Artificial Intelligence 241 (2016) 66–102

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

MultiWiBi: The multilingual Wikipedia bitaxonomy project

Tiziano Flati *, Daniele Vannella, Tommaso Pasini, Roberto Navigli

Dipartimento di Informatica, Sapienza Università di Roma, Italy

ARTICLE INFO

Article history:
Received 14 May 2015
Received in revised form 10 August 2016
Accepted 15 August 2016
Available online 8 September 2016

Keywords:
Taxonomy extraction
Taxonomy induction
Machine learning
Natural language processing
Collaborative resources
Wikipedia

ABSTRACT

We present MultiWiBi, an approach to the automatic creation of two integrated taxonomies for Wikipedia pages and categories written in different languages. In order to create both taxonomies in an arbitrary language, we first build them in English and then project the two taxonomies to other languages automatically, without the help of language-specific resources or tools. The process crucially leverages a novel algorithm which exploits the information available in either one of the taxonomies to reinforce the creation of the other taxonomy. Our experiments show that the taxonomical information in MultiWiBi is characterized by a higher quality and coverage than state-of-the-art resources like DBpedia, YAGO, MENTA, WikiNet, LHD and WikiTaxonomy, also across languages. MultiWiBi is available online at http://wibitaxonomy.org/multiwibi.

1. Introduction

Over recent decades, knowledge has increasingly become the fundamental "lubricant" of our society. The Web today is by far the largest repository of knowledge in history and, as it gradually creeps into all aspects of our everyday lives, the ability to manipulate and control its knowledge concerns everyone, both the great mass of general users and researchers [1–3], and the big industry players [4,5] that are called upon to process and deliver information in an efficient and accurate manner. With the exception of rare cases, such as WordNet [6], for which knowledge has been manually encoded, the building of big repositories of knowledge requiring human intervention, and the extended development times this entails, has now become, unfortunately, no longer feasible. Such approaches simply cannot cope with the high volume of information, its heterogeneity and the need to have knowledge available in as many languages as possible. Nevertheless, having such large repositories of knowledge embedded into intelligent systems would positively impact several Natural Language Processing (NLP) tasks, such as question answering [7–10], machine reading [11], entity linking [12,13], information extraction [14,15] and automatic reasoning [16–18]. For example, traditional open-domain Question Answering systems might not be able to answer questions such as "Which architect designed the Shard London Bridge?". Even in the case where the answer is in effect provided within the text (e.g., "Renzo Piano designed many skyscrapers, among which is the Shard London Bridge"), additional information is usually needed at the semantic type level [19] (e.g., "Renzo Piano" is an architect and "Shard London Bridge" is a skyscraper). As a further demonstration of concept, Word Sense Disambiguation [20,21] might also receive a significant boost. Consider, for instance, the sentence "The woman lit a match": by combining the information that i) a match can either be a lighter or a contest (contribution from the taxonomy) with ii) the fact that only lighters are usually lit (contribution from the disambiguation system), it should be possible to achieve higher disambiguation performance. Because of the usefulness of taxonomies, researchers and industrial stakeholders have been seeking for decades to design

* Corresponding author.
E-mail addresses: flati@di.uniroma1.it (T. Flati), vannella@di.uniroma1.it (D. Vannella), pasini@di.uniroma1.it (T. Pasini), navigli@di.uniroma1.it (R. Navigli).

http://dx.doi.org/10.1016/j.artint.2016.08.004
0004-3702/

novel mechanisms capable of automatically extracting valuable information which is both broad and accurate at the same time. This goal has been pursued in many different ways. In the early days (but such approaches remain as alive as ever) there was the conviction and the desire to extract knowledge from linguistic textual repositories alone: methods based on distributional word cooccurrence and statistical analysis over linguistic patterns relied on nothing but free text. Given the limited size and source of the textual corpora on which these systems relied, however, even when they proved to be accurate, they failed to serve as true general domain data providers. As time went by, though, collaborative efforts started to sprout spontaneously, with the aim of developing true encyclopedic stores in which users could actively contribute by enhancing the resource with additional information. Wikipedia, started in 2001, is one of the biggest such movements and currently the most active one, with knowledge available in 294 languages at the time of writing. A real added value brought by Wikipedia is the possibility to enrich text with hyperlinks: this feature, combined with the availability of tabular information, makes it possible to extract semi-structured information on a large scale [22,23].

Over time, systems have targeted very different types of relation, sometimes very general or open-domain (TextRunner [24], ReVerb [25], and approaches at the syntactic-semantic interface like [26] and DefIE [14]), sometimes very specific or bound to a particular domain. Semantic relations encode a large number of linguistic aspects, spanning from general relatedness (as is the case for links across Wikipedia pages) up to specific types, such as hypernymy, holonymy, meronymy, and so on [27]. It became increasingly clear that hypernymy relations represented one of the most important types which could be used to boost current artificial intelligent systems. Starting from the eighties, a whole branch of research had focused on this type of semantic relation, with the pioneering work of Hearst [28] laying the foundation for the forthcoming literature. Hearst's patterns, however, were designed to be applicable only to free text and did not exploit any specific feature of the collaborative machine-readable repositories yet to come. One of the first attempts to extract is-a information from Wikipedia dates back to WikiTaxonomy [29] which transformed the noisy network of Wikipedia categories into a structured taxonomy of concepts. Subsequently, the example of WikiTaxonomy inspired a full line of research, including YAGO [30], WikiNet [31], MENTA [32], and more recently LHD [33].

Many of the above-mentioned taxonomies are focused on English and do not easily scale to dozens of languages, due to their dependency on English corpora and tools. Nonetheless, the multilinguality issue has been addressed in some of the existing taxonomies in a number of ways: DBpedia is based on manual mappings of Wikipedia infoboxes across languages to concepts in a small upper ontology, MENTA combines the taxonomical information from WordNet with information coming from several elements of Wikipedia, such as infoboxes and categories. LHD relies on a simple, though general, linking approach based on string-matching rules. However, the type of knowledge extracted by these approaches is either partial (is-a information is provided only for Wikipedia pages or Wikipedia categories), incomplete (lacking full coverage) or heterogeneous (i.e., not drawn from a shared, standard repository).

In contrast, in this paper we present an approach to the automatic creation of an integrated bitaxonomy of Wikipedia pages and categories for multiple languages, called MultiWiBi, which addresses all the above-mentioned issues:

1. First, it does not focus on Wikipedia pages or categories on their own but taxonomizes the two sides together, showing that they are mutually beneficial for inducing a wide-coverage and fine-grained integrated taxonomy. In particular, hypernyms are returned in a coherent manner, such that a Wikipedia page (category) has a Wikipedia page (category) as hypernym. The rationale behind this decision is that not only has Wikipedia been designed with these two separate but interconnected structures in mind, but also the nature of the two sides of Wikipedia is very different, in that pages encode concepts and named entities while categories group pages into coherent sets. For these reasons it would be unnatural to merge these two types of item.
2. Second, our method is able to taxonomize Wikipedias in any language, in a way that is fully independent of additional resources. At the core of our approach, in fact, lies the idea that the English version of Wikipedia can be linguistically exploited as a pivot to project the taxonomic information to any other language offered by Wikipedia, in order to produce a bitaxonomy in arbitrary languages. English has been chosen as pivoting language because i) the quality of other Wikipedia languages is not comparable to the English version (see Section 12); ii) it is the language with the provable highest-performance syntactic parser, thus leading to the best hypernym lemmas; iii) English is the language which features by far the greatest number of pages in Wikipedia.[1] Nonetheless, our method can potentially pivot on any language, not only English; we chose English as pivoting language because it is the language with the highest amount of data and, presumably, also the highest quality.
3. Third, we prove that our approach overcomes the language barrier by extracting not only hypernyms for projectable concepts, but also for those concepts which do not have an English counterpart and therefore represent culture-specific bits of knowledge.

2. Background and contributions

In this section we introduce some background, explain our key idea of a Wikipedia bitaxonomy and clarify our contributions. We also summarize the assumptions our work relies on and introduce notation.

[1] Note that Swedish is the second most popular language in Wikipedia with 2,885,256 pages (i.e., less than half of the English knowledge).

2.1. Background

This work stems from the insight that the biggest collaborative encyclopedia, namely Wikipedia, can be used for automatically deriving hypernymy (i.e., is-a) information for the entities and concepts described therein. To do so, we exploit the dual nature of Wikipedia wherein both pages and categories are provided. The following paragraphs explain the differences between pages and categories in more detail and present some core terminology.

Wikipedia pages A Wikipedia page provides an encyclopedic description of a single entity or concept; for example the page Albert Einstein reports all the relevant known facts about the physicist, while the page Person describes the concept of *person*. The text is semi-structured, in that the information is available in an XML-like language and the information is divided into sections and paragraphs. Whenever possible, pages also contain dates, tables, biographies, citations, as well as media files and images. What makes Wikipedia so interesting, though, is the fact that pages are *interlinked*, so that words in a page are associated with other pages in Wikipedia. The resulting hypertext can therefore be viewed as a semantic network of Wikipedia pages. This network is dense and heterogeneous and the links, although unlabeled, implicitly encode not only is-a relations, but also many other types of semantic relations (e.g., *born-in*, *located-in*, etc.), up to, as in the common case, more general relatedness. For example the Wikipedia page Enrico Fermi contains a link to Physicist (a link which brings the reader to the type of Enrico Fermi) but also to Nobel Prize in Physics (which, indeed, is strongly related to the physicist, but does not represent an is-a relation).

Besides regular pages, Wikipedia also provides so-called *redirections*. Redirections are special pages which act as HTML redirections to other Wikipedia pages. For example redirections to the Wikipedia page Singing include, among others, Singer and Vocalist, while redirections to the Wikipedia page Headphones include Stereo headphones, Head phones and Headphone, among others. As should be clear from the foregoing examples, redirections include misspellings of the final Wikipedia page, as well as concepts which are related to the final page but do not necessarily convey the same meaning.

Wikipedia categories Wikipedia categories, instead, are separate entities which group pages into broader classes; for instance Theoretical physicists is a category of Albert Einstein, while Person is categorized, among others, into Concepts in ethics. Notably, the two sides are intertwined, as pages are usually associated with multiple categories and a category acts as a bucket for similar pages (we call these page-category associations "cross-links", see below). However, note that Wikipedia categories do not always represent a proper categorization for that page: for example Albert Einstein is associated with Theoretical physicists, but also with 1879 births (which does not characterize the physicist in a particular manner, apart from that of having been born in 1879) and also to Institute for Advanced Study faculty which is indeed related to, but does not say much about, Albert Einstein as a physicist, or, at least, as a person. Wikipedia categories can thus be seen as a mixed-label graph where category nodes are connected by both is-a and relatedness relationships, without explicit distinction between the two.

Cross-links One of the core elements of this work is represented by cross-links. These links are special relations which connect pages to categories and vice versa. Thanks to this particular type of link, in fact, the hypernymy information extracted automatically for the page side of Wikipedia can be transferred to the category side and vice versa. For example, knowing that many pages linked to the category American singers have been assigned the page Singer as hypernym is an important hint for increasing the strength of is-a association between the categories American singers and Singers. Wikipedia pages are usually connected to Wikipedia categories, but this might also not hold: there are, in fact, categories with no pages associated and pages which still need to be categorised; for example, the Wikipedia page MacQuarrie has no categories associated with it, while the Wikipedia category Transport disasters in Yemen has no pages associated. In English this happens about 1.6% and 13.6% of the time for the page and the category sides, respectively, of the English Wikipedia.

Sense inventories A sense inventory represents a predefined set of concepts. Two major schools of thought emerge in the literature: the first option, which we adopt in our work, is to use all the Wikipedia pages, redirections and categories to form the sense inventory. Specifically, in our work hypernyms for pages are drawn from the set of pages and redirections, while hypernyms for categories are drawn only from the set of categories. The second option is to draw on several resources external to Wikipedia (e.g., WordNet or the DBpedia ontology) and this is the case for many alternative approaches, such as MENTA, YAGO, DBpedia, etc. (see Section 9). For example, YAGO returns person^1_n for the Wikipedia category People from Barcelona, while DBpedia returns *http://dbpedia.org/ontology/Settlement* for the Wikipedia page Barcelona.[2]

Our idea Despite the inherent asymmetry between pages and categories, our hunch is that the two sides of Wikipedia can fruitfully be exploited mutually and synergistically to extract information about the generalization of both pages and categories. In fact, as a by-product, not only does our system acquire hypernymy information for each page, but it also infers generalizations for Wikipedia categories, and vice versa. As the two sides are connected, the output of our system can be

[2] We use the sense notation of [20]: w^i_p is the i-th sense of w with part of speech p.

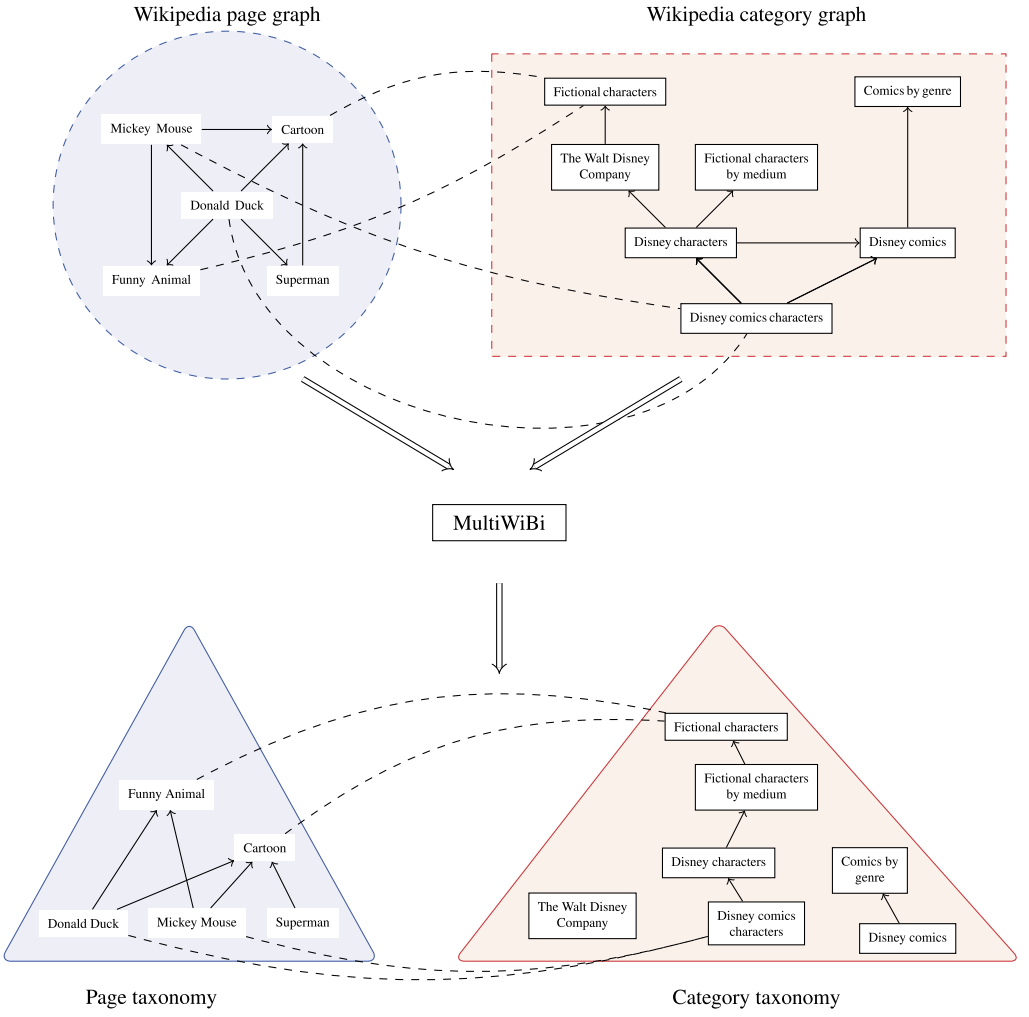


Fig. 1. Example of input and output of MultiWiBi.

seen as a pair of taxonomies, one taxonomy for the Wikipedia pages and one taxonomy for the Wikipedia categories, each taxonomy linked to the other via cross-links. We call this pair of taxonomies a *bitaxonomy*.

More formally, a bitaxonomy is a pair $B = (T_P, T_C)$ of taxonomies, where T_P is the taxonomy for the Wikipedia pages and T_C is the taxonomy for the Wikipedia categories. T_P (T_C) is defined as the set of hypernymy edges output by our algorithm for the page (category) side of Wikipedia, that is, $T_P = \{(p, p') \mid p, p' \in P$ and p is-a $p'\}$ ($T_C = \{(c, c') \mid c, c' \in C$ and c is-a $c'\}$), where P (C) is the set of all Wikipedia pages (categories). These edges represent the hypernymy information found by our algorithm; for instance, if the taxonomy for the Wikipedia pages contains the edge (Albert Einstein, Physicist) it means that we have automatically inferred that Albert Einstein is a physicist. In what follows, in the presence of an edge (p, p') we say that $p' = \text{isa}(p, h)$, where h is the hypernym lemma that generalizes p.

Fig. 1 provides a visual excerpt of the real input (top) and output (bottom) of our system. The page and the category sides are depicted with full lines and the cross-links drawn with dashed lines. For instance, consider the Wikipedia page Donald Duck, which in the Wikipedia page graph points to four pages, among which there are Mickey Mouse and Cartoon. Thanks to the application of MultiWiBi, Cartoon is promoted as hypernym of Donald Duck and as a result the first edge is discarded. On the other hand, the Wikipedia category Disney comics character which has two super categories (namely, Disney characters and Disney comics), is finally associated only with its hypernym category Disney characters, discarding the other super category.

The multilingual case Key to our approach is the idea of first acquiring a bitaxonomy of English and, then, *projecting* the information available in English *into* another language. English is seen as a pivot language which also allows us to infer facts known in English in other languages. Note that this does not mean that it will be possible to have all the English information transferred to all other languages. For the majority of the languages, in fact, the English Wikipedia contains many more concepts than Wikipedias in other languages taken individually. On the other hand, note also that the English

Wikipedia is far from being the union of the information found in all the other languages individually: each Wikipedia edition contains unique information which represents cultural concepts (such as food, dances, people native of a given country) often not available in English, such as the Italian page PACCHERI, a well-known type of pasta produced in Italy, or SAVARIN, a famous French sweet. With MultiWiBi we are also able to acquire this taxonomical information.

Data used in this paper All the data used in this paper for the examples and for the experimental setup is based on the English Wikipedia 2012 (for details, see Section 5). This was done to ensure a level playing field against alternative approaches that in general draw on a version of Wikipedia dating back to that year (see Sections 9 and 10).

2.2. Contributions

Our major contributions are the following:

- We provide a novel algorithm for inducing a taxonomy of Wikipedia pages and of Wikipedia categories. Starting from the raw dump of the English Wikipedia, this is performed in three steps. The first step produces a first taxonomy for the page side of Wikipedia; the second step, starting from a noisy category graph, iteratively isolates hypernyms for Wikipedia categories by discriminating is-a relations from general relations thanks to cross-links; the third step refines the bitaxonomy, improving the overall coverage, by solving structural flaws in the page and category graphs.
- We output a bitaxonomy, i.e., two taxonomies aligned to each other, meaning that concepts and entities in the page taxonomy are linked to categories in the category taxonomy, and vice versa. In contrast to several alternative approaches, which output hypernyms drawn from different sense inventories, the taxonomical information we output is homogeneous.
- Thanks to an advanced exploitation of the Wikipedia interlanguage links and link surface forms, we provide a probabilistic mechanism for obtaining translations of English hypernym lemmas in multiple languages. This is a crucial step in order to generalize MultiWiBi to languages other than English.
- We provide a method which is not bound to a particular language; in fact, MultiWiBi extracts a bitaxonomy on arbitrary Wikipedia languages, independently of additional resources. Notably, it also succeeds in covering those concepts which do not have an English counterpart, in marked contrast to all the alternatives which, instead, provide limited coverage.
- We have developed a novel approach for translating lemmas from any Wikipedia language to any other language in the encyclopedia. This is a statistical method which is able to provide for each lemma in a given language a distribution of translated lemmas in the target language, drawing only on information extracted from Wikipedia.
- As a result of this work, we release numerous gold standard datasets for pages and categories in four different languages that can be used as a benchmark for further experimentation and comparison purposes, for a total of 3850 ($1000 + 1000 + 256 + 155 + 436 + 232 + 140 + 500 + 131$) annotated items.

