standard resource they can be mapped to:

$$\mathit{ConceptPairs} = \left\{ \langle c_i, c_j \rangle \mid c_i \in \mathit{Senses}\big(\phi(\text{SUBCAT})\big),\ c_j \in \mathit{Senses}\big(\phi(\text{SUPERCAT})\big) \right\}.$$

We finally map the pair of categories to those senses of their descriptors in *ConceptPairs* such that these are in an *isa* relation within the subsumption or instantiation hierarchy of the gold standard[12]:

$$\big\langle \mu(\text{SUBCAT}), \mu(\text{SUPERCAT}) \big\rangle = \begin{cases} \langle c_i, c_j \rangle \in \mathit{ConceptPairs} & \text{if } \exists [c_i \in \mathit{Senses}(\phi(\text{SUBCAT})),\ c_j \in \mathit{Senses}(\phi(\text{SUPERCAT}))] \\ & \text{such that } c_i \sqsubseteq c_j, \\ \epsilon & \text{if no such pair } \langle c_i, c_j \rangle \text{ exists.} \end{cases}$$

In the case of a tie, namely when more than one such pair exists, we make a random choice for Cyc and select the more frequent word sense in WordNet.[13] For example, given a pair of categories ⟨STARS, ASTRONOMICAL OBJECTS⟩ we map in

[11] Note that by keeping the noun modifiers we reduce the amount of polysemy in the mapping. I.e., we avoid mapping ice hockey players by club in Canada to MUSICIAN via a polysemous mapping of player only.

[12] When traversing the gold standard hierarchy for determining the mapping we do not distinguish between subsumption and instantiation. In other words, we look for a connecting path from c_i to c_j along any edge denoting either an *isa* (Cyc's 'is generalized by', `#$genls`) or *instance-of* (Cyc's `#$isa`) relation. This is compatible with our broader definition of the *isa* relation from Section 2.

[13] Since senses in WordNet are ordered according to their frequency of occurrence in the manually sense-tagged SemCor corpus [64], this amounts to selecting the more frequent SemCor sense.

WordNet the former to star_n^1 – i.e. 'celestial bodies' rather than 'prominent actors' – and the latter to object_n^1, since star_n^1 *isa* object_n^1. Similarly, we map such pair in Cyc to `Star` (rather than, say, `FamousPerson`) and `PartiallyTangible`, since the latter dominates the former within its subsumption hierarchy.

Our method for taxonomy mapping makes a variety of very shallow approximations, since in the initial phase it considers all senses of a category descriptor and performs noun compound segmentation based only on the syntactic structure of the category label (rather than e.g. also imposing some semantic constraints). However, while simple, the method is able to yield high precision. We validated the output of the category mapper on a sample of 200 non-empty category pair mappings, i.e. 400 category mappings in total. A human validator with previous experience in ontology mapping and engineering was presented with each category pair one by one, and then labeled each category in the pair as correctly mapped or not. A mapping was deemed correct if the Wikipedia category and the concept it maps to in the external resource were judged to have the same meaning, e.g. RAYS refers to ray_n^7 in WordNet. Our method achieves a precision of 97.75% and 97.5% for Cyc and WordNet respectively, thus proving itself suitable for evaluating the coverage and novelty of our resource.

3.2.2. Evaluation metrics

Given the automatically generated mappings, we are able to compute the metrics of coverage and novelty against a gold standard. Let $G_{Wiki} = \langle V_{Wiki}, E_{Wiki} \rangle$ be our taxonomy, where the vertices represent the categories and the edges the automatically generated *isa* relations, and $G_S = \langle V_S, E_S \rangle$ a gold standard taxonomy. Given G_{Wiki} and the category mappings to the concepts in G_S, we can define a subgraph $G'_{Wiki} = \langle V'_{Wiki}, E'_{Wiki} \rangle$ of G_{Wiki} containing: (a) as vertices, all categories $v \in V_{Wiki}$ such that $\mu(v) \neq \epsilon$; (b) as edges, the *isa* relations $e \in E_{Wiki}$ between them. Finally, we can define *coverage* against a gold standard resource as the number of edges in the subgraph containing mapped categories to the number of edges in the gold standard graph:

$$\mathit{Coverage}\big(G'_{Wiki}, G_S\big) = \frac{|E'_{Wiki}|}{|E_S|}.$$

Coverage quantifies how many pairs of categories in our Wikipedia taxonomy can be mapped to concepts in a subsumption relation in the gold standard to the total number of pairs of concepts found in an *isa* relation in the latter. Coverage thus measures the size of the intersection between our taxonomy and another knowledge resource. However, when compared to gold standard resources, our taxonomy also contains novel concepts and relations. Accordingly, we can compute the novelty rate of our resource by calculating the proportion of how many pairs of Wikipedia categories are deemed to be in an *isa* relation but have no suitable mapping in the gold standard:

$$\mathit{Novelty}\big(G_{Wiki}, G'_{Wiki}\big) = \frac{|E_{Wiki} \setminus E'_{Wiki}|}{|E_{Wiki}|}.$$

Finally, we can also compute the 'gain' in knowledge provided by our resource with respect to existing knowledge bases by calculating the proportion of unmapped category pairs in an *isa* relation to the total number of semantic relations in the gold standard. We call this metric *extra coverage*:

$$\mathit{ExtraCoverage}\big(G_{Wiki}, G'_{Wiki}, G_S\big) = \frac{|E_{Wiki} \setminus E'_{Wiki}|}{|E_S|}.$$

Intuitively, novelty and extra coverage both quantify the proportion of pairs of categories $\langle c_i, c_j \rangle$ in our Wikipedia taxonomy for which no mapping can be established to the total number of semantic edges found in either the taxonomy itself or the gold standard, respectively.

In practice, while our metrics allow us to quantify the amount of novel and pre-existing information that our method is able to automatically generate, there are cases, such as when we approximate the category label using a *loose* or *head* category descriptor, where the question of whether an *isa* relation is in fact novel is not straightforward. Consider for instance the category pair ⟨STARS, ASTRONOMICAL OBJECTS⟩ from before. Here, we are indeed covering a pre-existing relation in both of our gold standard resources, namely that `Star` *isa* `PartiallyTangible` and star_n^1 *isa* object_n^1 in Cyc and WordNet respectively (since we can approximate the label astronomical object as object). However, we are also generating new information by saying that a (sense of) star is not a generic object, but rather an *astronomical object*. In order to be able to quantify these mixed novel/covered relations, we adopt the following solution: when we compute *coverage*, we use all three category descriptors, i.e. *strict*, *loose* and *head*, to find the candidate senses of a category in the gold standard; when we instead compute *novelty* and *extra coverage*, we only make use of a strict match, i.e. a relation is not novel in the case that we can find two senses of the exact-matching category descriptors such that the senses are found in an *isa* relation in the gold standard. This allows us to make the pair ⟨STARS, ASTRONOMICAL OBJECTS⟩ count as both a covered and a novel relation.

3.2.3. Results and discussion

Table 3 shows the results obtained by comparing the relations generated by our methods with ResearchCyc and WordNet. The evaluation is performed incrementally: we start with the syntax-based methods (i.e. head matching) and augment them with different sets of methods, namely our connectivity and pattern based methods. All differences in performance between different sets of methods are statistically significant at $p < 0.001$ with a McNemar test.

Table 3
Automatic evaluation. Comparison with ResearchCyc and WordNet (we report percentages).

	Cyc			WordNet		
	Coverage	Novelty	ExtraCov	Coverage	Novelty	ExtraCov
syntax (1–3)	0.8	99.6	19.2	4.2	99.6	144.5
connectivity (1–4, 6)	1.4	99.2	23.2	7.2	99.3	174.0
pattern-based (1–3,5–6)	1.4	99.2	25.9	7.8	99.3	194.4
all (1–6)	1.6	99.2	28.2	8.7	99.3	211.6

Regardless of the gold standard employed, Cyc and WordNet's results are consistent and allow us to conclude that the information contained in our taxonomy is indeed of a different kind than the one found in other knowledge sources. We first note that overall coverage is low, i.e. up to 1.6% and 8.7% for Cyc and WordNet respectively. This is due to the fact that categories in Wikipedia provide a thematic meta-classification scheme for the resource's encyclopedic entries, i.e. its pages, and accordingly have little overlap with the concepts found in manually built semantic networks. As a result, many categories cannot be found in the gold standards for a variety of specialized domains, e.g. there is no concept corresponding to BACTERIAL PROTEINS or USB (as in 'Universal Serial Bus') in either Cyc or WordNet. Besides, while the majority of category pairs can be successfully mapped to concept pairs in the target resources, i.e. 140,844 and 139,635 (67.6% and 67.0% of our taxonomy) for Cyc and WordNet respectively, many of these mappings end up covering the *same* relation and therefore do not count as multiple instances of a covered semantic edge. For instance, both ⟨EARLY MIDDLE AGES, HISTORICAL ERAS⟩ and ⟨LATE MIDDLE AGES, HISTORICAL ERAS⟩ cover the same *isa* relation in WordNet, namely age^2_n *isa* era^1_n. As a result, we are able to cover only 11,702 and 8465 unique relations for Cyc and WordNet, respectively.

The low coverage of our resource is counterbalanced by an extremely high novelty rate consistently above 99% for all methods and gold standards, as well as substantial extra-coverage for both Cyc (up to 28.2%) and WordNet (up to 211.6%). Using all our methods (1–6), we are in fact able to generate 206,635 and 206,673 *isa* relations which cannot be found in Cyc or WordNet respectively (based on our category mapping procedure).

The simple syntax-based methods achieve high novelty and extra-coverage with low coverage as a trade-off. This is because many of our categories can be mapped to concepts in the gold standard resources via an approximate matching but still provide finer-grained concepts than traditional knowledge repositories: for instance, by using head matching, we can find that HISTORICAL BUILDINGS are BUILDINGS. While this information is indeed novel, since neither Cyc nor WordNet contains a reference for historical building, only categories with identical heads are being connected. As a result, we do not create a single interconnected taxonomy but rather many separate taxonomic islands where the extracted information is trivial. By applying the connectivity and pattern based methods at different stages we are able to improve coverage and extra-coverage at practically the same novelty rate – up to +6.7% and +49.9% extra-coverage for Cyc and WordNet respectively. A closer look reveals that applying these methods on top of the syntax-based ones creates an interconnected taxonomy where concepts with quite different linguistic realizations are connected. The best results are obtained by combining all methods: +9.0% and +67.1% extra-coverage when compared with the syntax-based methods, thus indicating that connectivity and pattern based methods generate different sets of *isa* relations and are complementary.

3.3. Computing semantic similarity using Wikipedia

We extrinsically evaluate the quality of our taxonomy by computing semantic compatibility scores between pairs of words in benchmarking datasets: we test whether by coupling standard metrics to compute semantic similarity with our automatically generated taxonomy we are able to achieve results competitive with human annotated knowledge bases such as WordNet, the *de-facto* standard resource for this task.

In the WikiRelate! approach [104,85] we proposed using the Wikipedia categorization as a conceptual network for computing the semantic relatedness of words and were able to significantly outperform approaches using WordNet. However, when applied to computing semantic similarity, WikiRelate! performed significantly worse than approaches using WordNet. We believe that this is due to the fact that approaches for measuring semantic similarity that rely on lexical resources usually use paths based only on *isa* relations [13]. These, however, are available in the taxonomy we develop in this work. Accordingly, we take datasets modeling human judgments of semantic similarity and see whether computing semantic distances using the *isa* paths improves when compared to using semantically unspecified paths.

We perform an extrinsic evaluation by computing semantic similarity on two commonly used datasets, namely Miller and Charles' list of 30 noun pairs [61] and the 65 word synonymy list from Rubenstein and Goodenough [96]. We compare the results obtained by using Wikipedia with the ones obtained by using WordNet, which is the most widely used lexical taxonomy for similarity computation. We evaluate performance by taking the Pearson product-moment (r) and Spearman rank (ρ) correlation coefficients between the similarity scores and the corresponding human judgments. While a number of previous works made use of the Pearson correlation metric [42,41,104], others evaluated using the Spearman correlation [32,40,113]. In practice, we believe that both metrics are useful in quantifying the performance of a method to compute semantic similarity: ideally, we would like the output of our system to (i) have a strong linear relationship with human scores (as measured by r), as well as (ii) accurately reproduce the ranking of word pairs given by human annotators (as

Table 4
Results on correlation with human judgments of similarity measures. Best results are bolded for each dataset and evaluation measure.