This work is an extension of the conference paper "Two Is Bigger (and Better) Than One: The Wikipedia Bitaxonomy Project" [34]. The main novelty of this journal article is a new method for the automatic extension to the multilingual case. In striking contrast to the English case, in fact, the procedure does not rely on any existing resource or tool external to Wikipedia, making MultiWiBi virtually independent and replicable on any new version of Wikipedia, in any language. We performed a whole range of additional experiments to demonstrate the accuracy of our approach in and of itself, and also in comparison with several other resources.

2.3. Paper organization

The paper is organized as follows. Sections 3–8 present the construction of a multilingual bitaxonomy. Section 9 presents the related work and introduces the main competitors we compare against, while the comparative evaluation is reported in Section 10. The extension to the multilingual case is explained in Section 11 and the corresponding multilingual evaluation is presented in Section 12. Section 13 finally draws conclusions.

3. A Wikipedia bitaxonomy for English

In order to induce the English Wikipedia bitaxonomy, i.e., a taxonomy of pages and categories, we proceed in 3 phases:

1. **Creation of the initial page taxonomy:** we first create a taxonomy for the Wikipedia pages by i) parsing the textual definitions of each page and extracting the hypernym lemma(s) and ii) by disambiguating each hypernym lemma according to the Wikipedia sense inventory.
2. **Creation of the bitaxonomy:** we leverage the hypernyms in the page taxonomy, together with their links to the corresponding categories, to induce a taxonomy over Wikipedia categories in an iterative way. At each iteration, the links in the page taxonomy are used to identify category hypernyms and, conversely, the new category hypernyms are used to identify more page hypernyms.

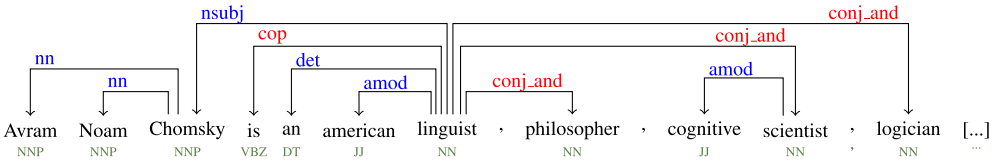


Fig. 2. The dependency tree for the Wikipedia definition of Noam Chomsky.

3. **Refinement of the bitaxonomy:** finally we employ structural heuristics to overcome inherent problems affecting certain classes of category and page hypernyms.

The output of our three-phase approach is a bitaxonomy of millions of pages and hundreds of thousands of categories for the English Wikipedia.

4. Phase 1: inducing the page taxonomy

The goal of the first phase is to induce a taxonomy of Wikipedia pages. Let P be the set of all Wikipedia pages and let $T_P = (P, E)$ be the directed graph of the page taxonomy whose nodes are pages and whose edge set E is initially empty ($E := \emptyset$). For each $p \in P$ our aim is to identify the most suitable generalization $p_h \in P$ so that we can create the edge (p, p_h) and add it to E. For instance, given the page Apple, which represents the fruit meaning of *apple*, we want to determine that its hypernym is Fruit and add the hypernym edge connecting the two pages (i.e., $E := E \cup \{(\text{Apple}, \text{Fruit})\}$). To do this, we proceed in two steps: i) a syntactic step, which extracts from a page's textual definition the lemma which best represents the hypernym for the page and ii) a semantic step, which identifies the most suitable sense for the lemma extracted in the syntactic step, according to our Wikipedia sense inventory.

4.1. Syntactic step: hypernym extraction

Given a page's textual definition, the aim of the syntactic step is to identify the lemma which best generalises the page's concept. To do this, for each page $p \in P$, we extract zero, one or more hypernym lemmas from the textual definition of p, that is, we output potentially ambiguous hypernyms for the page. The first assumption, which follows the Wikipedia guidelines[3] and is validated in the literature [35,36], is that the first sentence of each Wikipedia page p provides a textual definition for the concept represented by p. The second assumption we build upon is the idea that a lexical taxonomy can be obtained by extracting hypernyms from textual definitions. This idea dates back to the early 1970s [37], with further developments in the 1980s [38,39], the 1990s [40] and later [41–43].

To extract hypernym lemmas, we draw on the notion of copula, that is, "the relation between the complement of a copular verb and the copular verb itself".[4] Therefore, we apply the Stanford parser [44] to the definition of a page in order to extract all the dependency relations of the sentence. For example, given the definition of the page Noam Chomsky, i.e., "Avram Noam Chomsky is an American linguist, philosopher, cognitive scientist, logician, historian, political critic, and activist [...]", the Stanford parser outputs the set of dependencies shown in Fig. 2. The noun involved in the copula relation is *linguist* and thus it is taken as the page's hypernym lemma.

Finally, to capture multiple hypernyms, we iteratively follow the *conj_and* and *conj_or* relations starting from the initially extracted hypernym. For example, consider the definition of Noam Chomsky given above. Initially, the *linguist* hypernym is selected thanks to the copula relation; then, following the conjunction relations, also *philosopher*, *scientist*, *logician*, etc., are extracted as hypernyms. To understand the relevance of this step, consider that MultiWiBi succeeded in extracting more than one hypernym lemma for about 12% of all the English Wikipedia pages. We acknowledge that more sophisticated approaches like [35] or [45] could be applied, especially if we consider that this is a more light-weight solution than ours, which, instead, leverages a syntactic parser to extract the hypernym lemmas. Obtaining high coverage, though, is critical in our case, and we found that, in practice, our hypernym extraction approach is able to cover significantly more pages.

Handling special cases Words such as *one*, *kind*, *type*, etc., are often selected as hypernym lemmas. However, these are not always desirable lemmas, because they represent a class of objects. Consider, for instance, the definition of the page Tressette, "Tressette or Tresette is one of Italy's major national trick-taking card games, together with Scopa and Briscola"; the only copula relation extracted is between *is* and *one*, so the hypernym lemma which is extracted is *one*. Despite being correct in its most general sense, the latter, should be rejected in favour of *game*. Thus, to cope with this problem we use a specially designed class of nouns.[5] To avoid discarding valuable hypernyms, though, we handle only those cases in which the class term is followed by the preposition *of* (e.g., *"one of"*, *"a type of"*, etc.). Note that we identified the class terms

[3] See http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles.

[4] Cf. http://nlp.stanford.edu/software/dependencies_manual.pdf.

[5] *species*, *genus*, *one*, *list*, *term*, *part*, *form*, *type*, *collection*, *group*, *set*, *branch*, *order*, *class*, *family*, *series*, *name*, *style*, *variety*, *kind* and *pair*.

Frequency	Class term
96,706	species
54,970	one
53,128	genus
18,315	name

(a) Frequency of the 4 most frequent class terms across English glosses.

Level	# of lemmas	Percentage
Level 0	3,978,522	∼92.767%
Level 1	305,598	∼7.125%
Level 2	4,561	∼0.106%
Level 3	28	<0.001%

(b) Distribution of hypernym lemma nesting level.

Fig. 3. Statistics on nested hypernym lemmas.

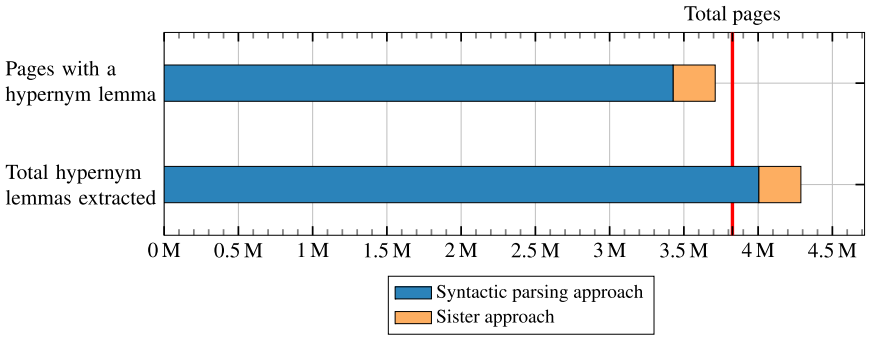


Fig. 4. Coverage of Wikipedia pages during the hypernym lemma extraction step.

independently of the underlying data distribution (so that these are neither data-oriented nor language-specific). Hence, when this occurs we replace the class term x with the noun n involved in the dependency relation $prep\_of(x, n)$. In the previous example, since the latter is involved in the dependency relation $prep\_of(\text{one}, \text{games})$, the lemma *one* is replaced with the more concrete and informative hypernym lemma *game*. Furthermore, we also cope with the problem of nested special nouns (e.g., as in the gloss '*[...] is one of a set of [...]*') by recursively applying the procedure explained above. For example, given the page HARMONIC MEAN and its definition "In mathematics, the harmonic mean is one of several kinds of average." the system is able to extract the lemma *average* as hypernym. The most frequent nested noun is *species*, as shown in Table 3a, followed by *one* and *genus*. Out of the total number of hypernym lemmas extracted, about 92% have nesting level 0, most of the remaining (around 7%) lemmas are nested at level 1 and only a negligible percentage have nesting level of 2 or 3, as can be seen from Table 3b.

Filling the gaps of the syntactic parser: the sister approach By analyzing the coverage of the lemmas extracted thanks to the syntactic parser, we found that for 400,286 of the English pages (about 10% of the total) no hypernym lemma could be extracted. We considered a sample of 100 pages for which the syntactic parser could not extract the hypernym. Out of the corresponding 100 textual definitions, we found that only 4 definitions contained the hypernym lemma in the copula relation, representing cases for which the syntactic parser failed to parse correctly, 8 were unrecognized disambiguation pages which we were not able to remove from the total list of pages, 18 contained the hypernym lemma expressed through relations other than copula (e.g., in the definition "Arthur Walworth is most noted as a biographer of Woodrow Wilson." the word *biographer* is only involved in a *prep_as* dependency relation and not in a *copula* relation), and 70 were ill-formed definitions which do not clearly define the concept represented by the Wikipedia page and briefly describe its history, its role in the world, or leave the generalization implicit. The latter class of ill-formed definitions include, for example, AUDI which is defined as "AUDI Aktiengesellschaft and its subsidiaries design, engineer, manufacture and distribute automobiles and motorcycles under the Audi, Ducati and Lamborghini brands".[6] In order to cover the pages affected by these problems, we applied an algorithm which is able to assign a hypernym lemma by inducing the information from other pages. Given a page p, the algorithm considers the so-called *sister pages* of p, i.e., pages which share with p at least one category, for which the syntactic parser has been able to provide a hypernym lemma. The algorithm then builds a distribution of such hypernym lemmas and selects the one which overlaps most with the lemmas of p's Wikipedia categories. For the above page, for instance, the selected hypernym lemma is *manufacturer* which overlaps with the AUDI categories MOTOR VEHICLE MANUFACTURERS OF GERMANY and CAR MANUFACTURERS OF GERMANY, among others. Thanks to the sister approach we are able to recover a hypernym lemma for about 70% of the pages which could not be covered by the syntactic parsing approach.

To visually grasp the impact of the application of the above two approaches, we report in Fig. 4 the coverage of Wikipedia pages. The bar on top represents the number of pages which have at least one hypernym lemma extracted thanks to the syntactic parsing (dark colour) and the sister approaches (light colour); as can be seen, 3,712,201 pages are covered overall, that is, approximately 97% of the total number of Wikipedia pages. The second bar represents, instead, the overall number

[6] Note that the definition for this page was improved in 2014 into "Audi AG is a German automobile manufacturer that designs, engineers, produces, markets and distributes luxury automobiles.", meaning that our syntax-based approach would have been effective.

of hypernym lemmas extracted with the two approaches: recall that our hypernym extraction procedure potentially extracts multiple hypernyms from a single definition, so that the total number of hypernym lemmas extracted can be much higher than the total number of Wikipedia pages (vertical red line in the figure); in fact, for the 3,712,201 Wikipedia pages covered, 4,288,709 hypernym lemmas were extracted in total.

4.2. Semantic step: hypernym disambiguation

Since our aim is to connect pairs of pages via hypernym relations, our second step consists of disambiguating the obtained hypernym lemmas of page p with their most suitable senses. For instance, given *fruit* as the hypernym for APPLE we would like to link APPLE to the page FRUIT as opposed to, e.g., FRUIT (BAND) or FRUIT (ALBUM). As explained in Section 2.1, going beyond previous work [36,46], as inventory for a given lemma we consider the set of pages and redirections whose main title is the lemma itself, except for the sense specification in parentheses. It is very important to do this because frequent concepts, such as SINGER, PHILOSOPHER, and VOLLEYBALL PLAYER, lack their own pages in Wikipedia. Even if, on the one hand, Wikipedians are continually mitigating this issue over time (e.g., PHILOSOPHER has its own Wikipedia page in 2014), on the other hand, this kind of problem is likely to persist in the future (e.g., SINGER does not exist as an independent page yet).

In order to disambiguate hypernym lemmas extracted in the previous step, we apply a battery of hypernym linkers, which output the most suitable sense for a given lemma, combined with two procedures which limit sense drifts during the application of the linkers.

4.2.1. Hypernym linkers

To disambiguate hypernym lemmas, we exploit the structural features of Wikipedia through a pipeline of hypernym linkers $\mathcal{L} = \{L_i\}$, applied in cascade order. We start with the set of page-hypernym pairs $H = \{(p, h)\}$ as obtained from the syntactic step. The successful application of a linker to a pair $(p, h) \in H$ yields a page p_h as the most suitable sense of h, resulting in setting $isa(p, h) = p_h$. At step i, the i-th linker $L_i \in \mathcal{L}$ is applied to H and all the hypernyms which the linker could disambiguate are removed from H. This prevents lower-precision linkers from overriding decisions taken by more accurate ones (cf. Section 5.3). Hypernym linkers are applied in the same order with which they are presented (for details, see Section 5.3).

In what follows we denote with $p \xrightarrow{h} p_h$ the fact that the definition of a Wikipedia page p contains an occurrence of h linked to page p_h. Note that we do not constrain p_h to be necessarily a sense of h and let it represent an arbitrary Wikipedia page; for instance, we allow the hypernym lemma *person* to be linked to the Wikipedia page INDIVIDUAL which is not a sense of *person* in Wikipedia.

Category linker Given the set $W \subset P$ of Wikipedia pages which have at least one category in common with p, we select the majority sense of h, if there is one, as hyperlinked across all the definitions of pages in W:

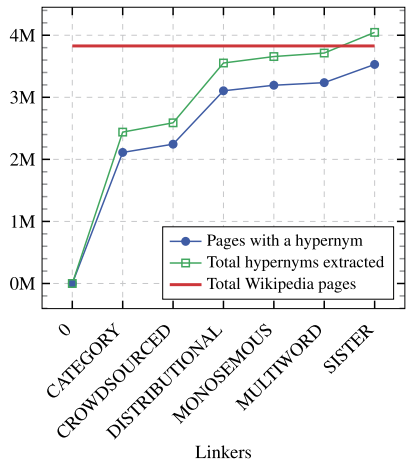
$$isa(p, h) = \arg\max_{p_h} \sum_{p' \in W} 1(p' \xrightarrow{h} p_h)$$

where $1(p' \xrightarrow{h} p_h)$ is the characteristic function which equals 1 if h is linked to p_h in page p', 0 otherwise. For example, the linker sets $isa(\text{EGGPLANT}, \textit{plant}) = \text{PLANT}$ because most of the pages associated with TROPICAL FRUIT, a category of EGGPLANT, contain in their definitions the term *plant* linked to the PLANT page.

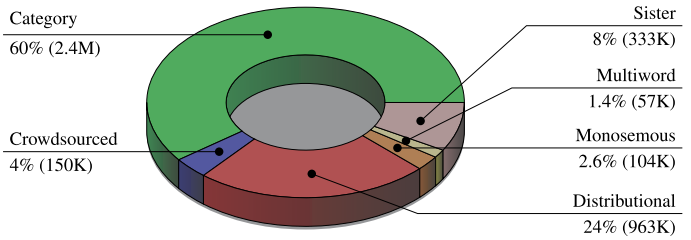
Crowdsourced linker If $p \xrightarrow{h} p_h$, i.e., the hypernym h is found to have been manually linked to p_h in p by Wikipedians, we assign $isa(p, h) = p_h$. For example, because *capital* was linked in the BRUSSELS page definition to CAPITAL CITY, we set $isa(\text{BRUSSELS}, \textit{capital}) = \text{CAPITAL CITY}$.

Distributional linker This linker provides a distributional approach to hypernym disambiguation. We represent the textual definition of page p as a distributional vector $\vec{v}_p$ whose components are all the English lemmas in Wikipedia (we consider nouns, adjectives, adverbs and verbs). The value of each component is the occurrence count of the corresponding content word in the definition of p. We perform no compounding, discard lemmas whose length is equal to 1 and discard the verb *to be* because it is contained in almost all Wikipedia definitions. The goal of this approach is to find the best link for hypernym h of p among the pages h is linked to, across the whole set of definitions in Wikipedia. Formally, for each p_h such that h is linked to p_h in some definition, we define the set of pages $P(p_h)$ whose definitions contain a link to p_h, i.e., $P(p_h) = \{p' \in P | p' \xrightarrow{h} p_h\}$. We then build a distributional vector $\vec{v}_{p'}$ for each $p' \in P(p_h)$ as explained above and create an aggregate vector $\vec{v}_{p_h} = \sum_{p'} \vec{v}_{p'}$. For discriminating among vectors, we also remove the target lemma from $\vec{v}_{p_h}$. Finally, we determine the similarity of p to each p_h by calculating the dot product between the two vectors $sim(p, p_h) = \vec{v}_p \cdot \vec{v}_{p_h}$. If $sim(p, p_h) > 0$ for any p_h we perform the following association:

$$isa(p, h) = \arg\max_{p_h} sim(p, p_h)$$



(a) Coverage of pages as linkers are applied.



(b) Distribution of disambiguated hypernyms by linker (displayed from top-left, counter-clockwise).

Fig. 5. Absolute number and distribution of hypernyms disambiguated by our hypernym linkers.

For example, consider the Wikipedia page ARISTOTLE and its hypernym lemma *teacher*. Among all Wikipedia textual definitions in which it occurs, teacher has been linked to several senses, among which there are TEACHER and PIANO TEACHER. The vectors for the starting page ARISTOTLE and these two senses are shown below:

$$\vec{v}_{\text{ARISTOTLE}} = (\text{student:1}, \text{philosopher:1}, \ldots, \text{polymath:1})$$
$$\vec{v}_{\text{TEACHER}} = (\text{student:30}, \text{philosopher:14}, \ldots, \text{polymath:1})$$
$$\vec{v}_{\text{PIANO TEACHER}} = (\text{pianist:1}, \text{virtuoso:1}, \ldots, \text{composer:1})$$

The similarities between the vector for the starting page and the vectors of the two senses are thus:

$$similarity(\text{ARISTOTLE}, \text{TEACHER}) = 1 \times 30 + 1 \times 14 + 1 \times 1 = 45.0$$
$$similarity(\text{ARISTOTLE}, \text{PIANO TEACHER}) = 0$$

In the first case the two vectors share lemmas such as *student* and *philosopher*, so their similarity is greater than zero, while in the second case the two vectors have no word in common. Hence, since TEACHER is the sense of *teacher* which maximises the similarity with ARISTOTLE, this linker sets $isa(\text{ARISTOTLE}, \textit{teacher}) = \text{TEACHER}$.

Monosemous linker If h is monosemous in Wikipedia (i.e., there is only a single sense p_h for that lemma), link it to its only sense by setting $isa(p, h) = p_h$. For example, the syntactic step extracted the hypernym lemma *businessperson* from the definition of MERCHANT and, since it is unambiguous, we link it to BUSINESSPERSON.

Multiword linker If $p \xrightarrow{m} p_h$ and m is a multiword expression containing the lemma h as one of its words, set $isa(p, h) = p_h$. For example, we set $isa(\text{AREA 51}, \textit{base}) = \text{MILITARY BASE}$, because the multiword expression *military base* is linked to MILITARY BASE in the definition of AREA 51.

Sister linker Finally, given the set $W \subset P$ of Wikipedia pages which have at least one category in common with p and share the hypernym lemma with it, we select the most frequent hypernym across these. For example we determine $isa(\text{GUITARIST}, \textit{person}) = \text{PERSON}$, thanks to the fact that seven pages (e.g., COMPOSER and DISC JOCKEY) have PERSON as common hypernym and share the category OCCUPATIONS IN MUSIC with the starting page GUITARIST.

Fig. 5a plots the coverage of the Wikipedia pages as hypernym linkers are applied in the presented order. Two lines are shown: the blue line plots the number of pages with at least one hypernym, the green line shows the number of total hypernyms found up to a certain phase. Again, since MultiWiBi extracts more than one hypernym lemma for any given page, the total number of hypernyms is higher than the total number of Wikipedia pages. Fig. 5b also shows the absolute number of the hypernym links found and the corresponding relative ratios. The first two heuristics provide, alone, about two thirds of the total hypernyms contained in the Wikipedia page taxonomy, while the others increasingly disambiguate hypernym lemmas, until 4,046,411 total hypernyms are found for 3,529,647 Wikipedia pages, covering more than 92% of the total Wikipedia pages.