Method	Pearson's r				Spearman's ρ			
	path	*wup*	*lch*	*res*	*path*	*wup*	*lch*	*res*
	Miller and Charles							
WordNet	0.76	0.76	0.78	0.81	0.72	0.74	0.72	0.75
WikiRelate!	0.67	0.68	0.71	0.44	0.63	0.58	0.63	0.50
WikiRelate! *isa*	0.75	0.81	0.80	**0.87**	0.78	0.78	0.78	**0.79**
	Rubenstein and Goodenough							
WordNet	0.78	0.80	**0.84**	0.82	**0.78**	0.77	**0.78**	0.76
WikiRelate!	0.65	0.68	0.69	0.34	0.62	0.61	0.62	0.49
WikiRelate! *isa*	0.70	0.77	0.75	0.78	0.74	0.74	0.74	0.75

measured by ρ). Accordingly, we adopt recent proposals [38] and provide both the Pearson and Spearman correlation metrics.

Table 4 reports the scores obtained by computing semantic similarity in WordNet and in Wikipedia using different path-length based measures, including the simple edge counting method of Rada et al. [88] (*path*, henceforth) and the more refined (normalized) measures from Wu and Palmer [112] and Leacock and Chodorow [48] (abbreviated as *wup* and *lch* respectively). We also use the information content based measure originally developed by Resnik [92] (abbreviated as *res*). In the case of Wikipedia however, it is difficult to see how to compute the information content from the probabilities of occurrence of the category labels in a corpus. This is because most of these category labels are multi-word expressions in contrast to the majority of words in WordNet Resnik's measure was originally developed on.[14] In order to apply Resnik's measure to Wikipedia we couple it accordingly with an intrinsic information content-based measure relying on the hierarchical structure of the category network [101]. For this same reason, we do not use other information content measures such as the ones from [43] and [53], which have nevertheless both been shown to correlate slightly better with human judgments than Resnik's measure.

We take as baseline the WikiRelate! method outlined in [104] and extend it by first computing only paths based on *isa* relations. The results indicate that using *isa* relations works better than the simple WikiRelate! baseline.[15] This is because we are able to filter out category relations which decrease similarity scores, i.e. *notisa* (e.g. meronymic, antonymic) semantic relations. Using only paths along the *isa* hierarchy produces similarity scores which correlate slightly better with human judgments than WordNet on the Miller and Charles data and slightly less on the Rubenstein and Goodenough word pairs.

Our results on the Miller and Charles data are competitive with the best ones from the literature – i.e. [51] report $r = 0.89$ by combining path and information content based measures, whereas [2] report $r = 0.93$ by combining distributional and WordNet-based scores in a supervised learning setting – and lie near the estimated upper bound for the performance on this task – namely a correlation of $r = 0.90$, based on the replication study by [92] of Miller and Charles's experiments. However, we also notice that the results are less competitive when evaluating based on the Spearman correlation metric – i.e. [40] report $\rho = 0.90$ and $\rho = 0.84$ for the Miller and Charles and Rubenstein and Goodenough data respectively, whereas [2] report up to $\rho = 0.92$ and $\rho = 0.89$ for these two datasets. This indicates that our method is able to produce scores which well quantify the strength of the similarity between word pairs, but which are less effective at generating a ranking consistent with the relative ordering of the word pairs found in the gold standard. Similarly to the performance figures obtained with the Miller and Charles data, our results on the Rubenstein and Goodenough dataset are near the upper bound of $r = 0.80$ reported by [81]. However, on these data the Wikipedia-based scores have lower correlation with human judgments than WordNet, which, in contrast, exhibits better performance when compared with the results obtained on the Miller and Charles dataset. Overall, the results indicate that our taxonomy can be used to robustly compute the semantic similarity of words, since it yields correlation scores competitive with those of WordNet when applied to datasets specifically designed for this task.

4. Related work

In this section we relate our work to the existing body of literature on knowledge acquisition (Section 4.1) and then give an overview of automatic methods for extracting knowledge from Wikipedia and its application to AI and NLP tasks (Section 4.2).

[14] We could compute the information content by counting the occurrence of the heads of the category labels, but this would also have the side-effect of assigning the same information content to all categories with the same head.

[15] Differences in performance are statistically significant at 95% significance level ($p = 0.05$). For computing statistical significance we performed a paired t-test on each dataset for pairs of corresponding relatedness measures (e.g. between the WordNet and Wikipedia path measures).

4.1. Automatic knowledge acquisition

There is a large body of work concerned with acquiring knowledge for AI and NLP applications. Many NLP components can get by with rather unstructured, associative knowledge as provided by the cooccurrence of words in large corpora, e.g., distributional similarity [52,49,75,108,107, inter alia] and vector space models [100]. Such unlabeled relations between words were proven to be as useful for disambiguating syntactic and semantic analyses as the manually assembled knowledge provided by WordNet.

However, the availability of reliable preprocessing components like POS taggers, syntactic and semantic parsers allows the field to move towards higher level tasks, such as question answering, textual entailment, or end-to-end dialog systems which require a tighter notion of similarity (as noted e.g. by [33]). This lets researchers focus (again) on taxonomic and ontological resources. The manually constructed Cyc and WordNet provide a large amount of domain independent knowledge. However, they both cannot (and are not intended to) cope with specific domains and current events. As a result, many researchers in NLP have concentrated on developing methods for automatic harvesting of lexical relations. Except for a few works based on *clustering* methods [17,77], most of the proposed approaches rely on *pattern-based* methods originally pioneered in [39] to extract *isa* relations. This approach has been shown to scale well for large repositories of textual data, e.g. the Web [20], and was extended in [7] to account for part-of relation extraction. [35] build on top of [7] and employ machine learning techniques to disambiguate *part-of* patterns using word senses from WordNet.

The limitation of the pattern-based approach lies in the amount of supervision to be provided, i.e. in the form of manually created input patterns or manual sense annotations. Besides, the availability of large repositories of text such as the Web makes it impractical to define an exhaustive list of extraction patterns. Consequently, the last decade has seen a large body of work on *weakly supervised bootstrapping algorithms* for information extraction. These methods all work by taking a small set of seed examples of the target extraction as input and iteratively enlarging that set by discovering new extractions. Originally proposed in [95] for mutually bootstrapping both semantic class induction and extraction patterns, unsupervised bootstrapping has been successfully applied to many information extraction tasks including the automatic extraction of binary relations [12,1,76], facts [26,79], semantic class attributes [78] and instances [74], as well as the acquisition of knowledge for question answering [90,54]. Recent advances have additionally concentrated on developing fully unsupervised methods that require no seed examples, e.g. [6] propose a self-supervised classifier that automatically learns 'trustworthy' extractions.

Lexical relation harvesting systems do not necessarily produce formal semantic repositories. The emerging field of ontology learning tries to overcome these problems by learning (mostly) domain dependent ontologies from scratch. However, the generated ontologies are relatively small and the results rather poor – e.g., [21] report an F_1 measure of about 33 with regard to an existing ontology of less than 300 concepts. It seems to be more promising to ontologize automatically discovered semantic relations [80] or to extend existing resources such as Cyc [57] or WordNet [102]. The examples shown in these works, however, seem to indicate that the extension takes place mainly with respect to named entities, a task which is arguably not as difficult as creating a complete (domain-) ontology from scratch.

Another approach to building large knowledge bases relies on input by volunteers, i.e., on collaboration among the users of an ontology [93]. However, the current status of the *Open Mind* [19] and *MindPixel*[16] projects does indicate that they are largely academic enterprises. Similar to the *Semantic Web* [8], where users are supposed to explicitly define the semantics of the contents of web pages, they may be hindered by too high an entrance barrier. In contrast, Wikipedia and its categorization system feature a low entrance barrier achieving quality through collaboration based on a large user base. This is compatible with other successful approaches to mass annotation such as Web-based games, e.g. Verbosity [5], harvesting knowledge and annotated data via crowdsourcing [103,16], as well as community efforts to populate structured knowledge bases such as *Freebase*.[17]

4.2. Using Wikipedia as a resource for AI and NLP

Since Wikipedia has only existed since 2001 and has been considered a reliable source of information for an even shorter amount of time [34], researchers in NLP have just recently begun to work with its content or use it as a resource. Wikipedia has been successfully used for a multitude of AI and NLP applications. These include both preprocessing tasks such as named entity [15,25] and word sense disambiguation [60,83], text categorization [31], computing semantic similarity of texts [32,65], coreference resolution [85] and keyword extraction [24,66], as well as full-fledged, end-user applications such as question answering [3,4,55, inter alia], topic-driven multi-document summarization [68], text generation [98] and cross-lingual information retrieval [22].

Researchers working in information extraction have also recently begun to use Wikipedia as a resource for automatically deriving structured semantic content. [9] present the DBpedia system which generates hundreds of millions of RDF statements by extracting the attribute-value pairs contained in the *infoboxes* of the Wikipedia pages (i.e. the tables summarizing the most important attributes of the entity referred to by the page), e.g. the entry `capital=[[Berlin]]` from the GERMANY page. But while this project has achieved the creation of a huge database of structured knowledge, the focus has been

[16] http://www.mindpixel.com.
[17] http://www.freebase.com.

mostly on developing infoboxes' parsers as well as semi-automatically linking the knowledge base to external resources such as e.g. the CIA World Factbook, Freebase and OpenCyc, rather than developing open-domain knowledge extractors.

A proposal to tackle this problem is described within the context of the 'Intelligence in Wikipedia' project [109], originally developed in [110] and [111], which presents a framework based on synergistic interaction between knowledge acquisition methods and user edits: labeled data from Wikipedia are first used to learn knowledge extractors which enrich encyclopedic entries with new types of structural information. These enriched entries are then validated by humans to iteratively provide new training data for the original extractors. [110] show how to augment Wikipedia with automatically extracted information by developing a self-supervised attribute extraction system based on the infobox data. They propose to 'autonomously semantify' Wikipedia by (1) extracting new facts from its text via a cascade of Conditional Random Field models; (2) adding new hyperlinks to the articles' text by finding the target articles nouns refer to. The Kylin Ontology Generator (KOG) developed by Wu and Weld [111] is the work closest to ours. Their system builds a subsumption hierarchy of classes by combining Wikipedia infoboxes with WordNet using statistical-relational learning. Each infobox template, e.g. `Infobox Country` for countries, represents a class and the slots of the template are considered to be the attributes of the class. KOG uses Markov Logic Networks [94] in order to jointly predict both the subsumption relation between classes and their mapping to WordNet. The results from [111] are highly competitive with the ones presented in this paper, i.e. KOG achieves up to 98.8% precision and 92.5% recall for the task of detecting subsumption relations between pairs of infobox classes. However, it is difficult to draw a comparison, given that the evaluation is performed using 5-fold cross validation on a dataset of only 563 classification instances, which are semi-automatically generated from the manually-created mappings from Wikipedia articles to WordNet synsets found in DBpedia. In general, we note that all methods relying on infobox data such as [9,110,111] potentially suffer from a lower rate of coverage when compared with our approach, since many entries in Wikipedia – in particular, common nouns such as, for instance, AUTOMOBILE or KNIFE – do not have an infobox at all. Specifically in the case of KOG, while it represents a theoretically sounder methodology than [84] and [114] – as it is based on a general method to statistically learn complex relational structures – the lightweight heuristics from the latter two are straightforward to implement and show that, when given high quality semi-structured input as in the case of Wikipedia, large coverage semantic networks can be generated by using simple heuristics which capture conventions governing its public editorial base.

As indicated by our results on comparing our taxonomy with Cyc and WordNet (Section 3.2), the information contained in Wikipedia is of a different kind than the one contained in existing gold standard knowledge resources. Consequently, researchers have developed methods to integrate Wikipedia with other knowledge repositories. Previous efforts aimed at automatically linking Wikipedia pages to WordNet synsets include a model based on vector spaces [97], a supervised approach using keyword extraction [91] as well as a probabilistic formulation based on structured overlap [83]. [105] build the YAGO system by merging WordNet's taxonomic hierarchy with Wikipedia's category system, in order to populate the former with millions of instances based on the heuristics presented in Section 2.4.1. Similar to DBpedia, YAGO provides a very large knowledge repository with a logically clean model compatible with RDFS. However, the mapping of Wikipedia categories to WordNet synsets is performed by relying only on the so-called most frequent sense heuristic – i.e. mapping a category to the first WordNet sense of its label – a method which has been shown by [82] to be outperformed by a graph-based technique based on structural information.

One of Wikipedia's most interesting features is its multilinguality, namely the fact that different versions of Wikipedia in different languages can be linked by means of so-called *inter-language links*. Mining multilingual content from Wikipedia has been performed both in the contexts of DBpedia and WikiNet [70], whereas recently, [72] presented BabelNet, a wide-coverage, multilingual semantic network which integrates the relational structure of WordNet with the semi-structured information from Wikipedia.