In order to limit the potential noise introduced by the linkers, after the application of each linker, we apply two special modules whose aim is to preserve quality during the linking pipeline and detect possible shifts in meaning.

```
Input: Strings s1, s2
Output: true if (s1, s2) is a semantic shift, false otherwise
 1: s1 ← normalize(s1);
 2: s2 ← normalize(s2);
 3: h1 ← get_head(s1);
 4: h2 ← get_head(s2);
 5: r ← shift_test(h1, h2);
 6: if r != undef then return r;
 7: t1 ← get_last_token(s1);
 8: t2 ← get_last_token(s2);
 9: r ← shift_test(t1, t2);
10: if r != undef then return r;
11: return false;
```

(a) The Semantic Shift Recognizer (SSR) Algorithm.

Freq.	s_1	s_2
47299	footballer	Association football
13321	painter	Painting
4871	singer	Singing
3671	fencer	Fencing
3388	boxer	Boxing
3360	athlete	Athletics (sport)

(b) Excerpt of the most frequent semantic shifts recognized by the SSR module.

Fig. 6. The SSR algorithm (a) and an excerpt of the most frequent shifts returned by the algorithm (b).

4.2.2. Preserving meaning between hypernym lemmas and hypernym senses

As a result of the application of the entire linking pipeline we obtain a large number of disambiguated hypernym lemmas. However, a non-negligible number of disambiguated hypernyms suffer from the problem of *semantic shift*. This phenomenon occurs when a page's hypernym lemma is linked to another page which is closely related to it but is not a sense for the hypernym at hand. Consider for example the textual definition "Heinrich von Tenner was an Austrian fencer$_{\text{Fencing}}$." in which the hypernym term *fencer* is linked to the page Fencing. This is not something inappropriate *per se*, but instead reflects a very common phenomenon which consists of annotating text with the domain rather than the word sense (i.e., Fencing can be considered as the topic or domain usually associated with *fencer* but not a sense of it).[7] Furthermore, this phenomenon involves different kinds of linguistic aspects, such as gender differentiation (e.g., *actress/actor*), distinction between an activity and the associated role (e.g., *singing/singer*, *painting/painter*), etc. In addition, it is important to point out that links in Wikipedia can be pages as well as redirections. As such, redirections include mispellings of the final Wikipedia page as well as concepts which are related to the final page but do not necessarily convey the same meaning. Note that redirections do not have any text associated with them, so it becomes hard to define solid linguistic rules which measure the relationship between a redirection and the target page.

Lemma preserver (LP) As a first simple attempt to cope with the semantic shift phenomenon we apply a procedure that we have called *Lemma preserver*. Whenever any of the linkers presented in Section 4.2.1 outputs a Wikipedia page p as the disambiguation of the hypernym lemma l this routine tries to preserve the meaning of l by looking at the possible redirections of p. More specifically, it detects cases in which the hypernym lemma l and its disambiguation p do not match. In these cases, a new candidate is searched across the redirections of p. A redirection p' of p is selected as the new disambiguation for l if the title of p' and l match (ignoring case, such as *linguist* and Linguist) or l is contained in the title of p' (such as the lemma *wrestler* in the title Professional wrestler). For example, the Category linker disambiguated the hypernym lemma *linguist* of the Wikipedia page Noam Chomsky with Linguistics. Of course, as explained above, this is a closely related page, but should not be considered a valid disambiguation for the hypernym lemma extracted. As a result of the LP procedure, instead, Linguistics is replaced by Linguist, which is a redirection to the former. This is a very frequent and important action to take; consider that about 17% of the links output by the first (i.e., category) linker are replaced by the lemma preserver.

Semantic shift recognizer (SSR) A second, more general and linguistically-bound strategy is represented by a module called *Semantic Shift Recognizer* (SSR) which, on the basis of English hand-crafted rules, automatically discriminates is-a relations from semantic shifts.

We now describe in detail the mechanism behind the SSR module, whose pseudocode is reported in Fig. 6a. To recognize if there is a semantic shift between the two concepts represented by two strings s_1 and s_2 we first normalize them (lines 1–2) so that i) all words within parentheses are removed (e.g. *Person (sport)* is cut to *Person*), ii) s_1 and s_2 are lowercased, iii) acronyms are normalized (e.g., $s_1 = ep$ and $s_2 = extended\ play$ get normalized into ep), and iv) several separators are normalized with a space (e.g., *business_man* and *business-man* get both normalized to *business man*).

The core of the SSR module consists of isolating the heads of the two strings (lines 3–4) and subsequently applying the following string matching rule (function *shift_ test* at line 5 of the algorithm in Fig. 6a):

> **shift test**: extract the stems r_1, r_2 of the two heads as well as their remaining suffixes x_1, x_2 (i.e., $h_1 = r_1x_1$ and $h_2 = r_2x_2$). If r_1 and r_2 do not coincide, suspend the judgement by returning *undef*. If they coincide and x_1, x_2 differ, return

[7] In fact, these edges bear very important semantics and could in principle be left in the taxonomy with an opaque RELATED-TO label. As for now, we have decided to discard them in order to provide a cleaner and more coherent taxonomy for the Wikipedia page side.

true, and *false* otherwise. For example, $s_1 = singer$ and $s_2 = singing$ is considered to be a shift because the two strings share the stem *sing*, but have different suffixes (i.e., *er/ing*); instead, $s_1 = plant$ and $s_2 = plant$ is not considered to be a shift because the two stems and the two (empty) suffixes coincide.

The above test has been designed to return *true* only for those cases which are likely to be semantic shifts. Therefore, in case the previous test does not detect any semantic shift (i.e., the variable r at line 6 is undefined), the shift test above is repeated on the last two respective tokens of s_1 and s_2 (lines 7–8). If no semantic shift has been detected yet, the SSR assumes that no shift occurs. For instance, the pair $s_1 = human$ and $s_2 = person$ is not detected as a shift. Fig. 6b reports the most frequent semantic shifts detected by the Semantic Shift Recognizer module; as can be seen, most of them consist of topic drifts.

5. Page taxonomy evaluation

For ease of reading, we describe experiments and results step by step. Thus, as soon as the taxonomy creation process has been described, we also provide the evaluation of its output. In this section we also introduce the measures and the datasets used to evaluate our techniques. All our experiments are based on the 2012 edition of Wikipedia in 4 different languages.[8] This was almost a forced choice, since nearly all the available taxonomic resources refer to a version of Wikipedia which dates back to 2012.[9]

5.1. Evaluation measures

Unfortunately, measuring the quality of a taxonomy is not a trivial task. Currently there is still no agreement on how to perform it [43]. On the one hand, performing a complete validation of all the edges contained in a taxonomy is unattainable, on the other hand, even when a smaller sample of the edges is validated, it is not clear which measures to use for a correct and fair evaluation. For these reasons, we defined three measures that take values between 0 and 1 and try to characterise three different dimensions of quality: precision, recall and coverage. Note that these measures are used to evaluate the page taxonomy as well as the category taxonomy (see Section 8.3).

We used macro precision, i.e. the average ratio of correct items to the total number of items returned. This measure is designed to count the average correctness of the information provided for each single node covered by the taxonomy. For example if we have only two nodes in our taxonomy, the first with one correct edge and the second with 19 edges, all of which are wrong, we estimate a precision of 50%. Note that if the taxonomy contains only one correct edge, its precision is 1; this means that this measure alone cannot truly grasp the overall quality of the taxonomy.

Given the wide range of possible answers that could be considered to be correct, standard recall across resources could not be calculated. We therefore defined a variant of recall (R*) as the ratio of the items for which the system outputs at least one correct answer. In order to calculate precision and recall, for each resource we therefore manually marked each hypernym returned as correct or not.

Another useful measure which acts as the upper bound to precision and recall is coverage, defined as the fraction of items for which at least one answer is returned, independently of its (or their) correctness; the rationale behind this measure is to have a rough idea of the amount of information provided by the taxonomy by considering the number of items covered.

5.2. Page dataset

To evaluate the quality of our page taxonomy we randomly sampled 1,000 Wikipedia pages. For each page we provided: i) a list of suitable hypernym lemmas for the page, mainly selected from its definition; ii) for each lemma the correct hypernym page(s).

5.3. Hypernym linker order

The optimal order of application of the above linkers is the same as that presented in Section 4.2.1. It was established by selecting the combination, among all possible permutations, which maximized macro precision on a tuning set of 100 randomly sampled Wikipedia pages, disjoint from our page dataset. The Sister Linker, instead, is employed as the last one, since it exploits hypernym links found by previous linkers.

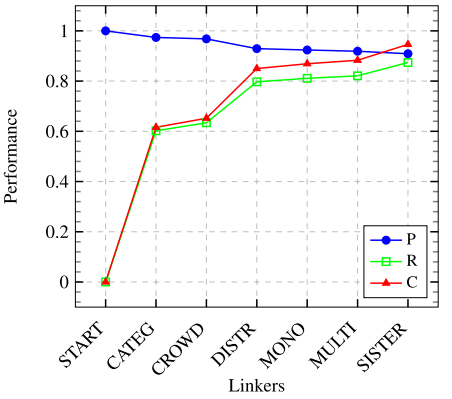
5.4. Results

Results, both at lemma and sense level, are reported in Fig. 7a. The first two lines show performance when considering the quality of the extraction of the ambiguous hypernyms. As can be seen, at lemma level, the configuration that exploits

[8] The exact timestamps for the different languages are: 2012/10/01 for English, 2012/10/07 for French, 2012/10/12 for Italian and 2012/09/27 for Spanish.
[9] All but MENTA, which, instead, extracts data from Wikipedia 2010. See Table 1 for details.

	P	R*	C	# items
Lemma (syntactic)	93.80	89.80	94.50	1,000
Lemma (syntactic + sisters)	**93.17**	**93.30**	**98.90**	
Sense (simple)	83.20	81.80	96.00	1,000
Sense (only LP)	85.89	84.20	96.00	
Sense (only SSR)	89.68	85.90	94.20	
Sense (LP + SSR)	**90.89**	**87.40**	**94.60**	

(a) Page taxonomy performance at lemma- and sense- level. Performance related to the chosen configurations are shown in bold.



(b) Page taxonomy performance as linkers are applied.

Fig. 7. Page taxonomy performance.

the sister pages in combination with the simple syntactic extraction phase produces a modest increment in both coverage and recall, to little detriment of precision. The final configuration is shown in bold (syntactic + sisters). The following lines in the table show results after i) the disambiguation step (vanilla), ii) when the LP module is used after the application of the linkers (only LP), iii) when only the SSR module is applied after the application of the linkers, and finally iv) when both modules are applied (LP + SSR). As can be seen, applying the LP module does not alter coverage, because this module does not filter out any linker's answer. In contrast, both precision and recall are boosted modestly. When the SSR module is applied, instead, coverage decreases to 94.20, but precision and recall receive an important increase. Finally, when the two modules are applied, the peak is reached for precision and recall, while coverage is somewhat in between the vanilla setting and the more restrictive one when using the LP or the SSR individually. In bold we have highlighted the final, chosen configuration, that is, the combination of the linkers, the LP and the SSR procedures. Fig. 7b shows the performance in terms of precision, recall and coverage as the hypernym linkers are applied. Precision, generally very high, has a positive spike after the application of the first linker and then decreases slowly as subsequent linkers are chained, measuring around 90%. Recall and coverage consistently increase when more linkers are considered, performing on a par with precision.

6. Phase 2: inducing the bitaxonomy

The Wikipedia page taxonomy built in Section 4 will now serve as a stable, pivotal input to the second phase, the aim of which is to build our bitaxonomy, that is, a taxonomy of pages aligned to a taxonomy of categories. Our key idea is that the generalization information available in each of the two partial taxonomies is mutually beneficial. We implement this idea by exploiting one taxonomy to add new hypernymy relations to the other, and vice versa, in an iterative way, until a fixed point is reached. The final output of this phase is, on the one hand, a page taxonomy augmented with additional hypernymy relations and, on the other hand, a category taxonomy which is built starting from the noisy category graph (see Section 2).

6.1. The bitaxonomy algorithm

We now describe in detail the bitaxonomy algorithm. To help the reader throughout the explanation, we will support the presentation by reference to Fig. 8, which shows the steps in which the algorithm is divided. We identified four steps (each step is represented by a number enclosed in a square) named as follows: Item switch (step 1), Taxonomy climbing (step 2), Candidate discovery (step 3) and Sanity check (step 4). Before going into the details of each single step, let us explain how the data structures are initialised.

6.2. Initialization

Our initial bitaxonomy $B = (T_P, T_C)$ is a pair consisting of the page taxonomy $T_P = (P, E)$, as obtained in Section 4, and the category taxonomy $T_C = (C, \mathbb{1}_{super})$, where C contains all the Wikipedia categories and $\mathbb{1}_{super} := \{e = (u, v) \in E(CG) \mid deg^+(u) = 1\}$, where CG is the Wikipedia category graph; in simpler words, the initialization of the category taxonomy considers all those nodes which have outdegree equal to 1 (i.e., which have only one super category in the noisy category graph) and adds these edges to the set $E(T_C)$. The algorithm is started on the category taxonomy with the (partial) page taxonomy as input (line 1).

In the algorithm we denote with T the taxonomy being refined and with T' the taxonomy that the algorithm draws on to update T. Initially $T = T_C$ and $T' = T_P$ (see line 1).

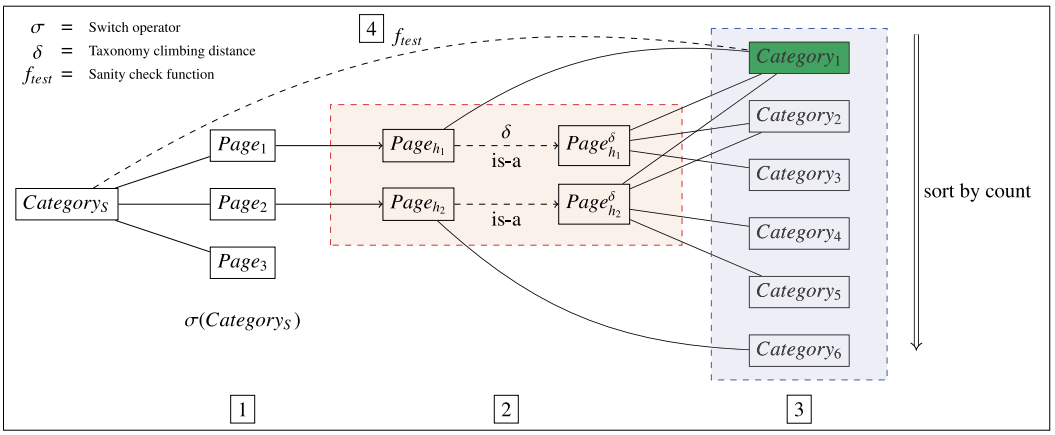


Fig. 8. Example of the application of the MultiWiBi iterative algorithm on the category side of Wikipedia. $Category_S$ and $Category_i$ denote the starting and the candidate categories, respectively.

6.3. The four steps

We now describe the core algorithm (Fig. 9) of our approach, which iteratively populates and refines the edge sets $E(T_C)$ and $E(T_P)$.

Item switch (step 1) In the first step we start by considering an uncovered node $t \in T$. Depending on the current iteration, t can be either a page or a category (line 5). We then apply an operator σ, that we call *switch operator*, which takes as input a Wikipedia item (either a page or a category) and returns the set of its Wikipedia counterpart elements, i.e., those items which belong to the other side of Wikipedia and are connected to it by means of a cross-link (see Section 2). In a few words, σ expresses the mutual membership relation existing between pages and categories. More formally, given $c \in C$, $\sigma(c)$ is the set of pages categorized with c, while given $p \in P$, $\sigma(p)$ is the set of categories associated with page p in Wikipedia. In this step, the algorithm starts from t and uses $\sigma(t)$ to switch from one taxonomy to the other (line 7 and Fig. 8, 1).

Example Consider the uncovered Wikipedia category $t =$ OLYMPICS (line 5). By applying the switch operator to OLYMPICS, we reach the following set of pages $\sigma(\text{OLYMPICS}) = \{$ PARALYMPIC GAMES, OLYMPIC GAMES, ..., OLYMPIC CUP$\}$ ($|\sigma(\text{OLYMPICS})| = 26$).

Taxonomy climbing (step 2) Given the dual Wikipedia items $\sigma(t) = \{t'_1, \ldots, t'_{|\sigma(t)|}\}$, the goal of this step is to harvest hypernyms of the dual nodes in $\sigma(t)$ which will then be switched back to the starting taxonomy. To do this, we build a set $H(\sigma(t))$ by "climbing" the taxonomy T', reaching all the hypernyms at distance less than or equal to the hypernymy distance parameter δ starting from each item $t'_i \in \sigma(t)$ (line 8). The maximum climbing distance changes during the iterations, so as to constrain the algorithm to favour closer hypernyms over the first iterations and allow it to reach further hypernyms as it proceeds (line 21 and Fig. 8, 2).

Example (cont'd) Out of the total 26 pages contained in $\sigma(\text{OLYMPICS})$, 23 pages come with a hypernym, discovered during the construction of the page taxonomy (line 8); for example, PARALYMPIC GAMES is a MULTI-SPORT EVENT. All the hypernyms at distance 1 are added to $H(\sigma(\text{OLYMPICS}))$, which is the set of the pages to project back to the category taxonomy; for example, MULTI-SPORT EVENT is contained in this set.

Candidate discovery (step 3) The goal of this step is to identify a set of candidate hypernyms for the starting node t. To this end, having $H(\sigma(t))$ as input to this step, we apply the *switch operator* to each t'_h in $H(\sigma(t))$ (lines 9–10) and we count the number of times we reach a node in T (line 11). As Wikipedia items in one taxonomy are usually associated with multiple items in the other taxonomy, items will be counted multiple times, so as to generate a distribution. The result of this step is thus a distribution over candidate nodes which notably belong to the same taxonomy given as input to the algorithm (cf. Fig. 8, 3). This is the core of the bitaxonomy algorithm, in which hypernymy knowledge is transferred from one taxonomy to the other.

Example (cont'd) For each hypernym page in $H(\sigma(t))$ we apply the *switch operator*, obtain the candidate categories and sum 1 for each of them. As a result we obtain the following distribution: {MULTI-SPORT EVENTS:4, ..., AWARDS:1, SWIMSUITS:1}, meaning that we end up counting the category MULTI-SPORT EVENTS four times (because four hypernymy paths in the page taxonomy led to MULTI-SPORT EVENT, which is in turn connected to this category) and other categories, such as AWARDS and SWIMSUITS, only once.

Algorithm 1 The bitaxonomy algorithm.

Input: T_P, T_C
1: $T := T_C$, $T' := T_P$, $\xi \leftarrow 1000$, $\lambda \leftarrow 1$, $\delta \leftarrow 1$, $\lambda_{max} \leftarrow 6$, $\delta_{max} \leftarrow 3$
2: **repeat**
3: $sizeT \leftarrow |E(T)|$
4: $convergence \leftarrow false$
5: **for all** $t \in V(T)$ s.t. $\nexists\ t_h \in T$, $(t, t_h) \in E(T)$ **do**
6: **reset** *candidate_count*
7: $\Sigma \leftarrow \sigma(t)$ ▷ step 1
8: $H \leftarrow get\_hypernyms(\Sigma, \delta, T')$ ▷ step 2
9: **for all** $t'_h \in H$ **do**
10: **for all** $t_h \in \sigma(t'_h)$ **do** ▷ step 3
11: $candidate\_count(t_h)$++
12: **end for**
13: **end for**
14: **for all** $t_h \in sort(candidate\_count)$ **do**
15: **if** $sanity\_check(t, t_h, T)$ **then** ▷ step 4
16: $E(T) := E(T) \cup \{(t, t_h)\}$
17: **break**
18: **end if**
19: **end for**
20: **end for**
21: **if** $T == T_C$ **and** $(|E(T)| - sizeT < \xi)$ **then** ▷ Parameter update and stop condition
22: $\lambda \leftarrow \lambda + 1$
23: **if** $\lambda >= \lambda_{max}$ **then**
24: $\lambda \leftarrow 1$
25: $\delta \leftarrow \delta + 1$
26: **end if**
27: **if** $\delta >= \delta_{max}$ **then** $convergence \leftarrow$ **true**
28: **end if**
29: **end if**
30: **swap** T and T'
31: **until** convergence
32: **return** $\{T, T'\}$

Fig. 9. The bitaxonomy algorithm.