5. Conclusions

In this paper we described our work on inducing a large scale domain independent taxonomy from the collaboratively generated encyclopedia Wikipedia. We first took the category system in Wikipedia as a conceptual network. Then, we labeled the relations between the categories as *isa* and *notisa* by applying syntax-based methods, using methods based on the connectivity of the network and applying lexico-syntactic patterns to large text corpora. The generated taxonomy comprises 209,919 nodes and 335,128 links between the nodes. Sizewise it ranks in between two other large knowledge resources used in AI and NLP, namely ResearchCyc and WordNet.

We compared our taxonomy with these two resources and a manually annotated set of concept pairs. This intrinsic evaluation showed results fully comparable to the state-of-the-art in taxonomy learning [21]. When evaluating the quality of our resource against a manually annotated dataset our methods were able to achieve high precision (i.e. up to 90.3%) for a satisfying recall (78.5%), yielding an F_1 measure of up to 84%. While syntax-based methods achieve very high precision and rather low recall, both connectivity-based methods and lexico-syntactic patterns increase recall considerably while decreasing precision as a trade-off. We additionally quantified how our resource compares with other large, manually-built knowledge resources such as Cyc and WordNet and found that by using our methods we are able to cover only a small portion of them. This suggests that the knowledge contained in our resource is indeed of a different kind than the one included in traditional knowledge bases. By using our lightweight heuristics, we are able to generate a resource where more than 99% of the semantic relations are novel, i.e. not found, in these gold standard resources: as a result, we are able to

provide an extra coverage of 28.2% for Cyc and of 211.6% for WordNet. Finally, we performed an extrinsic evaluation by computing semantic similarity where the Wikipedia-based taxonomy proved to be competitive with WordNet.

We showed that a taxonomy induced from a collaboratively constructed knowledge repository can achieve quality on par with manually created knowledge resources – and this at a fraction of the cost and time needed for creating them. The high quality of our taxonomy depends on – and benefits from – the time and labor volunteers spend on writing and maintaining articles in Wikipedia and structuring it via the categorization network. We believe that the Wikipedia model of collaborative editing provides us with already well-maintained knowledge. This semi-structured input in turn enables us to 'stake out a middle ground between manual and automatic knowledge acquisition'[18] to derive a high quality taxonomy.

Our work on deriving a taxonomy is the first step in creating a full-fledged ontology based on Wikipedia. We have already performed a few of the next steps, including the labeling of generic *notisa* relations with more specific ones such as *has-part*, *has-attribute*, etc. [69], the differentiation between concepts and instances in the taxonomy [114], as well as its mapping and integration with WordNet [82]. Furthermore, Wikipedia also has the potential to serve as a source for inducing a knowledge base in many languages. We have already shown that the methods described in this paper can easily be transferred to German [44].

Our methodology, albeit based on a set of heuristics manually developed for Wikipedia, can be applied to all wikis which have a categorization network and a certain amount of textual content (though the textual content could be replaced by corpora or search engine queries). Due to the popularity gained by the collaborative knowledge construction approach, many such resources exist – examples include other collaborative Web-based encyclopedias such as Baidu Baike[19] and wikis for specialized domains such as biological molecular structures (PDBWiki[20]) and securities (ValueWiki[21]) – and our methodology could be applied to create domain taxonomies (we leave such exploration for future work).

From a broader perspective, we argue that these large repositories of wide-coverage semantic knowledge can be expected to help overcome the knowledge bottleneck observed since the very dawn of AI research, and consequently open up a whole new world of possibilities to (again) develop knowledge-lean approaches for a variety of complex AI tasks. In accordance with this vision of knowledge-rich AI, our future work will concentrate on embedding machine readable knowledge within end-user applications such as automatic summarization, semantic information retrieval and statistical machine translation, all of which will benefit from the availability of knowledge induced from collaborative knowledge resources.

Acknowledgements

This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a PhD scholarship from KTF (09.003.2004). The second author has been partially supported by the European Commission through the CoSyne project (FP7-ICT-4-248531). We would like to thank Hinrich Schütze for his advice throughout the development of this work, Vivi Nastase for critically reading earlier versions of this paper, and AAAI-07 and AIJ reviewers for their helpful comments. We also appreciate feedback given by audiences at AAAI, the University of Zürich, the University of Stuttgart, the University of Mannheim, the University of Heidelberg, and Harvard University.

Appendix A. Head finding rules modification

Since we parse Wikipedia category labels which are mostly NP fragments rather than full sentences, we often need to recover from parsing errors and introduce the following two modifications to the head rules from [23, Appendix A] accordingly:

1. The rules for NPs are changed to search from *right to left* for the first child which is an NP. This is to recover from errors where flat NPs like

   `(NP(NNP Acorn)(NN operating)(NNS systems))`

   are output instead as non-base noun phrases like

   `(NP(NP(NNP Acorn))(NP(VBG operating)(NNS systems))).`

2. We constrain the output of the head finding algorithm to return a lexical head labeled either as a noun (`NN`, `NNP`, `NNPS`, `NNS`) or a 3rd person singular present verb (`VBZ`). This is to tolerate errors where plural noun heads have been wrongly identified as verbs as in e.g.:

   `(NP(NNP NBC)(NN network))(VP(VBZ shows)).`

   This way we return a noun even for category labels whose head is e.g. a gerund or present participle as in associated for `People/NNS associated/VBN with/IN religion/NN or/CC philosophy/NN`. We start by applying

[18] We are indebted to an anonymous AAAI-07 reviewer for such a perspicacious characterization of our work.
[19] http://baike.baidu.com.
[20] http://pdbwiki.org.
[21] http://valuewiki.wikia.com.

Collins' head rules and check that the part of speech of the lexical head belongs to one of these categories. If this is not the case we traverse the tree, find the first preterminal with such a label and return its child. If no such node can be found, we simply return the original lexical head.

In addition, for coordinated noun phrases such as $\langle \mathtt{NP} \rightarrow \mathtt{NP}_1\,\mathtt{CC}\,\mathtt{NP}_2 \rangle$ we find the head of both $\mathtt{NP}_1$ and $\mathtt{NP}_2$, rather than taking the leftmost coordinated NP as the head of the phrase. This way we return both nouns for NP coordinations, e.g. both buildings and infrastructure for `(NP(NNS Buildings)(CC and)(NN infrastructure))(PP in Japan)`.