Sanity check (step 4) The input for this step is the same in either of the two sides of the bitaxonomy, i.e., a starting node $t \in T$ and a candidate hypernym $t_h \in T$, belonging to the same taxonomy. The goal of this step is to select, whenever possible, the best hypernym amongst the candidate list found in the previous step. Such promotion is performed only if the candidate hypernym t_h passes a sanity check which guarantees the compatibility with the starting node t. Given the different nature of the two sides of Wikipedia, we devised specialised conditions; this step is thus the only one which depends on the current taxonomy being updated. As regards the page taxonomy, given the page $p \in T_P$ and the candidate hypernym page $p_h \in T_P$, the sanity check ascertains whether p_h is a sense for some of the hypernym lemmas extracted for p (see Section 4.1). As for the category taxonomy, given the category $c \in T_C$ and the candidate hypernym category $c_h \in T_C$, the sanity check ascertains whether c and c_h are connected by a path of length $\leq \lambda$ (see Section 6.4). If this holds, we then select the direct super-category of c lying on the shortest path between c and c_h. The rationale behind this asymmetry lies in the fact that only the category side of Wikipedia is backed with an underlying noisy graph, and connectivity techniques cannot also be generalised easily to the page side.

This fourth step considers the items contained in the distribution of step 3 in decreasing order, promotes the node t_h^* with the highest count which passes a sanity check, if any (line 15), and a new edge $e = (t, t_h^*)$ is finally added to the taxonomy (line 16). Note that as soon as a candidate node passes the sanity check a new edge is added to the taxonomy and all the remaining candidates are discarded (line 17). The sanity check has the aim of discriminating among the hypernym candidates contained in the set $H(\sigma(t))$, by checking whether it is safe to add an edge between the starting node and the candidate.

Example (cont'd) We proceed in decreasing order of vote and ascertain whether the sanity check for categories holds. As Multi-sport events has the highest count and is connected to the starting category Olympics by a path in the Wikipedia category network (in fact, the former is a direct super-category of the latter), we finally add the hypernym edge (Olympics, Multi-sport events) to T_C (line 16) and exit step 4 (line 17).

6.4. Parameter update and stop condition

At the end of each iteration the role played by the two taxonomies is swapped and the (partial) taxonomy becomes the new input for a new iteration (line 30). The four steps are repeated until a stop condition is satisfied (line 27). The

algorithm is governed by two parameters, the maximum path length parameter λ and the maximum hypernymy distance parameter δ. The former controls the maximum length of the path in the category sanity check; the latter regulates the maximum hypernymy distance in the taxonomy climbing step (step 2). We voluntarily let δ take smaller values than λ so as not to assign over-generalised hypernyms. At the end of a given iteration, whenever less than ξ edges have been added, λ is incremented. When a maximum value λ_{max} is reached, λ is reset to 1, in order that closer categories will henceforth be preferred, and δ is increased by one. As a safety stop condition, we also constrain the hypernymy distance parameter δ to a maximum value δ_{max}. Indeed, by starting from a page (or category) and climbing a taxonomy without such a maximum value, we would risk reaching the top of the taxonomy and assigning hypernyms which are too general (such as ENTITY or BEING) which would likely contribute to generating errors. In the case when the hypernymy distance parameter δ reaches the maximum value allowed, the algorithm is stopped and the two taxonomies returned. Note that the parameters are modified only when a temporary convergence with these parameters is reached: the fact that the algorithm assigns a small number of edges during any particular iteration, in fact, means that the path length parameter is not high enough to let the algorithm generalize sufficiently. Hence, the need to increase the path parameter and spin the algorithm through an additional iteration. Note also that, since λ depends on ξ, it is not possible to know *a priori* the number of iterations that the algorithm will have to perform.

6.5. Parameter tuning

The optimal values for the parameters used in the bitaxonomy algorithm (λ_{max}, δ_{max} and ξ) have been chosen according to two development sets containing 100 pages and 100 categories, disjoint from the two datasets used in sections 5.2 and 8.3. We let λ_{max} and δ_{max} range between 1 and 10, and ξ in $\{10, 100, 1000, 10000\}$.

7. Phase 3: bitaxonomy refinement

Despite the successful application of the bitaxonomy algorithm to the two taxonomies, they still suffer from structural shortcomings which we will now focus on.

As regards the page taxonomy, the algorithm crucially leverages two important features to discover the right hypernym to promote: first, a Wikipedia page needs to have categories associated with it, and, second, it also needs to provide at least one hypernym lemma. This means that there is a (small) class of Wikipedia pages to which the algorithm cannot be applied, and which are, in fact, left out. This class mainly contains redirections promoted to hypernym which, by construction, have neither categories nor definitions associated (cf. Section 2.1). For this reason we introduced a final refinement for the page taxonomy which addresses the problem of finding a proper generalization for this set of redirections.

As regards categories, the problem is very similar. Since the bitaxonomy algorithm crucially exploits the *switch operator* to harvest the pages associated with a certain category, it fails whenever the latter has no pages categorised under it. To this end we designed an ad-hoc procedure that overcomes this structural shortcoming.

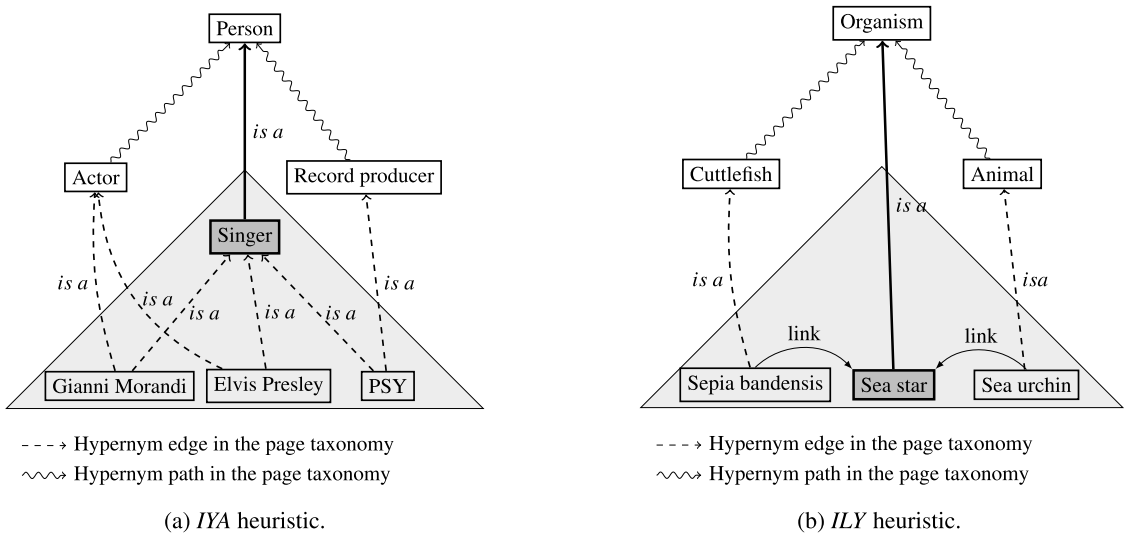
7.1. Page taxonomy refinement

At the core of the refinement of the page taxonomy there are two simple ideas which, applied in cascade order, both use a trivial taxonomical property: if a node in a taxonomy has two hypernyms, then these must reconcile somewhere up in the hierarchy, i.e., they must have a common ancestor. For example, in an ideal taxonomy the two hypernyms of ELVIS PRESLEY, namely SINGER and ACTOR, should both have PERSON or ARTIST as their lowest common ancestor. The two ideas, called IYA (I am if You Are) and ILY (I am Like You), both exploit this principle. The former, IYA (Fig. 10a), exploits the ancestors of the hyponyms of a given redirection. For example, in order to discover the hypernym for the redirection SINGER, we first consider, among others, the pages GIANNI MORANDI and PSY and consider in turn their alternative hypernyms, ACTOR and RECORD PRODUCER, respectively. We then climb the taxonomy until a common ancestor is encountered, i.e., PERSON, which is finally promoted as hypernym of the initial redirection SINGER. Ancestors which are met more frequently are preferred. The latter, ILY (Fig. 10b), contrarily to IYA, leverages the ancestors of those pages which have an outgoing link to the redirection considered, choosing the most voted ancestor. For example, in order to find the hypernym for the redirection SEA STAR, we consider all the pages pointing to the redirection, among which there are SEPIA BANDENSIS and SEA URCHIN. Similarly to the previous procedure, ILY determines the common ancestor ORGANISM and sets *isa*(SEA STAR, ORGANISM).

Note that the two procedures differ only in the set of starting Wikipedia pages considered. In the first case preference is given to pages which have the redirection as direct hypernym; in the second case, instead, the condition is relaxed and all the pages that contain an outgoing link to the redirection are considered. In order to evaluate the edges extracted thanks to this phase, we validated 100 randomly sampled relations, obtaining 93% accuracy.

7.2. Category taxonomy refinement

The refinement of the category taxonomy aims to address a structural weakness, represented by the fact that for a given Wikipedia category the cross-links are missing or limited in number. For this reason it is very difficult to provide hypernyms for this type of category on the basis of the cross-links, which are thus not sufficient to infer all the hypernymy



(a) *IYA* heuristic. (b) *ILY* heuristic.

Fig. 10. Patterns for the coverage refinement of the page taxonomy. Edges in bold represent inferred hypernymy relations.

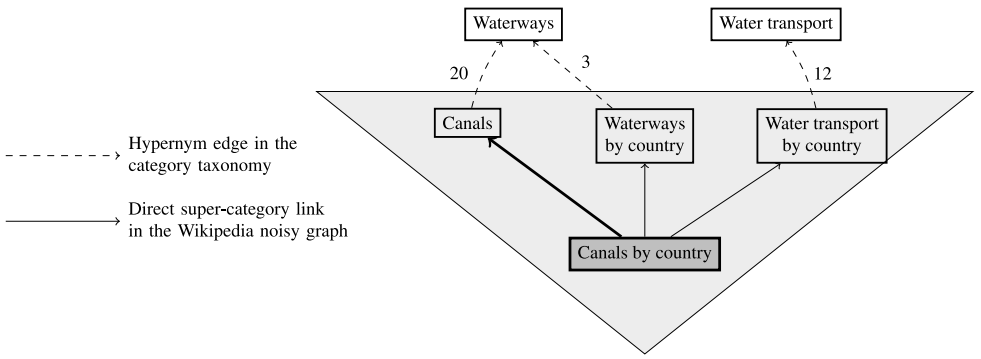
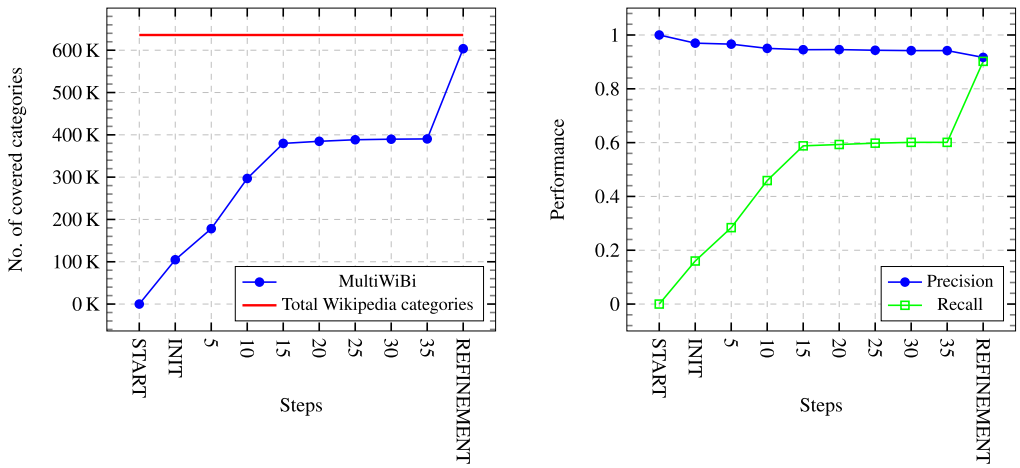


Fig. 11. Pattern for the category taxonomy refinement.

information required. For example, note that the English categories that are associated with 5 or less pages represent 40% of total number of categories in Wikipedia.

We thus designed a simple enrichment heuristic which, applied iteratively until convergence, adds hypernyms to those categories c for which no hypernym could be found in phase 2, i.e., $\nexists c'$ s.t. $(c, c') \in E(T_C)$ (see Fig. 11). Note that this heuristic does not leverage the cross-links but only the information learned during the application of the algorithm. Given an uncovered category c, we consider its direct Wikipedia super-categories and let each of them vote for their direct hypernym categories. Then we proceed in decreasing order of vote and select the highest-ranking category c' which is connected to c in T_C. We finally pick up the direct super-category c'' of c which lies in the path from c to c' and add the edge (c, c'') to $E(T_C)$. In the case of ties, categories which contributed most to score of c'' are favoured. For example, as shown in Fig. 11, given the category CANALS BY COUNTRY, we take all its super-categories (namely CANALS, WATERWAYS BY COUNTRY and WATER TRANSPORT BY COUNTRY) and let each of them vote according to their hypernym categories in T_C. For example WATERWAYS accumulates a score of 23 because during the bitaxonomy algorithm 20 pages contributed to the insertion of the edge (CANALS, WATERWAYS) and 3 pages contributed to the insertion of the edge (WATERWAYS BY COUNTRY, WATERWAYS). Given that WATERWAYS is the most voted hypernym, the algorithm chooses CANALS as hypernym because it is the category which contributes most to the score of WATERWAYS and therefore adds the edge (CANALS BY COUNTRY, CANALS) to T_C.

As previously done for the pages, we evaluated this procedure in the same manner, that is, by randomly sampling 100 edges and validating them by hand. This resulted in 81% accuracy, demonstrating that this approach consistently increases the category coverage, while keeping the quality of the extracted relations very high.



(a) Category coverage trend.

(b) Category precision and recall trend.

Fig. 12. Category taxonomy evaluation.

8. English bitaxonomy evaluation

8.1. Page taxonomy improvement

After the application of the first phase, in the Wikipedia page taxonomy 359,925 items out of 3,889,572 were still uncovered, i.e., had no hypernym(s) associated (cf. Section 4.2.1). After phases 2 and 3, however, 59,303 total edges were added to the page taxonomy, covering 58,113 nodes, about 15% of the total uncovered pages after the first phase.

8.2. Category taxonomy statistics

We applied phases 2 and 3 to the output of phase 1, which was evaluated in Section 5. In Fig. 12a we show the increase in category coverage at each iteration as well as after phase 3. The final outcome is a category taxonomy which includes 603,590 hypernymy links between categories, covering about 95% of the 635,972 categories in the 2012 English Wikipedia dump. The graph shows the steepest slope in the first iterations of phase 2, which converges around 400k categories at iteration 30, and a significant boost of another 213k hypernymy edges as the result of the refinement (phase 3).

8.3. Category taxonomy quality

To estimate the quality of the category taxonomy, we randomly sampled 1,000 categories and, for each of them, we manually associated the super-categories which were deemed to be appropriate hypernyms. We calculated precision and recall in the same way as we did for pages (see Section 5.1). Fig. 12b shows the performance trend as the algorithm iteratively covers more and more categories. Phase 2 is particularly robust across iterations, as it leads to increased recall while retaining very high precision. As regards phase 3, the refinement leads to only a slight precision decrease, while improving recall considerably. Overall, the final taxonomy T_C achieves 91.67% precision, 90.20% recall and 98.40% coverage on our dataset.

8.4. Quality of the upper taxonomies

To assess the quality of the top-level edges of our page and category taxonomy, we extracted and manually validated the top-most 100 edges. The evaluation resulted in 76% accuracy for the page side and 65% for the category side. This is an excellent result if we consider that these edges come from the upper level of the taxonomy which, by definition, contains very general concepts (often not well defined in Wikipedia). Note also that a manual validation of the upper taxonomy is also feasible, since it involves only a small number of edges and can thus easily be done in a relatively short time.

9. Related work

Although the extraction of taxonomies from machine-readable dictionaries was already studied in the early 1970s [37], pioneering work on large amounts of data only appeared in the early 1990s [28,40]. More recently, approaches based on hand-crafted patterns and pattern matching techniques have been developed to provide a supertype for the extracted terms [35,47–49,41,43, inter alia]. However, most of these methods do not link terms to existing taxonomies, whereas those that

explicitly link do so by adding new leaves to the existing taxonomy instead of acquiring wide-coverage taxonomies from scratch [50,51].

The recent upsurge of interest in collaborative knowledge curation has enabled several approaches to large-scale taxonomy acquisition [23]. Most approaches initially focused on the Wikipedia category network, an entangled set of generalization-containment relations between Wikipedia categories, to extract the hypernymy taxonomy as a subset of the network. The first approach of this kind was WikiTaxonomy [29,52], based on simple, yet effective lightweight heuristics, totalling more than 100k is-a relations. Another approach of this type was YAGO [30,53] which yields a taxonomical backbone by linking Wikipedia leaf categories to the first (i.e., most frequent) sense of their category heads in WordNet.

Interest in taxonomizing Wikipedia pages, instead, developed with DBpedia [54], which pioneered the current stream of work aimed at extracting semi-structured information from Wikipedia templates and infoboxes. In DBpedia, entities are mapped to a coarse-grained ontology which is collaboratively maintained and contains only about 270 classes corresponding to popular named entity types. Freebase [55] is a later development and a merger of other resources, such as MusicBrainz and ChefMoz. While also being based on infoboxes, it is a set of more than four million topics loosely organized and relies on the collaborative editing of what is now a mixture of both strict and informal relations. The Linked Hypernym Dataset (LHD) [33] is the most recent effort which tries to taxonomize the Wikipedia encyclopedia by associating Wikipedia pages with a DBpedia entity or a DBpedia ontology concept as their type. The types are hypernyms mined from pages' free text using hand-crafted lexico-syntactic patterns. Furthermore, LHD has been released in two versions: LHD 1.0 provides hypernyms which can be either DBpedia entities or concepts drawn from the DBpedia ontology (version 3.9), while LHD 2.0 contains concepts drawn from the smaller DBpedia ontology only. To our knowledge LHD is also the only other approach which, in addition to providing hypernyms for Wikipedia pages, also attaches the corresponding ambiguous hypernym lemmas. A few notable efforts to reconcile the two sides of Wikipedia, i.e., pages and categories, have been put forward only very recently: WikiNet [31,56] is a project which heuristically exploits different aspects of Wikipedia to obtain a concept network by deriving not only is-a relations, but also other types of relations. A second project, MENTA [32], creates one of the largest multilingual lexical knowledge bases by interconnecting more than 13M pages in 271 languages. Hypernym extraction, though, is supervised in that decisions are made on the basis of labelled training examples and requires a reconciliation step owing to the heterogeneous nature of the hypernyms. A totally different approach is proposed by [57], which exploits the hierarchical layouts of Wikipedia pages to extract hypernym/hyponym candidates. An SVM classifier is then trained in order to recognize which candidate is a real hyponym (e.g., the sub-heading *Earl Grey tea* is a hyponym of the page Black tea). The approach is quite complementary to ours, in that it focuses on the hierarchical structure of Wikipedia. However, it is based on hand-written and language-specific patterns to recognize hyponym candidates. We plan to investigate in the future how to generalize this approach to multiple languages and exploit it in our pipeline. Finally, our work differs substantially from the others in several respects: first, in marked contrast to most other resources, but similarly to WikiNet and WikiTaxonomy, our resource is self-contained and does not depend on other resources such as WordNet; second, similarly to MENTA and differently from all others, we address the taxonomization task on both sides, i.e., pages and categories, by providing an algorithm which mutually and iteratively transfers knowledge from one side of the bitaxonomy to the other; third, we provide a wide coverage bitaxonomy closer in structure and granularity to a manual WordNet-like taxonomy, in contrast, for example, to DBpedia's flat type-oriented hierarchy.

Another related task is sense alignment. The goal of this task is to link the same concept appearing in two distinct resources. Pilehvar et al. [58] propose a unified approach which uses a novel similarity measure to compare definitions across lexical resources. Nieman and Gurevych [59], instead, construct a multilingual lexical resource from Wiktionary by disambiguating semantic relations and translations. This new resource is then used to measure monolingual and cross-lingual verb similarity. The very same task was tackled by Gurevych et al. [60] who perform the alignment by means of a supervised classifier. Meyer et al. [61], instead, built a new multilingual resource based on Wiktionary and demonstrated that it could be used to cope with monolingual and multilingual sense alignment tasks for verbs. However, even if related, the task of sense alignment is rather different from the task of taxonomy building. In fact, the former is the task of establishing horizontal links between entries across resources which contain the same information (e.g. aligning WordNet's synset $athlete^{n}_{1}$ to the Wikipedia page Jock (athlete)) [59]; extracting a taxonomy from Wikipedia, instead, means establishing vertical links between a page (or category) and its most suitable generalization *within* the exact same resource. In contrast to sense alignment algorithms, the alignment between the two bitaxonomies returned by our algorithm is only a by-product result and not the aim of the research. In addition, approaches for cross-resource sense alignment generally rely on rich context-based features (such as textual definitions associated with the resource entries), something which is not easily obtainable in this case (e.g. categories do not have definitions).