Appendix B. Guidelines for the manual annotation of *isa* relations

Below is a list of pairs of words. For each pair $\langle a, b \rangle$, please assign one of the following relations:

IOF a is an INSTANCE-OF b corresponds to set membership. a must refer to a (unique) individual and b must refer to a set such that a is a member of b. Examples:
North Korea IOF country
Errol Morris IOF film director

ISA a ISA b corresponds to set inclusion. a and b must refer sets such that a is a (proper) subset of b. Examples:
physicists ISA scientists
football ISA sport

NOT If none of the above apply.

When annotating the pairs, please follow these guidelines:

1. Annotate the pairs $\langle a, b \rangle$ while answering the question: is a a (kind of/form of) b?
2. Reify the concepts, that is, consider abstract concepts as made up of material objects. E.g. disciplines are made up of things like publications, concrete theories, therefore:
   psychology ISA social science
   natural language processing ISA artificial intelligence
3. Disregard the number of the phrases (i.e. singular or plural), e.g. theoretical physicists ISA scientist, although scientist is singular and would denote a single individual rather than a set;
4. If a phrase has multiple senses, consider all senses of the phrase, e.g. school can refer to both the building and the institution so these pairs would be tagged as follows:
   school ISA building
   school ISA institution
   In other words, tag with a relation if there is *at least one* pair of senses of the two phrases which is in the relation. Note therefore that the same phrase can be in an ISA relation with two distinct sets (i.e. sets whose intersection is empty, as in the example above).

Appendix C. Downloads

The Wikipedia taxonomy is made available in RDFS format and can be downloaded at http://www.h-its.org/nlp/download/wikitaxonomy.php. Further details on the taxonomy generation process and data format can be found in [86].

References

[1] E. Agichtein, L. Gravano, Extracting relations from large plain text collections, in: Proceedings of the Fifth ACM Conference on Digital Libraries, San Antonio, TX, 2–7 June 2000, pp. 85–94.
[2] E. Agirre, A. Soroa, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A study on similarity and relatedness using distributional and WordNet-based approaches, in: Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, Col., 31 May–5 June 2009, pp. 19–27.
[3] D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, S. Schlobach, Using Wikipedia at the TREC QA track, in: Proceedings of the Thirteenth Text REtrieval Conference, Gaithersburg, MD, 16–19 November 2004.
[4] K. Ahn, J. Bos, J.R. Curran, D. Kor, M. Nissim, B. Webber, Question answering with QED at TREC-2005, in: Proceedings of the Fourteenth Text REtrieval Conference, Gaithersburg, MD, 15–18 November 2005.
[5] L. von Ahn, M. Kedia, M. Blum, Verbosity: A game for collecting common-sense facts, in: Proceedings of the Conference on Human Factors in Computing Systems, Montréal, Québec, Canada, 22–27 April 2006, pp. 75–78.
[6] 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.
[7] M. Berland, E. Charniak, Finding parts in very large corpora, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD, 20–26 June 1999, pp. 57–64.
[8] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Scientific American 284 (2001) 34–43.
[9] 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.
[10] J. Brank, M. Grobelnik, D. Mladenić, A survey of ontology evaluation techniques, in: Proceedings of the 8th International Multiconference Information Society, Ljubljana, Slovenia, 10–17 October 2005, pp. 166–169.

[11] T. Brants, TnT – A statistical Part-of-Speech tagger, in: Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle, Washington, 29 April–4 May 2000, pp. 224–231.
[12] S. Brin, Extracting patterns and relations from the World Wide Web, in: Proceedings of the Workshop on the Web and Databases at the 6th International Conference on Extending Database Technology, Valencia, Spain, 23–27 March 1998, pp. 172–183.
[13] A. Budanitsky, G. Hirst, Evaluating WordNet-based measures of semantic distance, Computational Linguistics 32 (2006) 13–47.
[14] P. Buitelaar, P. Cimiano, B. Magnini (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands, 2005.
[15] 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.
[16] C. Callison-Burch, M. Dredze (Eds.), Proceedings of the NAACL 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk, 2010.
[17] S.A. Caraballo, Automatic construction of a hypernym-labeled noun hierarchy from text, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD, 20–26 June 1999, pp. 120–126.
[18] J. Carletta, Assessing agreement on classification tasks: The kappa statistic, Computational Linguistics 22 (1996) 249–254.
[19] T. Chklovski, R. Mihalcea, Building a sense tagged corpus with Open Mind Word Expert, in: Proceedings of the ACL-02 Workshop on WSD: Recent Successes and Future Directions, pp. 116–122.
[20] P. Cimiano, S. Handschuh, S. Staab, Towards the self-annotating web, in: Proceedings of the 13th World Wide Web Conference, New York, NY, 17–22 May 2004, pp. 462–471.
[21] P. Cimiano, A. Pivk, L. Schmidt-Thieme, S. Staab, Learning taxonomic relations from heterogenous sources of evidence, in: P. Buitelaar, P. Cimiano, B. Magnini (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands, 2005, pp. 59–73.
[22] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, S. Staab, Explicit vs. latent concept models for cross-language information retrieval, in: Proceedings of the 21th International Joint Conference on Artificial Intelligence, Pasadena, CA, 14–17 July 2009, pp. 1513–1518.
[23] M. Collins, Head-driven statistical models for natural language parsing, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1999.
[24] A. Csomai, R. Mihalcea, Linking documents to encyclopedic knowledge, IEEE Intelligent Systems 23 (2008) 34–41.
[25] S. Cucerzan, Large-scale named entity disambiguation based on Wikipedia data, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, Prague, Czech Republic, 28–30 June 2007, pp. 708–716.
[26] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.M. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates, Web-scale information extraction in KnowItAll (Preliminary results), in: Proceedings of the 13th World Wide Web Conference, New York, NY, 17–22 May 2004, pp. 100–110.
[27] J. Euzenat, P. Shvaiko, Ontology Matching, Springer Verlag, Berlin, Germany, 2007.
[28] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998.
[29] J.R. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by Gibbs sampling, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, 25–30 June 2005, pp. 363–370.
[30] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (1971) 378–382.
[31] E. Gabrilovich, S. Markovitch, Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge, in: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, MA, 16–20 July 2006, pp. 1301–1306.
[32] 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.
[33] M. Geffet, I. Dagan, The distributional inclusion hypotheses and lexical entailment, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, 25–30 June 2005, pp. 107–114.
[34] J. Giles, Internet encyclopedias go head to head, Nature 438 (2005) 900–901.
[35] R. Girju, A. Badulescu, D. Moldovan, Automatic discovery of part-whole relations, Computational Linguistics 32 (2006) 83–135.
[36] N. Guarino, C. Welty, Evaluating ontologies with OntoClean, Communications of the ACM 45 (2002) 61–65.
[37] D. Harman, M. Liberman, TIPSTER Complete, LDC93T3A, Philadelphia, PE, Linguistic Data Consortium, 1993.
[38] S. Hassan, R. Mihalcea, Cross-lingual semantic relatedness using encyclopedic knowledge, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 July 2009, pp. 1192–1201.
[39] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proceedings of the 15th International Conference on Computational Linguistics, Nantes, France, 23–28 August 1992, pp. 539–545.
[40] T. Hughes, D. Ramage, Lexical semantic relatedness with random graph walks, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, Prague, Czech Republic, 28–30 June 2007, pp. 581–589.
[41] A. Islam, D. Inkpen, Second order co-occurrence PMI for determining the semantic similarity of words, in: Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 22–28 May 2006, pp. 1033–1038.
[42] M. Jarmasz, S. Szpakowicz, Roget's Thesaurus and semantic similarity, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria, 10–12 September 2003, pp. 212–219.
[43] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proceedings of the 10th International Conference on Research in Computational Linguistics, Taipei, Taiwan, 22–24 August 1997, pp. 19–33.
[44] L. Kassner, V. Nastase, M. Strube, Acquiring a taxonomy from the German Wikipedia, in: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May–1 June 2008.
[45] D. Klein, C.D. Manning, Fast exact inference with a factored model for natural language parsing, in: S. Becker, S. Thrun, K. Obermayer (Eds.), NIPS 2002, in: Advances in Neural Information Processing Systems (NIPS 2002), vol. 15, MIT Press, Cambridge, MA, 2003, pp. 3–10.
[46] S. Kripke, Naming and Necessity, Basil Blackwell, Oxford, 1980.
[47] T. Kudoh, Y. Matsumoto, Use of Support Vector Machines for chunk identification, in: Proceedings of the 4th Conference on Computational Natural Language Learning, Lisbon, Portugal, 13–14 September 2000, pp. 142–144.
[48] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, in: C. Fellbaum (Ed.), WordNet. An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998, pp. 265–283.
[49] L. Lee, Measures of distributional similarity, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD, 20–26 June 1999, pp. 25–31.
[50] D.B. Lenat, R.V. Guha, Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project, Addison–Wesley, Reading, MA, 1990.
[51] Y. Li, Z.A. Bandar, D. McLean, An approach for measuring semantic similarity between words using multiple information sources, IEEE Transactions on Knowledge and Data Engineering 15 (2003) 871–882.
[52] D. Lin, Automatic retrieval and clustering of similar words, in: Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada, 10–14 August 1998, pp. 768–774.
[53] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the 15th International Conference on Machine Learning, Madison, WI, 24–27 July 1998, pp. 296–304.
[54] L.V. Lita, J. Carbonell, Instance-based question answering: A data-driven approach, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004, pp. 396–403.

[55] K.K. Lo, W. Lam, Using semantic relations with world knowledge for question answering, in: Proceedings of the Fifteenth Text REtrieval Conference, Gaithersburg, MD, 14–17 November 2006.
[56] M.P. Marcus, B. Santorini, M.A. Marcinkiewicz, Building a large annotated corpus of English: The Penn treebank, Computational Linguistics 19 (1993) 313–330.
[57] C. Matuszek, M. Witbrock, R.C. Kahlert, J. Cabral, D. Schneider, P. Shah, D. Lenat, Searching for common sense: Populating Cyc from the web, in: Proceedings of the 20th National Conference on Artificial Intelligence, Pittsburgh, PE, 9–13 July 2005, pp. 1430–1435.
[58] M.T. Maybury (Ed.), New Directions in Question Answering, AAAI Press, Menlo Park, CA, 2004.
[59] J. McCarthy, Programs with common sense, in: Proceedings of the Teddington Conference on the Mechanization of Thought Processes, Her Majesty's Stationary Office, London, UK, 1959, pp. 75–91.
[60] R. Mihalcea, Using Wikipedia for automatic Word Sense Disambiguation, in: Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, NY, 22–27 April 2007, pp. 196–203.
[61] G.A. Miller, W.G. Charles, Contextual correlates of semantic similarity, Language and Cognitive Processes 6 (1991) 1–28.
[62] G.A. Miller, F. Hristea, Towards building a WordNet noun ontology, Revue Roumaine de Linguistique LI (2006) 405–413.
[63] G.A. Miller, F. Hristea, WordNet nouns: Classes and instances, Computational Linguistics 32 (2006) 1–3.
[64] G.A. Miller, C. Leacock, R. Tengi, R. Bunker, A semantic concordance, in: Proceedings of the 3rd DARPA Workshop on Human Language Technology, Plainsboro, NJ, 1993, pp. 303–308.
[65] 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, IL, 13 July 2008, pp. 25–30.
[66] D. Milne, I.H. Witten, Learning to link with Wikipedia, in: Proceedings of the ACM 17th Conference on Information and Knowledge Management, Napa Valley, CA, 26–30 October 2008, pp. 1046–1055.
[67] G. Minnen, J. Carroll, D. Pearce, Applied morphological processing of English, Natural Language Engineering 7 (2001) 207–223.
[68] V. Nastase, Topic-driven multi-document summarization with encyclopedic knowledge and activation spreading, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Waikiki, Honolulu, Hawaii, 25–27 October 2008, pp. 763–772.
[69] 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. 1219–1224.
[70] 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, Valletta, Malta, 19–21 May 2010.
[71] R. Navigli, Word Sense Disambiguation: A survey, ACM Computing Surveys 41 (2009) 1–69.
[72] 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.
[73] A. Oltramari, A. Gangemi, N. Guarino, C. Masolo, Restructuring WordNet's top-level: The OntoClean approach, in: Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain, 29–31 May 2002, pp. 17–26.
[74] M. Paşca, B. Van Durme, Weakly-supervised acquisition of open-domain classes and class attributes from Web documents and query logs, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, OH, 15–20 June 2008, pp. 19–27.
[75] P. Pantel, D. Lin, Discovering word senses from text, in: Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002, pp. 613–619.
[76] P. Pantel, M. Pennacchiotti, Espresso: Leveraging generic patterns for automatically harvesting semantic relations, 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. 113–120.
[77] P. Pantel, D. Ravichandran, Automatically labeling semantic classes, in: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Boston, MA, 2–7 May 2004, pp. 321–328.
[78] 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.
[79] M. Paşca, D. Lin, J. Bigham, A. Lifchits, A. Jain, Organizing and searching the world wide web of facts – Step one: The one-million fact extraction challenge, in: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, MA, 16–20 July 2006, pp. 1400–1405.
[80] M. Pennacchiotti, P. Pantel, Ontologizing semantic relations, 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. 793–800.
[81] G. Pirrò, N. Seco, Design, implementation and evaluation of a new similarity metric combining feature and intrinsic information content, in: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems: OTM 2008, in: Lecture Notes in Computer Science, vol. 5332, Springer, Heidelberg, 2008, pp. 1271–1288.
[82] S.P. Ponzetto, R. Navigli, Large-scale taxonomy mapping for restructuring and integrating Wikipedia, in: Proceedings of the 21th International Joint Conference on Artificial Intelligence, Pasadena, CA, 14–17 July 2009, pp. 2083–2088.
[83] S.P. Ponzetto, R. Navigli, Knowledge-rich Word Sense Disambiguation rivaling supervised system, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 1522–1531.
[84] 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, BC, Canada, 22–26 July 2007, pp. 1440–1445.
[85] S.P. Ponzetto, M. Strube, Knowledge derived from Wikipedia for computing semantic relatedness, Journal of Artificial Intelligence Research 30 (2007) 181–212.
[86] S.P. Ponzetto, M. Strube, WikiTaxonomy: A large scale knowledge resource, in: Proceedings of the 18th European Conference on Artificial Intelligence, Patras, Greece, 21–25 July 2008, pp. 751–752.
[87] M. Porter, An algorithm for suffix stripping, Program 14 (1980) 130–137.
[88] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric to semantic nets, IEEE Transactions on Systems, Man and Cybernetics 19 (1989) 17–30.
[89] A. Radford, Syntax: A Minimalist Introduction, Cambridge University Press, Cambridge, UK, 1997.
[90] D. Ravichandran, E. Hovy, Learning surface text patterns for a question answering system, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PE, 7–12 July 2002, pp. 41–47.
[91] N. Reiter, M. Hartung, A. Frank, A resource-poor approach for linking ontology classes to Wikipedia articles, in: J. Bos, R. Delmonte (Eds.), Semantics in Text Processing, in: Research in Computational Semantics, vol. 1, College Publications, London, England, 2008, pp. 381–387.
[92] 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, pp. 448–453.
[93] M. Richardson, P. Domingos, Building large knowledge bases by mass collaboration, in: Proceedings of the 2nd International Conference on Knowledge Capture, Sanibel Island, FL, 23–25 October 2003, pp. 129–137.
[94] M. Richardson, P. Domingos, Markov logic networks, Machine Learning 62 (2006) 107–136.