The next section presents evidence for the above assertions by comparing statistics about the structure of the taxonomies and by presenting experiments that assess their quality across all other resources.

10. Comparative evaluation

In this section, using the same measures as those described in Section 5, we provide a thorough comparison of MultiWiBi's English bitaxonomy against all the major alternatives in the literature. Section 10.1 presents and discusses several dimensions characterizing the different resources. In Section 10.2 we report on different measures concerning the taxonomical structure. In Section 10.3 we describe the selection and construction of the datasets used, while in Section 10.4 we

Table 1
Features of the main taxonomic resources. *WN* stands for WordNet, *P* for Wikipedia pages, *R* for Wikipedia redirections, *C* for Wikipedia categories, *D* for DBpedia ontology, *H* for Human effort and *WKT* for Wiktionary.

Resource	Timestamp (dd/mm/yy)	Pages	Categories	Page hyp. sense inventory	Category hyp. sense inventory	Type of additional sources	# Languages
MultiWiBi	01/10/12	✓	✓	P + R	C	H, Syntax	271
WikiNet	04/01/12	✓	✓	P + C	P + C	H, Syntax	1 (EN)
DBpedia	01/06/12	✓	✗	D	–	D	125
LHD	10/12	✓	✗	D	–	H, PoS tagger, D	3 (EN, DE, NL)
WikiTaxonomy	01/10/12	✗	✓	–	C	H, Syntax	1 (EN)
YAGO	01/12/12	✓	✓	C + WN	WN	H, Syntax, WN	1 (EN)
YAGO3	04/14	✓	✗	C + WN	–	H, Syntax, WN	10
MENTA	10/04/10	✓	✓	P + C + WN	P + C + WN	H, Syntax, WN, WKT	271

report and discuss the results obtained. Note that YAGO3 doesn't provide taxonomic information for Wikipedia categories, so, for its evaluation we used YAGO.

10.1. Features of taxonomic resources

In order to examine the differences between MultiWiBi and all other resources analyzed in this work, we first present several features in Table 1. For each resource we report i) the data timestamp, i.e., the most accurate date of the Wikipedia dumps from which data has been derived, ii) whether the resource provides hypernyms for Wikipedia pages, iii) whether the resource provides hypernyms for Wikipedia categories, iv) the inventory from which the hypernyms are drawn, v) the type of dependencies upon language, tools and other resources, and finally vi) the degree of multilingualism, measured as the number of languages covered by the resource at the time of writing.

Timestamp First, as can be seen from Table 1 (second column), all resources but MENTA and YAGO3 are isochronous: apart from small differences in the reference month, they all come from 2012. This makes comparison much easier.

Wikipedia sides covered and the hypernym inventories For years, resources have been covering only one of the two sides and approaches could mainly be divided into two groups: those which provide hypernyms only for pages and those which provide hypernyms only for categories. Within this classification we can further discriminate on the basis of the inventories from which hypernyms are drawn: i) DBpedia and LHD are consistent and return hypernyms drawn from the DBpedia upper ontology, while ii) WikiTaxonomy is also consistent by returning Wikipedia categories. In contrast with the three above resources, the other systems presented in Table 1 try to cover both the two sides of Wikipedia; WikiNet, MENTA, YAGO and MultiWiBi are, in fact, the only resources which provide hypernyms for both pages and categories. As regards the hypernym inventory, though, they differ substantially. On the one hand WikiNet mixes the two sides of Wikipedia together, and MENTA returns hypernyms in which Wikipedia pages, Wikipedia categories and WordNet synsets are all amassed together. On the other hand, instead, YAGO outputs WordNet synsets as hypernyms for Wikipedia categories, while YAGO3 outputs both WordNet synsets and Wikipedia categories for Wikipedia pages. MultiWiBi, in distinct contrast, by returning two separate (but aligned) taxonomies with two disjoint hypernym sets, associates hypernyms in a coherent and consistent manner; as a result, Wikipedia pages have pages and categories have categories as hypernyms.

Dependency on additional sources Another feature, which separates the systems into different classes, is the need for any sort of human intervention, additional resource or sense-tagged corpora. The second to last column of Table 1 shows for each resource the type of such dependency. The degree to which each resource is tied to human effort is in turn linked to the ease of converting that resource into another language and, of course, the less human effort required, the better. As regards human intervention, MultiWiBi depends only on the list of stopwords introduced in the syntactic step and does not need any additional human effort. LHD, WikiNet and WikiTaxonomy, instead, rely heavily on lexico-syntactic patterns (e.g., *X by Y* or *X [VBN IN] Y*): LHD learns the patterns using 600 manually annotated training examples for each language; in WikiNet and WikiTaxonomy, instead, patterns are defined by hand, making such pattern-based models at least laborious to generalize across languages. YAGO and YAGO3 involve human effort because the category-to-WordNet mappings were corrected by hand, also making it difficult to generalize to many languages automatically. Finally, in the case of MENTA: i) Wikipedia-to-WordNet mappings are established by a supervised linker, trained on 200 manually labelled training examples, ii) the Category-WordNet subclass relationship is learnt thanks to a supervised learning model, trained on 1,539 labelled training mappings. This, however, is done only once for all the languages.

As regards the amount of dependency on external tools, we note that MultiWiBi needs a syntactic parser only for extracting hypernym lemmas in English and for extracting heads from the strings passed to the SSR module. LHD requires only a PoS-tagger in order to train the transducer which learns lexical-syntactic patterns. WikiNet, WikiTaxonomy, YAGO, YAGO3 and MENTA all need a syntactic parser to extract heads from categories. Needless to say, the dependency on tools which are language-specific limits the applicability of a system to only those languages having such tools. Even though MultiWiBi relies on syntax in the English case, in Section 11 we introduce a new mechanism for extracting hypernym lemmas in other languages, which requires nothing but the raw Wikipedia dump in the desired language.

Table 2
Structural analysis of the taxonomies in the literature.

Feature	Resources							
	MultiWiBi	WikiNet	DBpedia	MENTA	LHD 1.0	LHD 2.0	YAGO3	WordNet
# nodes	3,600,781	2,949,685	1,906,274	2,936,667	3,038,604	2,960,780	3,628,053	82,115
coverage	92.45%	75.51%	49.18%	66.16% (80.54%)	67.14%	65.94%	72.86%	–
# edges	4,100,634	14,280,200	2,112,468	2,958,235	3,013,824	2,960,508	20,469,826	84,427
avg. height	5.89	1.71	4.11 (4.23)	1.76 (4.90)	1.02 (2.25)	1 (2.95)	1.01(5.91)	8.07

(a) Page taxonomies.

Feature	Resources					
	MultiWiBi	WikiNet	WikiTax	YAGO	MENTA	WordNet
# nodes	605,887	487,469	389,027	385,657	559,530	82,115
coverage	94.91%	63.39%	55.97%	52.10%	56.18%	–
# edges	603,557	834,837	568,987	378,942	555,155	84,427
avg. height	16.69	3.11	3.46	1 (6.75)	1.33 (3.73)	8.07

(b) Category taxonomies.

As regards, instead, the dependency on external resources, we can distinguish between two types of dependency: some approaches use an external resource only as hypernym inventory. This is the case for DBpedia, LHD, YAGO and YAGO3: DBpedia and LHD use the DBpedia Ontology as hypernym inventory (letter D in Table 1), YAGO links Wikipedia categories to WordNet synsets and YAGO3 draws out hypernyms for Wikipedia pages from the sets of Wikipedia categories and WordNet synsets. These systems, however, do not exploit the external resource any further. In contrast, MENTA makes heavy use of external resources for different purposes: not only is countability information about category heads based both on both WordNet and Wiktionary, but its is-a classifier also takes as input the hypernymy information already contained in WordNet. In marked contrast, WikiNet, WikiTaxonomy and MultiWiBi do not rely on any additional information: all these are self-contained taxonomies which exploit data coming solely from Wikipedia itself.

Degree of multilingualism Finally, another dimension we considered, tightly intertwined with that of dependency on external resources, is multilingualism: first of all, note that all the resources are (or could be made) multilingual, thanks to the interlanguage links which connect the different editions of Wikipedia (see Section 11 for details). By merely analysing the resources as they have been publicly released, instead, we note that three resources rely only on the English Wikipedia, namely WikiNet, WikiTaxonomy and YAGO. The possibility of an extension to other languages for these three resources is at the very least questionable, owing to their dependency on language tools. LHD has been applied to 3 languages, and separate data repositories are released independently; The YAGO3 page taxonomy has been released in 10 different languages; a category taxonomy, instead, is not available. DBpedia has been released in 125 languages, managed by 18 isolated chapters; MultiWiBi and MENTA are the only two resources which are applicable to every Wikipedia language, making them the only truly language-indepedent approaches (even though the latter also relies on WordNet, Wiktionary and labelled training examples).

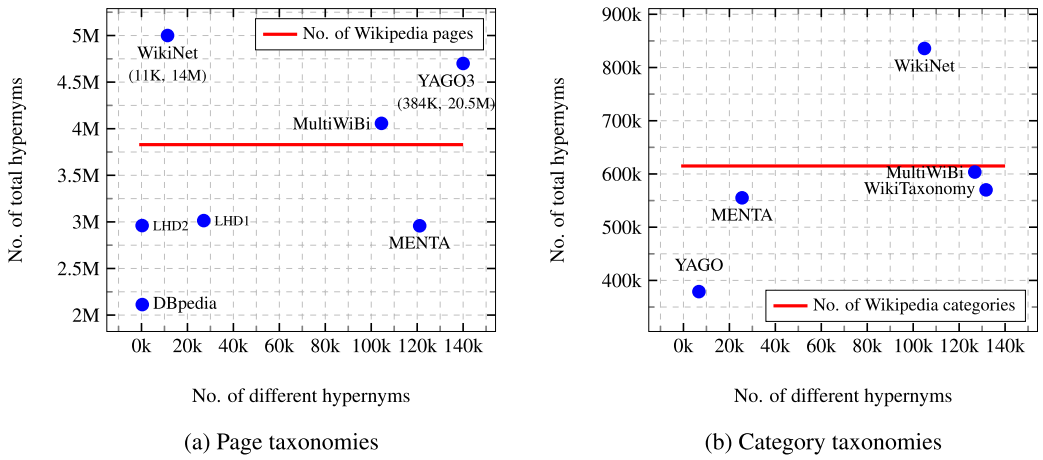
10.2. Structural analysis of the taxonomic resources

In addition to all the aspects considered in the previous section, we present some comparative evaluation concerning the structure of the taxonomies. As already mentioned in Section 5.1, it is not easy to find valid measures capable of evaluating a taxonomy in a comprehensive manner. Before reporting on precision, recall and coverage we wish to present and discuss several structural measures for all the considered resources, both on the page and the category side of Wikipedia (Tables 2a and 2b, respectively). We take into account several indicators, among which are: the number of nodes and edges contained in the taxonomies, the coverage of Wikipedia pages and categories, the average height, and a new measure called *granularity*.

10.2.1. Structural features of taxonomies

Page-based resources Table 2a reports the statistics concerning Wikipedia pages. In order to have a reference point, we report the same measures also for WordNet (note, however, that WordNet contains far fewer nominal concepts, so the number of nodes and edges is not informative). The first measure that we consider important is the number of nodes (first row), along with coverage, calculated as the number of pages with respect to the whole Wikipedia for which a hypernym is returned, regardless of its correctness (second row). Since MENTA builds upon a Wikipedia dump dating back to 2010, in parentheses we report the coverage with respect to the page inventory at that time. As can be seen, MultiWiBi is the best resource in terms of coverage, with 92.45% of nodes covered by a hypernym. Since WordNet does not contain Wikipedia pages, coverage cannot be calculated and is thus not shown in the table.

The third dimension considered is the number of edges (third row), which expresses the quantity of hypernymy information contained in a resource. As can be seen, MultiWiBi provides the largest number of edges, surpassed only by WikiNet. The high number of edges contained in WikiNet, however, does not correlate with quality, and we will study this further in Sections 10.2.2 and 10.4.



(a) Page taxonomies (b) Category taxonomies

Fig. 13. Hypernym granularity for the resources.

The most important feature is probably the average height of a taxonomy (last row), measured as the average length of hypernymy paths linking leaves to any root. This feature gives an idea of the generalization power of each taxonomy, since the longer the path, the more fine-grained the generalization. For DBpedia, LHD 1.0, LHD 2.0, YAGO, YAGO3 and MENTA, which rely on external taxonomies, in parentheses we also report the average height adjusted by including the external taxonomy. The first three use the DBpedia Ontology for returning the hypernyms for the Wikipedia pages; the last two resources, instead, also use WordNet and Wikipedia categories as their sense inventory. The average height of these augmented resources is obviously greater; nonetheless, with an average height of 5.89, MultiWiBi surpasses all other approaches, making it the resource structurally closest to WordNet, which has height 8.07.

Category-based resources In general the scenario for categories is not very different from the one illustrated for the page-based resources. As regards categories (see Table 2b), MultiWiBi is again the resource with the best coverage. YAGO is the resource with the lowest coverage, due to the fact that attention has been paid only to leaf categories. MultiWiBi exhibits the maximum average height, more than five times greater than any other resource. We also note that the average height of the category taxonomy T_C is much greater than that of the page taxonomy T_P, due to the fact that the category taxonomy distinguishes between very subtle classes (such as Albums by artists vs. Albums by recording location, etc.), which, instead, all get merged into the same concept Album in the page taxonomy.

10.2.2. Taxonomy granularity

A second important aspect that we analyzed was the granularity of each taxonomy, determined by drawing each resource on a bi-dimensional plane with the number of distinct hypernyms (i.e., non-leaf nodes) on the x axis and the total number of hypernyms (i.e., edges) on the y axis. Figs. 13a and 13b show the position for the page-based and the category-based resources, respectively. Two baselines that use two opposite strategies (not displayed in the figure) are essential for determining the differences among the different systems' positions in the two-dimensional plane: the first represents the baseline system which always assigns the same hypernym to all the Wikipedia items (achieving minimum granularity), while the second represents the system which assigns a different fictitious hypernym to each Wikipedia item (achieving maximum granularity). As can be seen, MultiWiBi, as well as the page taxonomy of MENTA, is the resource with the best granularity, as not only does it attain high coverage, but it also provides a larger variety of classes as generalizations of pages and categories. Specifically, MultiWiBi provides hypernyms for over 4M hypernym pages, chosen from a range of 104k distinct hypernyms, while others exhibit a considerably smaller range of distinct hypernyms (e.g., DBpedia by design, but also WikiNet, with around 11k distinct page hypernyms). The large variety of classes provided by MENTA, however, is due to it providing more than 100k Wikipedia categories as page hypernyms (among which categories about *deaths* and *births* alone represent more than 3% of the distinct hypernyms). Finally, YAGO3 exhibits the highest number of distinct hypernyms (about 380k). As regards categories, while the number of distinct hypernyms of MultiWiBi and WikiTaxonomy is approximately the same (around 130k), the total number of hypernyms returned by WikiTaxonomy (around 580k for both taxonomies) refers to half of the categories covered by MultiWiBi (see row 'coverage' in Table 2a). As regards WikiNet, its large number and variety of category hypernyms is, instead, counterbalanced by low precision and recall, as we show in the experimental results (Section 10.4).

10.3. Experimental setup

We compared MultiWiBi against the Wikipedia taxonomies of the major knowledge resources in the literature providing hypernym links, namely DBpedia, WikiNet, MENTA, WikiTaxonomy, YAGO and YAGO3 (see Section 9). As datasets, we used

Table 3
Taxonomy comparison at lemma level.

	P	R*	C	# items
Lemma	94.83	90.20	98.50	1,000
Lemma LHD	92.78	81.00†	87.30	

† denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and LHD.

Table 4
Page and category taxonomy evaluation.

Dataset	System	P	R*	C	# items
Pages	MultiWiBi	90.76	87.48	94.78	767
	WikiNet	56.86†	71.32†	82.01	
	DBpedia	77.87†	54.11†	58.27	
	MENTA	81.52†	72.49†	88.92	
	YAGO3	86.44	83.57	83.57	
	LHD 1.0	76.20†	53.85†	70.66	
	LHD 2.0	91.57	63.75†	69.62	
Categories	MultiWiBi	90.65	89.06	98.26	631
	WikiNet	64.05†	49.92†	71.16	
	WikiTax	89.68	55.15†	59.43	
	YAGO	93.58	53.09†	56.74	
	MENTA	87.11	84.63	97.15	
	MENTA$^{-\text{ENT}}$	85.18	71.95†	84.47	

† denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and the daggered resource.

our gold standards of 1,000 randomly-sampled pages (see Section 5) and categories (see Section 8.3). In order to ensure a fair playground evaluation we decided to reannotate each item in both datasets, considering for inclusion all competitors' hypernyms for that item. Moreover, given the heterogeneity in the release date of the resources, we detected those pages (categories) which do not exist in any of the above resources and removed them to ensure (potential) full coverage of the dataset across all resources. As already explained in Section 10.1, in fact, MENTA is the only resource based on a dump dating back to 2010, a bit far from the others. However, if on the one hand we acknowledge its performance might be relatively higher on a 2012 dump (though potentially counterbalanced by higher ambiguity), on the other hand, the software for generating MENTA over a different Wikipedia dump is not available.[10] WikiTaxonomy, originally based on a 2009 dump, was, instead, re-implemented in order to align it to the same dump used by MultiWiBi. The last column of Table 4 reports the size of the levelled datasets after the item deletion.[11]

10.4. Results

Lemma taxonomy Thanks to our procedure (see Section 4.1), which exploits dependencies extracted by the Stanford parser, the lemma taxonomy achieves very good scores in all three measures. As can be seen in Table 3, in fact, MultiWiBi scores more than 90% in precision, recall and coverage, overtaking LHD, which loses around 10 points in recall and coverage despite having a good precision. Not only does this confirm our assumption that a Wikipedia taxonomy can be extracted by page definitions, but also shows that most of the pages in the encyclopedia have a well formed definition.

Wikipedia pages We first report the results of the knowledge resources which provide page hypernyms, i.e., we compare against WikiNet, DBpedia, MENTA, YAGO3 and LHD. We show the results on our page hypernym dataset in Table 4 (top). As can be seen, all systems but WikiNet exhibit very good precision. WikiNet on one side and LHD 2.0 on the other side stick to the two opposite poles of the precision-recall trend: the former achieves high recall (around 71%) at the cost of a much lower precision (around 57%) due to the high number of hypernyms provided, many of which are incorrect, whereas the latter is characterised by high precision, but low recall. LHD 2.0, instead, is the system with the highest precision, it shows only modest coverage and recall and an inspection of the answers returned revealed that 32% and 11% of the hypernyms were http://dbpedia.org/ontology/Agent and http://dbpedia.org/ontology/Place, respectively, which, despite being correct, are very general. In the case of DBpedia, even considering both types of edge provided for the hypernym relation, a modest precision (77.87) (similar to that of LHD 1.0) and low coverage (55.93) are shown, due to the dependency on the availability of infoboxes in Wikipedia pages. Here we do not report individual performances for the several types of edges contained in DBpedia because the coverage did not change significantly when changing the source of the information

[10] Personal communication with the authors.
[11] We wish to make it very clear that these datasets are the same as those presented in Section 5.2 and Section 8.3, but are different only as regards size, because of the item deletion.

Table 5
Excerpt of the answers given by the different systems on the category dataset.

Category	Resources				
	MultiWiBi	WikiNet	WikiTax	YAGO	MENTA
Nigerian culture	Culture by nationality	Cultures of Africa	African culture by nationality	–	entity_n^1
Racism in Russia	Racism by country	Black-on-black racism	–	–	entity_n^1
Orellana Province	Provinces of Ecuador	Province capitals of Ecuador	Provinces of Ecuador	–	entity_n^1
Salvadoran cuisine	Latin American cuisine	Latin American culture	Cuisine by nationality	politician_n^1	entity_n^1
Turkish artists	Artists by nationality	–	Artists by nationality	artist_n^1	person_n^1
People from Campania	People by region in Italy	People by region in Italy	People by region in Italy	person_n^1	person_n^1

(ranging between 10.30% and 55.93%). MENTA and YAGO3 are the resources closest to ours; however, we remark that the hypernyms output by both resources are very heterogeneous. MENTA, in fact, has 48% of the hypernyms represented by a WordNet synset, 37% by Wikipedia categories and only 15% by Wikipedia pages. YAGO3 exhibits similar behaviour, giving as output hypernym categories, WordNet synsets or YAGO3 and OWL classes (such as *yagoGeoEntity* or *owl:Thing*). In contrast to all other resources, MultiWiBi outputs hypernyms in a coherent manner, by linking pages to hypernym pages, while at the same time achieving the highest performance of 90.76% precision, 87.48% recall and 94.78% coverage.

Wikipedia categories We then compared MultiWiBi with all the knowledge resources which deal with categories, i.e., WikiNet, WikiTaxonomy, YAGO and MENTA.

We show the results on our category dataset in Table 4 (bottom). MultiWiBi is the best resource, achieving the second highest precision (90.65%) and the highest recall (89.06%) and coverage (98.26%). WikiNet is characterised by the lowest precision and recall. The lowest coverage, between 56% and 59%, is attained by WikiTaxonomy and YAGO: in the former case this is likely due to the inadequacy of lexical-syntactic patterns which do not succeed in capturing all category variants, whereas in the latter case this is due to the fact that only leaf categories are considered. MENTA is, again, the closest resource to ours, obtaining comparable overall performance. Notably, however, MENTA outputs the first WordNet sense of *entity* for 13% of all the given answers which, despite being correct and counted in precision and recall, is uninformative. Since a baseline system which always outputs *entity* would maximise all the three measures, we also calculated the performance for MENTA when discarding *entity* as an answer; as Table 4 shows (bottom, $\text{MENTA}^{-\text{ENT}}$), recall drops to 71.95%. Note that ENTITY is never returned by MultiWiBi as hypernym of a dataset item. To investigate this phenomenon in more detail, we also identified the most general Wikipedia pages (ENTITY and BEING) and categories (OBJECTS, CONCEPTS, HUMANS, CULTURE) and calculated the number of times they were given as output as hypernyms in the two taxonomies. In marked contrast with MENTA, which provides a general hypernym for 130,408 nodes in its taxonomy, MultiWiBi outputs general hypernyms only for 1,808 (0.05% of the total) pages and just 38 (0.006% of the total) categories.

Table 5 shows, by way of example, the different answers given by the systems for some items in the category dataset. As can be seen, MENTA's answers are quite general and much less specific than those returned by other systems. Further analysis, presented below, shows that the specificity of the hypernyms returned by the other systems is considerably lower than that of MultiWiBi.

10.5. Taxonomy specificity

To get further insight into our results we also performed an additional analysis by means of a last quality measure. We estimated the level of specialization of the hypernyms in the different resources on our two datasets. The idea was that a hypernym should be valid while at the same time being as specific as possible (e.g., SINGER should be preferred over PERSON, if they both apply). We therefore calculated a measure, which we called specificity, that computes the percentage of times a system outputs a more specific answer than another system. To do this, we manually ranked the answers of all the systems for the 767 items in the page dataset (described in Section 10.3). Each hypernym returned was evaluated according to the degree of specificity, by comparatively associating each valid answer with a score $0 < score_S(x) \leq 1$. In the case when the answer was missing or wrong, $score_S(x)$ was set to 0. For example, given the page ALAN EDMONDS, we assigned the score .5 to MENTA's hypernym 1930S BIRTHS and 1 to MultiWiBi's hypernym REPORTER because the former is less specific than the latter. However, certain systems often return categories as hypernyms, which are thus more likely to be more specific than MultiWiBi. This was the case with YAGO3, for example, which assigned several categories as hypernym (as well as WordNet synsets) to each page.

Since a system S is allowed to return more than one hypernym per item, for each system we considered only the most specific answer. When comparing two systems S_1 and S_2 on an item x, we say that S_1 is more specific than S_2 whenever $score_{S_1}(x) > score_{S_2}(x)$. We then calculate three types of configuration, depending on $score_{S_1}(x)$ being equal to, greater than or less than $score_{S_2}(x)$ and denote these configurations with $S_1 = S_2$, $S_1 > S_2$ and $S_1 < S_2$, respectively. More formally:

$$S_1 \odot S_2 := |\{a \in D : score_{S_1}(x) \odot score_{S_2}(x)\}| / |D|$$

where $\odot \in \{=, >, <\}$ and a is an item (i.e., a page or a category) of the dataset D. Table 6 shows the results for all the resources and for both the page and category taxonomies: MultiWiBi consistently provides considerably more specific

Table 6
Specificity comparison. $MultiWiBi \odot S := |\{a \in D : score_{MultiWiBi}(x) \odot score_S(x)\}| / |D|$ and $\odot \in \{=, >, <\}$.

Dataset	System (X)	MultiWiBi = X	MultiWiBi > X	MultiWiBi < X
Pages	WikiNet	21.51	75.36	3.13
	DBpedia	29.99	59.19	10.82
	MENTA	20.47	54.24	25.29
	LHD 1.0	46.68	46.41	6.91
	LHD 2.0	25.68	64.41	9.91
	YAGO3	0.00	5.00	95.00
Categories	WikiNet	42.00	45.32	12.68
	WikiTax	42.31	40.25	17.43
	YAGO	9.51	86.05	4.44
	MENTA	10.62	78.61	10.78

hypernyms than any other resource (middle column), quantitatively corroborating our qualitative insight based on example inspection.

To further investigate the results of systems with heterogeneous hypernyms we considered only the items in the dataset for which a given system output a Wikipedia category and recalculated the specificity for all such systems (i.e., MENTA, YAGO3, WikiNet and LHD 1.0).[12] Results show that only MENTA has better performance, being more specific than MultiWiBi 44% of times. WikiNet improves its scores but still remains less specific than MultiWiBi 52% of times. The same happens for LHD 1.0, which is more specific than MultiWiBi just 14% of the times. YAGO3, instead, remains more specific than MultiWiBi 90% of the time. This is because, for each item in our dataset, the most specific answer from YAGO3 is most of the time a category. Overall this experiment demonstrates that MultiWiBi is still more specific than the other systems even if categories are finer-grained than pages.

11. Projecting the bitaxonomy

We now describe a method for obtaining a bitaxonomy in any language other than English. The general idea is that of combining the bitaxonomy obtained in English and the exact same methodology outlined in Section 3, in order to obtain a bitaxonomy in a second, arbitrary language. To do this we will crucially leverage a very important element characterizing Wikipedia, namely the interlanguage links. By linking concepts in a Wikipedia language to their equivalent concepts in another language (when present), the interlanguage links play a key role in that they allow MultiWiBi, as well all other approaches, to make the taxonomic information available across languages. We inform the reader in advance, though, that MultiWiBi goes beyond the direct integration of such interlanguage links, by means of an innovative approach that is able to cover even Wikipedia items which do not have an English counterpart.

Interlanguage links and the projection rule "Interlanguage links are links from a page in one Wikipedia language to an equivalent page in another language. [...] For example, the Irish Wikipedia has a page on Ireland titled "Éire", so the English Wikipedia page on Ireland will link to the Irish one, and vice versa".[13] Thanks to the interlanguage links it is possible to align pages contained in the English Wikipedia to pages in another language, while preserving the original meaning in the target language. Notably, interlanguage links are also present between the categories of Wikipedias in different languages. This kind of link paves the way for a simple, yet effective mechanism which enables us to project the hypernymy information coming from one language onto another language. We call this mechanism the *projection rule*: we will exploit this rule to project the bitaxonomies across languages. Simply put, by means of the interlanguage links, this rule checks whether a given Wikipedia item in a source language (a page or category) and its hypernym also exist in the target language. More formally, the *projection rule* is defined as follows:

$$X_E \textit{ is-a } Y_E \ \wedge \ X_E \,||\, X_F \ \wedge \ Y_E \,||\, Y_F \ \Rightarrow \ X_F \textit{ is-a } Y_F \qquad \forall X_E, Y_E \in T_E, X_F, Y_F \in T_F \qquad \text{(Projection rule)}$$

According to this rule, given the English language E and another arbitrary language $F \neq E$, if we know that i) an English page X_E in the taxonomy T_E has a hypernym (X_E *is-a* Y_E), that ii) the English page has an equivalent page X_F in language F ($X_E \,||\, X_F$) and that iii) the English hypernym has an equivalent in language F ($Y_E \,||\, Y_F$), then we can safely infer that the latter is a valid hypernym for the foreign page (X_F *is-a* Y_F). This idea is also depicted in Fig. 14 with an example. We can see the English page Subapical consonant has Consonant as hypernym and, furthermore, we know that the corresponding Italian equivalents are Consonante subapicale and Consonante, respectively; given these facts, then, we can derive the fact that the same is-a relation holds between the two corresponding Italian Wikipedia pages (i.e., Consonante subapicale is-a Consonante).[14] We will draw upon the projection rule basically on two occasions: i) for projecting the English bitaxonomy

[12] LHD 2.0 and DBpedia do not give categories as hypernym for pages, so we could not recalculate the results against them.

[13] http://en.wikipedia.org/wiki/Help:Interlanguage_links.

[14] The projection rule, though, despite being correct in principle, might not always hold in practice. It might happen, in fact, that interlanguage links do not preserve meaning across two languages because they do not align exactly the same concept (e.g., very specific types of snow might not be represented

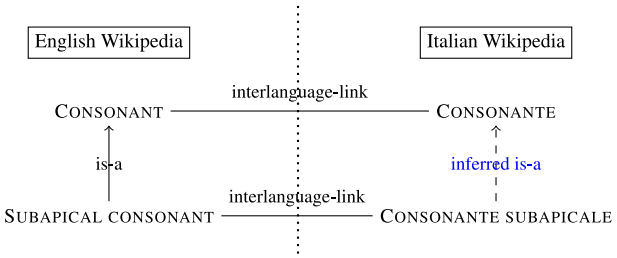


Fig. 14. Example of the application of the projection rule. The dashed edge in the Italian Wikipedia (right) represents new is-a information that can be inferred from English to its counterpart (left).

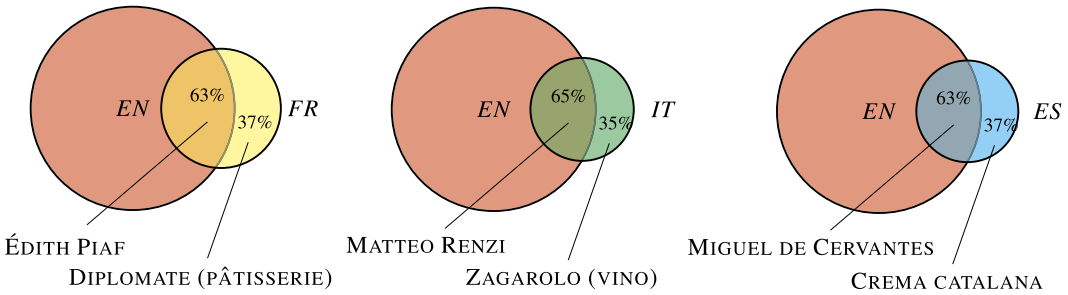


Fig. 15. Relationship between the English Wikipedia and the versions in other languages.

(see Section 11.2) and ii) for building a multilingual gold standard which will enable a fair comparison across languages (see Section 12).

A limitation of the interlanguage links As can be seen in Fig. 15, though, the English Wikipedia overlaps Wikipedias in other languages to a certain extent only. For example, only 65% of the Italian pages have an equivalent in English. With regard to English, Wikipedias in other languages contain additional concepts which, either are typical of that particular culture and often exist only in that language (such as the Italian page CASTAGNOLE (DOLCE), a typical Italian sweet), or, despite not being culture-specific, represent some other culture's concept (e.g., the French page TEATRO DEL GIGLIO about a famous Italian theatre, which could also exist in English, but has not been written yet). From here on we will call this set of pages *WEE pages* (pages Without English Equivalent). Therefore, any procedure which relies only on interlanguage links for producing a multilingual taxonomy will have the major drawback that its application will be limited only to those pages which have an equivalent in English.

Going beyond the interlanguage links We now present an innovative approach which will overcome the above limitation and will also provide hypernyms for those pages which do not have a corresponding page in the English Wikipedia. Our method is general and can be applied to any version and any language of Wikipedia, having as system input only the respective XML Wikipedia dump: none of the algorithmic procedures presented from here on is bound to any language whatsoever. However, for our convenience we present, discuss and evaluate the bitaxonomies in three languages: French (FR), Italian (IT) and Spanish (ES) (see details in Section 12).

In order to obtain a full bitaxonomy in a language other than English, we put forward a mechanism which compensates for the lack of a syntactic parser in another language (used in the syntactic step, cf. Section 4.1) and proceed in three steps:

1. **Construction of a Translation Table (TT)**: we provide a mechanism to build a translation table for a large number of lemmas contained in Wikipedia;
2. **Extraction of multilingual hypernym lemmas**: we exploit the translation table built in the previous step to associate each Wikipedia page of another language with one hypernym lemma;
3. **Application of WiBi**: we apply exactly the same procedure (hypernym lemma disambiguation, bitaxonomy algorithm and bitaxonomy refinement) presented for the English case (cf. Sections 4–7).

in all the Wikipedias or they might be translated as general snow, without fine-grained distinction). Thus, in order to evaluate, in general across the whole Wikipedia, whether quality was preserved across languages by means of the projection rule, we randomly sampled 500 pages of the English Wikipedia which presented an interlanguage link in Italian, and then evaluated the correctness of these links. We found that in only 2 cases (i.e., 0.004% of the total) did the equivalence between the aligned concepts not hold. Note that wrong cases also include alignments which are not completely incorrect: for example the English page MYTHOLOGICAL HYBRID is linked to the Italian page CECAELIA which is a particular mythological hybrid.

Table 7
Excerpt of the English–Italian translation table (numbers indicate translation confidence).

English lemma	Translations
plane	piano_cartesiano:0.20 piano:0.15 pialla:0.04 aeroplano:0.03 aereo:0.023 piano_astrale:0.02 …
car	automobile:0.33 autovettura:0.11 automobili:0.05 auto:0.02 autovetture:0.01 vettura:0.01 …
key	chiave:0.37 chiavi:0.03 chiave_crittografica:0.001 chiave_segreta:0.0005 …

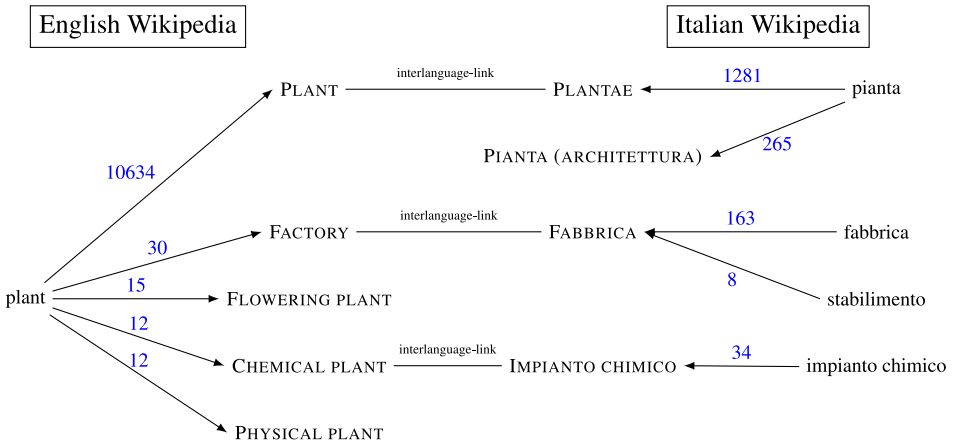


Fig. 16. Paths connecting the surface anchor *pianta* in Italian to the surface anchor *plant* in English.

11.1. Construction of translation tables

In this phase we show, starting from the English Wikipedia, how to build a Translation Table (TT) for an arbitrary Wikipedia language.

A translation table can be seen as a sort of bilingual dictionary in which words of the source language are translated into words of a target language. In contrast to standard bilingual dictionaries, however, translation tables contain explicit probabilities associated with the translations of a given lemma. An excerpt of the Italian translation table (obtained as a result of this step) is shown in Table 7. Here, the (ambiguous) English lemma *plane* is translated into Italian as *piano cartesiano* (the x–y plane) with probability 0.20, as *piano* (the metaphoric sense of *plane*) with probability 0.15, as *pialla* (the carpenter's plane) with probability 0.04, as *aeroplano* (airplane) with probability 0.03, and so on.

In order to build the Translation Table (TT) for a given language, we consider all the linked tokens of Wikipedia. More formally, the TT in language L contains translation information for each English lemma l such that i) Wikipedia contains a linked occurrence of l (to a page p) and ii) p has an equivalent page p' in language L' and p' is linked by some foreign term l'. For instance, the Italian Translation Table contains translations for the lemma *plant* (cf. Fig. 16), but does not contain translations for the lemma *saucepan* (since this is always linked to the page SAUCEPAN, but the latter is a redirection and thus has no Italian equivalent). The input of the procedure is a word l_E, in the source language E and the output is a probability distribution $P(\cdot \mid l_E)$ over words in the target language F. We set up the problem of finding suitable translations for a given word by exploiting, on the one hand, the interlanguage links provided by Wikipedia, and, on the other hand, the association between Wikipedia pages and the associated textual anchors occurring in the whole Wikipedia. Fig. 16 shows this by means of an example. The data on the left side of the figure belong to the English Wikipedia, while the data on the right side belong to the Italian Wikipedia. Edges between a surface form and a page represent the fact that the former has been linked to the latter and numbers report the times the link occurs in Wikipedia. The English lemma *plant* on the left, for example, is linked 10,634 times to the Wikipedia page PLANT, 30 times to FACTORY and so on. The pages linked by an anchor represent the meanings that the given surface form can have in different contexts. A similar configuration can be seen on the right side of the figure, where Italian surface forms are linked to the corresponding meanings. Note, however, that in general an anchor can link to different meanings (*plant* pointing to PLANT, FACTORY, etc.) and a given page is linked by many anchors (the Italian page FABBRICA is pointed to by both *fabbrica* and *stabilimento*). Finally, interlanguage links are shown as undirected edges linking the two sides of the figure; for example the English page PLANT is aligned to the Italian PLANTAE and FACTORY to FABBRICA.

Our hunch is that this network can be exploited to derive translation probabilities. Starting from a given English lemma, in fact, it is possible to reach all its translations in another language by following the paths which join the two sides. For example, in order to infer that *pianta* is a valid translation for *plant*, it is sufficient to follow the graph pattern *plant* → PLANT – PLANTAE ← *pianta*. Each path is made of exactly three edges which represent, respectively, i) the association between the source anchor and one of the meanings it is linked to, ii) the interlanguage link between the source and the target meanings, and iii) the association between the target meaning and the target anchor. By calculating the paths between any pair of anchors (i.e., if we do this for all the source and target anchors) we can then obtain all the possible

Table 8
Top 10 languages in term of extracted lemmas.

Language	# of extracted lemmas	Average number of translations
DE	414.810	2.36
FR	381.845	1.99
IT	307.293	2.08
ES	265.629	2.15
JA	246.455	1.64
NL	223.528	1.50
RU	206.213	3.21
PL	195.017	2.91
PT	188.094	1.81
SV	165.207	1.59

translations. Note that in general, however, there might be more than a single path between any two anchors (for instance when two ambiguous anchors share the same senses across the two languages) and it is thus necessary to take into account all the paths which link the source anchor in a language to another anchor in the other language.

We are now ready to formally define the probabilities provided by a translation table. Given a lemma l_E in the source language E, we define the probability that the lemma l_F in the target language F is a translation for l_E as:

$$P(l_F|l_E) := \frac{1}{Z} \cdot \sum_{\substack{p_E \in O(l_E) \\ p_F \in O(l_F)}} P(p_E|l_E) \cdot P(p_F|p_E) \cdot P(l_F|p_F) \tag{1}$$

with:

$$P(p_E|l_E) := \frac{c(l_E \rightarrow p_E)}{\sum_{p'_E \in O(l_E)} c(l_E \rightarrow p'_E)} \tag{2}$$

$$P(p_F|p_E) := \begin{cases} 1 & \text{if there exists an interlanguage link from } p_E \text{ to } p_F \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$P(l_F|p_F) := \frac{c(l_F \rightarrow p_F)}{\sum_{l'_F \in I(p_F)} c(l'_F \rightarrow p_F)} \tag{4}$$

where Z is a normalization constant, $O(l_X)$ denotes the set of pages linked by l_X in language X, $I(p_X)$ denotes the set of surface forms pointing to p_X in language X and $c(l_X \rightarrow p_X)$ is the count of how many times l_X points to p_X in language X.