[95] E. Riloff, R. Jones, Learning dictionaries for information extraction by multi-level bootstrapping, in: Proceedings of the 16th National Conference on Artificial Intelligence, Orlando, FL, 18–22 July 1999, pp. 474–479.
[96] H. Rubenstein, J.B. Goodenough, Contextual correlates of synonymy, Communications of the ACM 8 (1965) 627–633.
[97] M. Ruiz-Casado, E. Alfonseca, P. Castells, Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets, in: Advances in Web Intelligence, in: Lecture Notes in Computer Science, vol. 3528, Springer Verlag, 2005, pp. 380–386.
[98] C. Sauper, R. Barzilay, Automatically generating Wikipedia articles: A structure-aware approach, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, 2–7 July 2009, pp. 208–216.
[99] L.K. Schubert, Turing's dream and the knowledge challenge, in: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, MA, 16–20 July 2006, pp. 1534–1538.
[100] H. Schütze, Automatic word sense discrimination, Computational Linguistics 24 (1998) 97–123.
[101] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in WordNet, in: Proceedings of the 16th European Conference on Artificial Intelligence, Valencia, Spain, 23–27 August 2004, pp. 1089–1090.
[102] 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.
[103] R. Snow, B. O'Connor, D. Jurafsky, A. Ng, Cheap and fast – But is it good? Evaluating non-expert annotations for natural language tasks, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Waikiki, Honolulu, Hawaii, 25–27 October 2008, pp. 254–263.
[104] M. Strube, S.P. Ponzetto, WikiRelate! Computing semantic relatedness using Wikipedia, in: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, MA, 16–20 July 2006, pp. 1419–1424.
[105] F.M. Suchanek, G. Kasneci, G. Weikum, YAGO: A large ontology from Wikipedia and WordNet, Journal of Web Semantics 6 (2008) 203–217.
[106] E.F. Tjong Kim Sang, S. Buchholz, Introduction to the CoNLL-2000 Shared Task: Chunking, in: Proceedings of the 4th Conference on Computational Natural Language Learning, Lisbon, Portugal, 13–14 September 2000, pp. 127–132.
[107] P.D. Turney, Similarity of semantic relations, Computational Linguistics 32 (2006) 379–416.
[108] J. Weeds, D. Weir, Co-occurrence retrieval: A flexible framework for lexical distributional similarity, Computational Linguistics 31 (2005) 439–475.
[109] D.S. Weld, F. Wu, E. Adar, S. Amershi, J. Fogarty, R. Hoffmann, K. Patel, M. Skinner, Intelligence in Wikipedia, in: Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence, Chicago, IL, 13–17 July 2008, pp. 1609–1614.
[110] F. Wu, D. Weld, Automatically semantifying Wikipedia, in: Proceedings of the ACM 16th Conference on Information and Knowledge Management, Lisbon, Portugal, 6–9 November 2007, pp. 41–50.
[111] 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, pp. 635–644.
[112] Z. Wu, M. Palmer, Verb semantics and lexical selection, in: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM, 27–30 June 1994, pp. 133–138.
[113] T. Zesch, C. Müller, I. Gurevych, Using Wiktionary for computing semantic relatedness, in: Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence, Chicago, IL, 13–17 July 2008, pp. 861–867.
[114] C. Zirn, V. Nastase, M. Strube, Distinguishing between instances and classes in the Wikipedia taxonomy, in: Proceedings of the 5th European Semantic Web Conference, Tenerife, Spain, 1–5 June 2008, pp. 376–387.