Equation (1) is a probability over all the paths going from l_E and ending in l_F. Each of the terms in the sum represents the probability of a single path and is made up of three terms: i) the probability of having l_E linking to p_E (first term $P(p_E|l_E)$, Equation (2)), ii) the probability of having p_E aligned to p_F (second term $P(p_F|p_E)$, Equation (3)), and iii) the probability of having p_F linked by a lemma l_F in that language (third term $P(l_F|p_F)$, Equation (4)).

For example, given the term $l_E = plant$, the probability $P(p_E = \text{PLANT} \mid l_E = plant)$ is .99, the probability $P(p_E = \text{FACTORY} \mid l_E = plant)$ is .001, while the probability $P(p_E = \text{PLANT (PERSON)} \mid l_E = plant)$ is .00023. Equation (1) in the case of $l_E = plant$ and $l_F = pianta$ (i.e., the probability that the English surface form *plant* translates into the Italian *pianta*) is:

$$P(pianta_{IT} \mid plant_{EN}) = \frac{1}{Z} \cdot \sum_{\substack{p_E \in O(plant_{EN}) \\ p_F \in O(pianta_{IT})}} P(p_E \mid plant_{EN}) \times P(p_I|p_E) \times P(pianta_{IT} \mid p_I)$$

The set $O(plant_{EN})$ includes for example PLANT, FACTORY and FLOWERING PLANT, among others. The only path with three non-zero factors is *plant* $\rightarrow$ PLANT – PLANTAE $\leftarrow$ *pianta*. Since $P(\text{PLANT} \mid plant) = .99$, $P(\text{PLANTAE}|\text{PLANT}) = 1$ and $P(pianta \mid \text{PLANTAE}) = .38$, the overall product $P(pianta \mid plant) = \frac{1}{1.0015} \times .99 \times 1 \times .38 = .37$.

In conclusion, as a result of the systematic application of our technique, we obtain one TT for each Wikipedia language. Each TT will then be used to associate hypernym lemmas with Wikipedia pages in the particular target language, as explained in the next subsection.

11.1.1. TT statistics

We now report some statistics for the TT tables obtained. Since the method is applicable to any language, we report the statistics also for languages for which an evaluation is not available. As can be seen in Table 8, German is the language with the highest number of extracted lemmas with an average of 2.36 translations for each given lemma. We also calculated the score distribution for the three languages evaluated in the paper (Italian, French and Spanish). We grouped the scores in buckets (0–0.1], (0.1, 0.2], …(0.9, 1] and plotted the distribution obtained in Fig. 17. On average, 29.55% of the translations have score in the (0.9, 1] interval and 43.39% of the translations have a score in the (0–0.1] interval; scores in (0.1–0.9] are almost equally divided among the remaining translations.

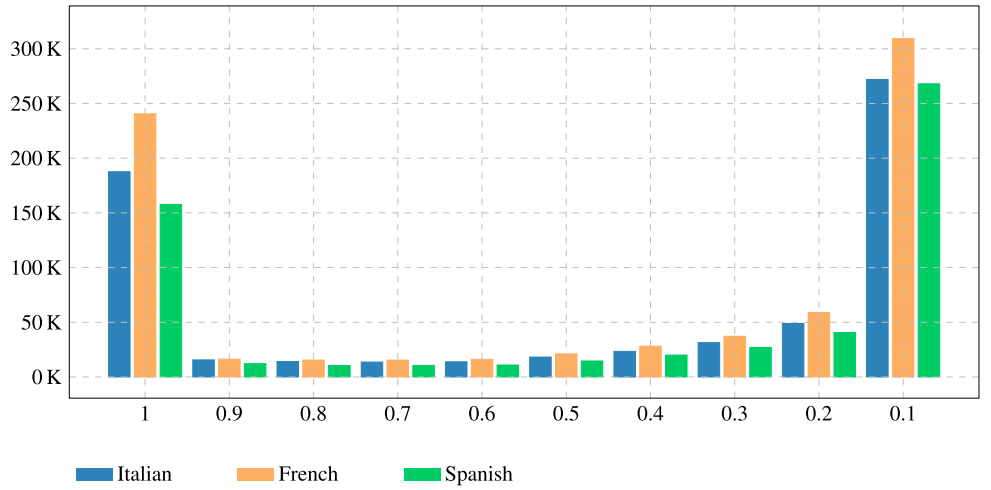


Fig. 17. Score distribution for Italian, French and Spanish translations.

11.2. Extraction of multilingual hypernym lemmas

We now describe how to exploit the Translation Tables in conjunction with the interlanguage links to provide hypernym lemmas for any page in an arbitrary Wikipedia language. The assignment of hypernym lemmas is based on strategies which exploit four different sources of information: i) the interlanguage links between two Wikipedia languages; ii) the TT and the local context of a given page (such as its textual definition, its categories, etc.); iii) the context provided by the sisters of a given page (basically, the distribution of hypernym lemmas of the sister pages); iv) global features of both Wikipedia and the hypernym lemmas discovered up to this point across all the Wikipedia pages. These strategies are applied in the same order of presentation, in cascade order.

Exploiting the interlanguage links (PROJECTED strategy) The first heuristic exploits the interlanguage links by means of the application of the projection rule. The hypernym lemma assigned to the page in the target language is the title lemma of the projected hypernym page. Thanks to this heuristic, for example, the Spanish Wikipedia page Madrid is assigned the hypernym lemma *ciudad*: the English Wikipedia page Madrid, aligned to the Spanish Madrid, has City as hypernym. The latter is, in turn, aligned to Ciudad and thus the title lemma of Ciudad (i.e., *ciudad*) is assigned to the Spanish starting page. However, note that this heuristic cannot cover concepts which are not encoded in the English Wikipedia (which are covered, instead, by subsequent strategies).

Exploiting the translation tables and the local context (TT strategy) This strategy draws on the TT presented in Section 11.1 and thus represents the first effort to automatically translate a hypernym lemma associated with an English page into its equivalent in a given language. At this step only local features are exploited, such as the page's textual definition and the titles of its categories. Starting from the English hypernym lemma l_E, this heuristic considers in decreasing order of probability all the translations of l_E and checks whether any of these is contained within the definition of p_F or in some of the category titles of p_F. For instance, the hypernym lemma for the Italian page Karl Popper is *filosofo*; since in English Karl Popper has *philosopher* as hypernym lemma, the heuristic considers all its translations, including *filosofo*, *filosofia*, *filosofi*, *filosofa*, etc. Given that the Italian definition for Karl Popper *"Popper è anche considerato un filosofo politico di statura considerevole, difensore della democrazia e del liberalismo [...]"* contains the translation *filosofo*, the latter is promoted to hypernym lemma of this page.

Exploiting context provided by sister pages (SISTER strategy) In order to also cover those pages of a language which do not have an equivalent in English, we designed another heuristic which draws on the sister pages of a given page and exploits the distribution of hypernym lemmas already discovered for these sister pages. The strategy considers in decreasing importance the distribution of hypernym lemmas of p_F's sisters and assigns to p_F the most frequent hypernym lemma which is contained in the definition of p_F or in some of the categories of p_F. For example, with this heuristic the Wikipedia French page Yahoo! Messenger is assigned the hypernym lemma *logiciel*, because it is contained in the following categories: Logiciel propriétaire, Logiciel de messagerie instantanée, Logiciel pour Mac OS and Logiciel pour Unix.

This strategy provides two added values: first of all, since it exploits sister pages, those pages which do not have an equivalent in English can also be covered. For example, this strategy succeeds in identifying *actrice* as hypernym lemma for the French page Stéphanie Reynaud, even though this page is available in French only. The other added value is that, because it exploits features which go beyond the mere textual definition of a page, it is able to extract suitable hypernyms even when the pages' definition does not contain the hypernym lemma, or contains a lemma which is less specific than expected (e.g., the definition for the page Platinum Tower is "El Platinum Tower es una lujosa edificación [...]" and the

lemma contained therein is *edificación*, which is less specific than the expected *rascacielos*, which, instead, is found thanks to the SISTER strategy).

Exploiting global features (F-IDF strategy) There is still a non-negligible fraction of Wikipedia WEE pages which, however, are as yet in their early stage of development, and thus suffer from lack of content. For example, more than 30% of Italian WEE pages lack a Wikipedia category, and 25% of these do not even have a definition. Note that the first strategy above cannot be applied to this class of pages, since there is no English equivalent. The TT and SISTER strategies are, instead, applicable: however, the former is effective only when some translation (provided by the TT) is contained in the textual definition of the target page p_F, while the latter is applicable only on those pages which have categories related to the hypernym lemma to be discovered, i.e., such that they bring in sister pages with a valid hypernym lemma. For example, the SISTER strategy is not useful for the Spanish page HEMISFERIO NORTE (i.e., NORTHERN HEMISPHERE in English), because its only category is GEOGRAFÍA (i.e., GEOGRAPHY in English).

To overcome this limitation, we introduce a measure called *f-idf* that resolves this problem by considering global information. This heuristic takes into account all the possible content available for a given page, by considering i) the page's textual definition (when present), ii) its categories (when present) and iii) the words of the title appearing within parentheses (e.g., the word *fiume* in the title TICINO (FIUME)). Given this context, all the possible n-grams are then collected (with $n \leq 5$), and the n-gram that maximises the following formula is promoted to hypernym lemma:

$$score(w) = f_w \cdot idf_w$$

where f_w is the frequency of the n-gram w as hypernym lemma (as obtained from the application of the previous three strategies to the whole Wikipedia) and idf_w (inverse definition frequency) is the logarithm of the inverse ratio of Wikipedia definitions containing w. The former prefers n-grams which are common hypernym lemmas across the whole Wikipedia lemma taxonomy built so far, the latter favours specific n-grams (such as the Italian *brano musicale*, whose idf is $\frac{1}{2761}$, vs. *brano*, whose idf is $\frac{1}{4298}$). For instance, consider the Italian page CERCAMI (RENATO ZERO), about a famous song of an Italian singer, whose textual definition is "Cercami è un famoso brano di Renato Zero, secondo singolo estratto dall'album Amore dopo amore del 1998". The n-grams extracted for this page include, among others, *Renato Zero*, *Cercami*, *brano*, *famoso brano*, *secondo*, etc. Among these, only four have also been assigned as hypernyms in the Italian taxonomy, namely *brano* (98 times, i.e., $f_{brano} = 98$), *singolo* ($f_{singolo} = 85$), *secondo* ($f_{secondo} = 16$) and *zero* ($f_{zero} = 7$). The corresponding idf are $\frac{1}{4298}$ for *brano*, $\frac{1}{9501}$ for *singolo*, $\frac{1}{15667}$ for *secondo* and $\frac{1}{660}$ for *zero*; *brano* is thus the n-gram which is finally preferred, with a score of $98 \cdot \frac{1}{4298} = 0.023$ and thus it becomes the candidate which is promoted as hypernym lemma of CERCAMI (RENATO ZERO).

Finally, for all those pages having an equivalent in English for which none of the above heuristics succeeded in assigning a hypernym lemma, we backoff to the MFT, the Most Frequent Translation l_F of l_E. For instance, the lemma assigned to the Spanish page FREDERICK LUGARD is *explorador*, since the latter is the translation with the highest probability for the English hypernym lemma *explorer*.

As a result of the application of the above four strategies to a non-English Wikipedia we are able to associate a hypernym lemma with almost the totality of the foreign pages.

11.3. Application of WiBi in the multilingual setting

Now that hypernym lemmas have also been extracted for each page in a specific non-English language, our aim is to disambiguate the hypernym lemmas so as to build a page taxonomy at the sense level, as well. Notably, we are in the same situation as we were in the English case, right before the application of the semantic step (see Section 4.2). We can therefore re-apply the very same hypernym linkers used for the English page taxonomy (cf. Section 4.2.1), with the exception of the distributional linker which assumes the availability of a PoS-tagger in the target language.

Thus, as was done in the English case (cf. Section 6), starting from the resulting page taxonomy we build a bitaxonomy in a non-English language (i.e., a taxonomy for the page side and a taxonomy for the category side of Wikipedia) by applying the iterative algorithm presented in Section 6.1.

In order to assess the impact of the interlanguage links over the final category taxonomy, we perform two kinds of experiment, yielding two different category taxonomies: in the first experiment we initialize the category taxonomy in exactly the same manner as explained in Section 6.2 (from here on, *plain* category taxonomy); in the second experiment we first apply the projection rule to all the hypernymy edges contained in the English category taxonomy and then perform the same initialization performed to obtain the plain category taxonomy (from here on, *projected* category taxonomy).

Finally, exactly as done in Section 7, we apply the refinement step on the non-English bitaxonomy obtained as a result of the application of the bitaxonomy algorithm.

11.4. Statistics for the multilingual setting

We now present the main statistics for the multilingual bitaxonomies obtained in the three languages, at lemma and sense level for the page taxonomies and for the two types of category taxonomies considered.

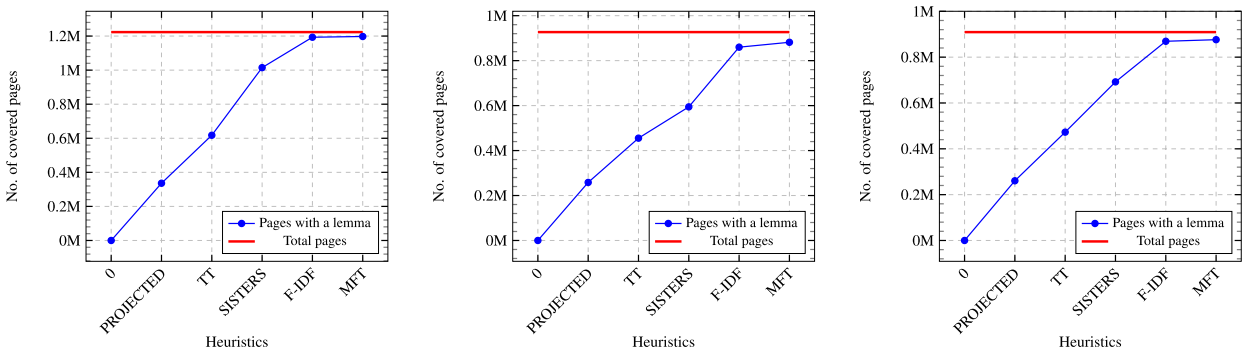


Fig. 18. Multilingual lemma taxonomy coverage with the different strategies.

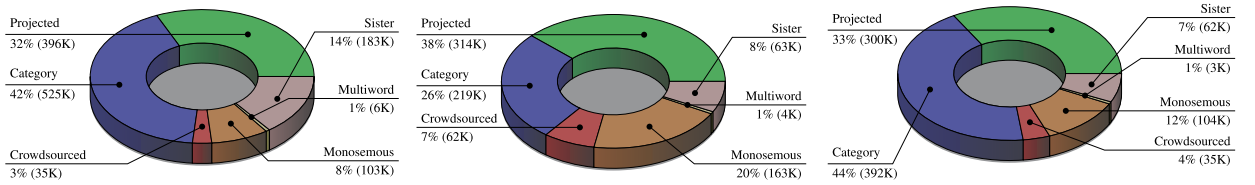


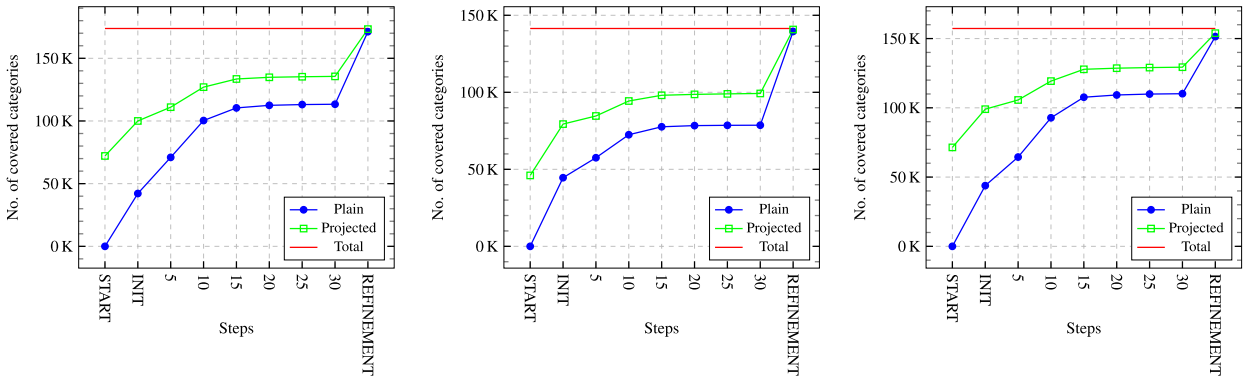
Fig. 19. Distribution of multilingual disambiguated hypernyms.

Hypernym lemmas Fig. 18 shows the coverage of the above heuristics for each considered language (French, Italian and Spanish). As can be seen, coverage increases consistently as more heuristics are applied, until approximately full coverage at lemma level is achieved in each language. The trend is very similar across languages and we attribute this phenomenon to a similar distribution of hypernym lemmas across the editions of Wikipedia.

Hypernym links Fig. 19 shows the distribution of the hypernym lemmas disambiguated by the various linkers. In contrast to the case of English, the pies include, on the one hand, the hypernym pages coming from the application of the projection rule, while, on the other hand, they do not display information regarding the distributional linker. Apart from the projected edges, the distribution of hypernyms by linker varies depending on the language considered. This does not hold for the projected is-a information, since we have seen that the overlap between English and the three languages does not change significantly and we expect the same amount of information to be transferred across languages. As regards the other linkers, the is-a edges returned by the category linker represent a substantial fraction of the total. In terms of number of links, the impact of the monosemous linker is comparable to that of the distributional linker in the English case, while the multiword linker proves to be marginally contributing. Overall, we extracted 1,246,524 is-a relations for French, 825,465 for Italian and 895,301 for Spanish, providing hypernym pages respectively for 1,116,330 pages out of 1,221,845 (91%), 742,796 pages out of 926,129 (80%) and 809,410 pages out of 908,820 (89%).

Bitaxonomies In Fig. 20 we show the coverage trend of the categories when applying the iterative algorithm to the plain (blue, lower line) and the projected (green, upper line) category taxonomy. Note that, similarly to the case of English, coverage progressively increases until iteration 30, where it finally reaches a plateau. Thanks to the application of the category refinement step, the gap with respect to the total is greatly reduced, reaching approximately full coverage of all Wikipedia categories. As can be seen in the figure (x-label START), in the projected category taxonomies about one third of the categories are already covered for all the languages. Interestingly, though, after the bitaxonomy algorithm and the category refinement step have been applied, the two lines reconcile approximately at the same point, meaning that starting with a projected category taxonomy does not necessarily result in a significantly greater coverage. However, in Section 12.4 we show that the projected taxonomy leads to generally better performance overall.

Analysis of the page taxonomies across languages After running MultiWiBi on a non-English edition of Wikipedia what we obtain is an augmented bitaxonomy which overlaps with the concepts contained in the English Wikipedia to a certain extent, but which also differs significantly from it. We thus distinguished four types of taxonomised pages:



(a) French (b) Italian (c) Spanish

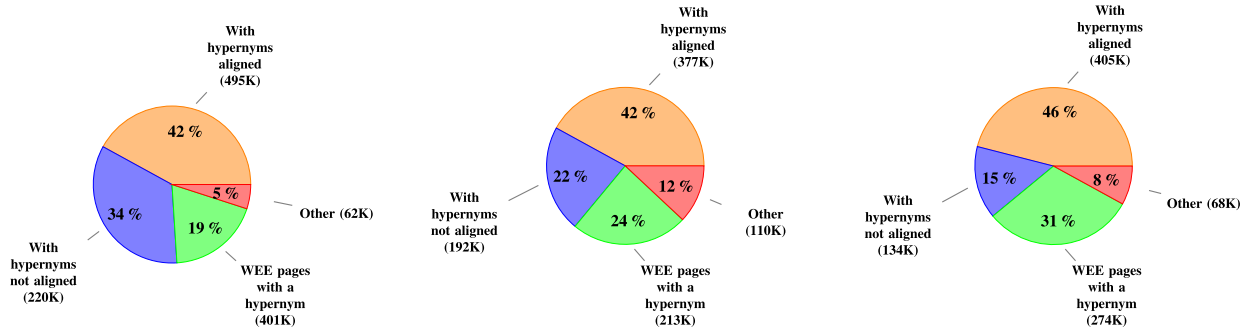
Fig. 20. Multilingual category taxonomy coverage over the iterations. (For interpretation of the colours in this figure, the reader is referred to the web version of this article.)

- **Pages with hypernyms aligned**. Pages which *do* have a corresponding page in English and whose hypernyms all coincide with those found in English (i.e., have the same hypernym aligned across languages). For example, the Italian page SCIROPPO D'ACERO, aligned to the English MAPLE SYRUP, has the Italian page SCIROPPO as hypernym, aligned in turn to the English SYRUP.
- **Pages with hypernyms not semantically aligned**. Pages which *do* have a corresponding page in English, but the concept expressed by their hypernyms differs with respect to that represented by the equivalent concept in English. This happens when the hypernym in English is a redirection (which is not possible to project) or has a different specificity. An example for the first case is the French page ARNOLFO DI CAMBIO which has the hypernym SCULPTEUR in French and the hypernym SCULPTOR in English, but the two are not aligned across the two languages because they are both redirections; the hypernyms returned in the two languages, while expressing exactly the same concept, are considered different at the surface level only because no equivalence has been established in Wikipedia so far. For the second case, the French page RICHARD II D'ANGLETERRE has MONARQUE as hypernym but KING OF ENGLAND in English (note that KING OF ENGLAND has no equivalent in French, but MONARQUE aligns to MONARCH in English).
- **WEE pages with a hypernym**. Pages which *do not* have a corresponding page in English and for which one or more hypernyms have been found. This class is of extreme importance, because hypernyms belonging to this class are unique, in the sense that this information could not be derived from English by means of the simple projection rule. The pages included in this class are usually (though not always, see discussion in Section 11) culture-specific: here we find the Italian Wikipedia page SAN GIMIGNANO (VINO) about a famous Italian wine which has no English counterpart and which has been taxonomised as VINO (WINE).
- **Other**. Finally, there is a fraction of pages which have not been taxonomised; for example, the Italian E PENSO A TE (ALBUM) has no English counterpart and MultiWiBi did not succeed in taxonomising it (even though the hypernym lemma extraction step managed to associate the lemma *album*).

In Fig. 21 we plot the distribution of the four types of page in the different languages: roughly 60% of the taxonomised pages are also lexicalized in English (orange and blue slices) and, within this set we distinguish i) the pages whose hypernyms are aligned with English (orange sector) and ii) pages whose hypernyms are not semantically aligned (blue sector). As regards WEE pages, instead, MultiWiBi managed to disambiguate the hypernym lemmas for 401k, 213k and 274k pages in French, Italian and Spanish, which represent 86%, 66% and 80%, respectively, of the total number of WEE pages in the three languages. This is a highly significant piece of information as it quantifies the amount of potentially culture-specific knowledge that MultiWiBi is able to extract. Just to give the reader a clearer grasp of the ground-breaking effect achieved by covering this type of concept, in Table 9 we report a sample list of WEE pages for French, Italian and Spanish for which MultiWiBi found suitable hypernyms. Even though these might also be contained in other non-English Wikipedias, they have the characteristic of not being lexicalized in English and thus represent additional concepts which MultiWiBi succeeded in taxonomizing.

12. Multilingual evaluation

We now present the experimental setup of our multilingual experiments and report the results of the multilingual bitaxonomies, evaluated with the same measures described in Section 5. We describe the creation of the gold standards in Section 12.1, present and discuss results at lemma level in Section 12.2 and at sense level in Sections 12.3 and 12.4 for the page and the category sides, respectively.



(a) French (b) Italian (c) Spanish

Fig. 21. Characterization of pages and their hypernyms in other languages.

Table 9
Example of culture-specific concepts.

French		Italian		Spanish	
Page	Hypernym	Page	Hypernym	Page	Hypernym
Langhe Chardonnay	Vin	Trenette al pesto	Pasta	Crema catalana	Postre
Diplomate (pâtisserie)	Pâtisserie	Lasagne (gastronomia)	Pasta	Baldomero Fernández Moreno	Poeta
Savarin	Pâtisserie	Maltagliati	Pasta	Luis García Montero	Poeta
Le Brebiou	Fromage	Piemonte Brachetto	Vino	El Rey (canción)	Canción
Fromages en Provinois	Fromage	Nebbione	Vino	Cantares (canción de J. M. Serrat)	Canción
Chécy (fromage)	Fromage français	Zagarolo (vino)	Vino	Mediterráneo (canción)	Canción

12.1. Experimental setup

In order to guarantee comparability of results across languages, we decided to start from the datasets presented in Section 5. For each language we thus created a dataset over two types of page:

- *Projected pages*: this part of the dataset was obtained by automatically projecting the English dataset and then manually fixing projection errors. Since not all the pages are aligned across all language editions of Wikipedia, full projection coverage could not be obtained. After the projection we thus obtained 256 pages for French, 218 for Italian and 205 for Spanish.
- *WEE pages*: the second part of the dataset was obtained by sampling a certain number of Wikipedia pages without English equivalent. This number was chosen in order to preserve the balance between concepts expressed in both languages (i.e., the set $P_E \cap P_F$) and concepts existing only in the other language (i.e., the set $P_F \setminus P_E$). From now on, these datasets will be called D_F^{WEE} (dataset of WEE pages in language F).

For building datasets at lemma level, due to the lack of interlanguage links between lemmas, we manually provided hypernym lemmas for each page in the dataset, exactly as was done for English. At sense level, instead, we decided to double the size of the page dataset (with and without English equivalent) for Italian. Italian was chosen because it was the most problematic language among the three languages considered, having many ill-formed definitions that lack an explicit hypernym. Thus the Italian dataset at sense level numbered 436 items for pages with English equivalent and 232 for WEE pages. We also computed the inter-annotator agreement (using Cohen's kappa coefficient) for the Spanish and French datasets, scoring 0.825 and 0.811, respectively.

12.2. Results for multilingual hypernym lemmas

In this section we provide experimental results about the quality of the taxonomies at lemma level. In particular, we i) compare the quality of the automatic translation procedure (described in Section 11.2) against a tool-based syntactic lemma extraction (similarly to what was done in the English setting, see Section 4.1) and ii) analyse and discuss the results when considering pages with and without English equivalents (Projected pages and WEE pages, respectively).

As can be seen from Table 10, MultiWiBi achieves very good results. As regards the projected datasets, we can see that consistently more than 99% of pages have at least one lemma, which means that nearly all the pages in the three target languages are covered at lemma level. Also quality is very high: while in Italian precision and recall are around 70%, in the other two languages they are both very high, between 80% and 85%. Lower performances in Italian are mostly due to pages with no categories associated and ill-formed textual definitions lacking an explicit hypernym lemma (e.g., Flavio Crespi has no Wikipedia categories and is defined by means of the textual definition "Pratica l'arrampicata in falesia e ha gareggiato nelle competizioni di difficoltà." in which the hypernym lemma *arrampicatore* is totally implicit). This means that the strate-

Table 10
Multilingual lemma taxonomy quality.

Language	Setting	Approach	P	R*	C	# items
FR	Projected pages	MultiWiBi	81.76	82.42	99.61	256
		Syntactic	82.41	69.53	84.38	
		Syntactic + sisters	78.69	75.00	95.31	
	WEE pages	MultiWiBi	76.67	74.19	96.77	155
		Syntactic	90.68	69.03	76.13	
		Syntactic + sisters	86.90	81.29	93.55	
IT	Projected pages	MultiWiBi	76.46	77.98	99.08	218
		Syntactic	–	–	–	
		Syntactic + sisters	–	–	–	
	WEE pages	MultiWiBi	58.77	57.76	98.28	116
		Syntactic	–	–	–	
		Syntactic + sisters	–	–	–	
ES	Projected pages	MultiWiBi	83.95	84.39	99.51	205
		Syntactic	70.00	65.37	90.24	
		Syntactic + sisters	67.57	68.78	98.54	
	WEE pages	MultiWiBi	74.55	66.67	89.43	123
		Syntactic	73.37	52.85	70.73	
		Syntactic + sisters	70.22	62.60	87.80	

gies presented in Section 11.2 managed either to translate (whenever possible) or extract meaningful hypernym lemmas for the considered pages. As regards WEE datasets, we witness a general decrease in performance; here Italian achieves performances below 60%, while French and Spanish achieve precision around 75% and recall between 66% and 74%.

Now, what if, as we did for the English version, we used a language-specific syntactic parser to extract hypernym lemmas for the Wikipedia pages in another language? To test whether the lemma taxonomy actually benefits from using a syntactic parser, for each specific version of Wikipedia we syntactically extracted all the terms involved in a dependency relation corresponding to the English *copula*. Since the names of the dependency relations change across languages, label mappings were provided manually: in French, for instance, we identified the *ast* relation of the Malt syntactic parser[15]; in Spanish we used the *att* relation output by the FreeLing syntactic parser.[16] As was done in English, whenever possible, we also employed the list of class hypernym lemmas (cf. Section 4.1), also manually translated from English (e.g., *variety* was translated into *variedad* in Spanish and *variété* in French), and exploited the relations corresponding to the English *conj_ and* and *conj_or* relations as well (*coord* in French and *co-n* in Spanish). Unfortunately, it was not possible to find a syntactic parser for Italian. Rows 'Syntactic' and 'Syntactic + sisters' in Table 10 report the performances when using the *vanilla* and the *sister* syntactic settings (cf. Section 4.1 and Fig. 7a). Similarly to the English case, including the context coming from the sister pages greatly boosts coverage in all languages while improving recall at the same time, to the slight detriment of precision. Except for the French WEE pages, where using a syntactic parser seems to improve MultiWiBi performance, all other scenarios show that MultiWiBi yields comparable or even better results than using a syntactic parser, for all the quality measures. In conclusion, we can say that the automatic inference of hypernym lemmas from the English taxonomy provides better hypernym translations overall than using a monolingual approach, with a significant increase in precision and recall when compared to the setting in which a syntactic parser in the target language is exploited. This phenomenon is likely due to two factors: on the one hand, the heuristics used to project the taxonomical information exploit more context than that made available to the syntactic parsers at the monolingual level, on the other hand, these syntactic parsers might not be as mature as the English Stanford counterpart, e.g., in need of more extensive training data as input.

Translation table evaluation In order to assess the quality of our translation table we decided to compare it against Moses, a statistical machine translation tool, trained on the Europarl parallel corpus[17] from English to Italian. We built a new version of the Italian lemma taxonomy using Moses's translation table as candidate hypernym lemma repository and we evaluated it considering only the hypernyms assigned by heuristics which exploit the translation table. The taxonomy built with our TT achieves 71.23% precision, 27.98% recall and a coverage of 38.53%, while using the TT obtained with Moses results in performances that are somewhat lower: 61.36% precision, 25.23% recall and 40.37% coverage. While we acknowledge that Moses could be trained on larger parallel corpora, the issues of domain specificity (e.g. law and institutional topics) and lack of coverage of many areas of knowledge would need to be dealt with. Instead, not only does our approach not depend on the availability of parallel corpora, which are generally missing for each Wikipedia language pair, but it also performs equally well across domains.

[15] http://www.maltparser.org.
[16] http://nlp.lsi.upc.edu/freeling/.
[17] http://www.statmt.org/europarl/v7/it-en.tgz.

Table 11
Multilingual page taxonomy evaluation.

Language	Resource	P	R*	C	# items
FR	MultiWiBi	84.51	80.86	94.14	256
	MENTA	81.37	48.83†	59.77	
	DBpedia	94.08	28.52†	29.69	
	YAGO3	94.00	80.86	80.86	
IT	MultiWiBi	80.06	79.36	96.33	436
	MENTA	79.73	53.21†	66.74	
	DBpedia	98.47	59.17†	60.09	
	YAGO3	92.38	83.49	83.49	
ES	MultiWiBi	86.98	81.95	93.66	205
	MENTA	81.02	42.93†	52.68	
	DBpedia	76.77†	37.07†	48.29	
	YAGO3	85.86	78.54	78.54	

† denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and the daggered resource.

Table 12
Multilingual WEE page taxonomy evaluation.

Language	Setting	P	R*	C	# items
FR	MultiWiBi	76.39	70.97	92.90	155
	MENTA	85.71	34.84†	40.65	
	DBpedia	100.00	16.77†	16.77	
	YAGO3	89.33	58.71	58.71	
IT	MultiWiBi	59.88	44.40	74.14	232
	MENTA	87.91	34.48	39.22	
	DBpedia	99.08	45.26	45.69	
	YAGO3	98.40	22.41	22.41	
ES	MultiWiBi	69.47	53.66	77.24	123
	MENTA	76.19	26.02†	34.15	
	DBpedia	66.13	33.33†	50.41	
	YAGO3	91.33	35.77	35.77	

† denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and the daggered resource.

12.3. Results for multilingual page taxonomies

We now move from evaluation of the hypernym lemmas to evaluation of their associated pages. Results for the multilingual page taxonomies are presented in Table 11 for the three languages, also compared with the alternative approaches, namely DBpedia, YAGO3 and MENTA. LHD was excluded from the comparison because it was not available in any of the three languages, while WikiNet was excluded because the average coverage on the normal datasets and the D_F^{WEE} datasets was 49.51% and 5.46%, respectively.

As can be seen, performances are very high for all the languages when the projected datasets are considered and, to a certain extent, they are comparable to the results obtained in English. For all the three languages we observe around 85% precision, 80% recall and 93% coverage. This result might be expected considering that these datasets are the projected version of their English equivalents. Nevertheless, it needs to be borne in mind that, after the application of the hypernym linkers, the hypernyms found by MultiWiBi in English and in another language might differ considerably (see Section 11.4), causing relatively small differences in the evaluation results. YAGO3 performs well, but even if its precision is very high, its coverage is lower than MultiWiBi. Furthermore, YAGO3's high recall is due to the fact that at least one very generic hypernym (e.g. *yagoGeoEntity*, *owl:Thing*) is given as output for each page, which we evaluated as correct independently of its generality (see Section 10.4 for an assessment of YAGO3's degree of specificity). Other systems, instead, are seriously affected by low coverage and, even when achieving precision comparable to (or higher than) MultiWiBi, exhibit a recall which is well below 60%.

We also show results for the D_F^{WEE} datasets in Table 12. We observe that it was not possible to obtain performance comparable to that of English. First of all we point out that results vary depending on the target language: for example, French exhibits very good results, achieving 76% precision and around 71% recall. Italian and Spanish have lower coverage than French and also lower performance in general. Spanish, however, despite being affected by low recall, stands up well, with a quite high 69.47% precision.

These results are, however, very promising because MultiWiBi is the very first resource providing such information for concepts with no English equivalent. As can be seen from the table, in fact, all the alternative approaches suffer from critical coverage problems. Even YAGO3, which performed well on the pages with English equivalent, has lower recall and coverage than MultiWiBi. In addition to this, the other approaches' answers, even when correct, are often either very general

Table 13
Multilingual category taxonomy evaluation.

Language	Resource	P	R*	C	# items
FR	MultiWiBi (plain)	78.07	74.17	95.00	140
	MultiWiBi (projected)	80.71	80.71	100.00	
	MENTA	82.61	55.00†	65.71	
IT	MultiWiBi (plain)	89.70	89.00	99.20	500
	MultiWiBi (projected)	83.54	83.54	100.00	
	MENTA	77.13	25.40†	32.80	
ES	MultiWiBi (plain)	84.75	81.97	96.72	131
	MultiWiBi (projected)	84.43	84.43	100.00	
	MENTA	80.46	54.20†	66.41	

† denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and the daggered resource.

(e.g., DBpedia returning http://dbpedia.org/ontology/Monument instead of ÉGLISE (ÉDIFICE)) or uninformative (e.g., MENTA returning 1970 BIRTHS instead of ÉCRIVAIN FRANÇAIS). MultiWiBi thus represents the first approach which manages to extract language-specific information automatically from Wikipedia, with performances, however, that are dependent upon the quality of the individual Wikipedias.

12.4. Results for multilingual category taxonomies

In Table 13 we report the performance when evaluating the plain and the projected MultiWiBi category taxonomy against the gold standard (see Table 13). The only alternative resource we compared with is MENTA, because it proved closest in terms of performance in the English experimental setup. Note also that YAGO, YAGO3 and WikiTaxonomy could not be compared, because no multilingual version of these resources exists (see Section 10.1).

We can see that projecting the category taxonomy from English greatly benefits the category taxonomies in the other languages. First, by means of the automatic projection, we achieved full coverage on the category dataset. Second, except for precision in Italian and Spanish, all other measures exhibit a remarkable increase, ranging between 1.27 and 5.83 percentage points.

13. Conclusions

In this paper we have presented MultiWiBi, a new approach for constructing bitaxonomies of Wikipedia in arbitrary languages, where each bitaxonomy is made up of two taxonomies which establish is-a relations between Wikipedia pages and categories, respectively. For each language, the approach is mainly divided into three phases. The first phase aims at building a taxonomy for the page side of Wikipedia; the second phase triggers an iterative algorithm that incrementally populates a taxonomy for the category side of Wikipedia by exploiting the interlanguage links existing between the two sides; the third phase is aimed at solving some problems affecting the structure of Wikipedia categories so as to output a polished category taxonomy.

Our contribution is three-fold. First, the two taxonomies of each bitaxonomy are aligned (pages are aligned to categories) and the bitaxonomies are also aligned across languages (concepts in English are aligned to the corresponding concepts in all other languages). Second, in marked contrast to other approaches, our work crucially pivots on the English edition of Wikipedia for inducing bitaxonomies in the other languages, without relying on any external resource (like WordNet or manual upper ontologies), parallel corpus or tool. Third, our experiments show that our bitaxonomies are characterized by higher accuracy and specificity than all other alternatives, making MultiWiBi the best set of taxonomies in the literature at the time of writing.

We have also integrated MultiWiBi into BabelNet,[18] thanks to which a full-fledged taxonomy of lexicographic and encyclopedic knowledge is now available. In order to maintain a high precision we integrated all those edges in MultiWiBi for which taxonomies in different languages agree. We plan to exploit MultiWiBi in several applications such as multilingual distributional semantics [62,63] and question answering.

Acknowledgements

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

We thank Luca Telesca for his implementation of WikiTaxonomy and Jim McManus for his comments on the manuscript.

[18] http://babelnet.org.

References

[1] T. Mitchell, Reading the Web: a breakthrough goal for AI, AI Mag. (2005) 1517–1519.
[2] S. Mirkin, I. Dagan, E. Shnarch, Evaluating the inferential utility of lexical-semantic resources, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09, Athens, Greece, 2009, pp. 558–566.
[3] H. Poon, J. Christensen, P. Domingos, O. Etzioni, R. Hoffmann, C. Kiddon, T. Lin, X. Ling, Mausam, A. Ritter, S. Schoenmackers, S. Soderland, D. Weld, F. Wu, C. Zhang, Machine reading at the University of Washington, in: Proceedings of the 1st International Workshop on Formalisms and Methodology for Learning by Reading in Conjunction with NAACL-HLT 2010, Los Angeles, California, USA, 2010, pp. 87–95.
[4] A. Singhal, Introducing the knowledge graph: things, not strings, Tech. rep., Official Blog (of Google), 2012. Retrieved May 18, 2012.
[5] D.A. Ferrucci, Introduction to "This is Watson", IBM J. Res. Dev. 56 (3) (2012) 1.
[6] C. Fellbaum (Ed.), WordNet: An Electronic Database, MIT Press, Cambridge, MA, 1998.
[7] P. Mcnamee, R. Snow, P. Schone, Learning named entity hyponyms for question answering, in: Proceedings of the Third International Joint Conference on Natural Language Processing, 2008, pp. 799–804.
[8] D. Moldovan, A. Novischi, Lexical chains for question answering, in: Proceedings of the 19th International Conference on Computational Linguistics, COLING '02, Taipei, Taiwan, 24 August–1 September 2002, 2002, pp. 1–7.
[9] H. Cui, M.-Y. Kan, T.-S. Chua, Soft pattern matching models for definitional question answering, ACM Trans. Inf. Syst. 25 (2) (2007) 1–30.
[10] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A.A. Kalyanpur, A. Lally, J.W. Murdock, E. Nyberg, J. Prager, et al., Building Watson: an overview of the deepqa project, AI Mag. 31 (3) (2010) 59–79.
[11] O. Etzioni, M. Banko, M.J. Cafarella, Machine reading, in: Proc. of AAAI, 2006, pp. 1517–1519.
[12] T. Lin, O. Etzioni, et al., Entity linking at web scale, in: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction, AKBC-WEKEX '12, Association for Computational Linguistics, 2012, pp. 84–88.
[13] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: a unified approach, Trans. Assoc. Comput. Linguist. 2 (2014) 231–244.
[14] C. Delli Bovi, L. Telesca, R. Navigli, Large-scale information extraction from textual definitions through deep syntactic and semantic analysis, Trans. Assoc. Comput. Linguist. 3 (2015) 529–543.
[15] A. Moro, H. Li, S. Krause, F. Xu, R. Navigli, H. Uszkoreit, Semantic rule filtering for web-scale relation extraction, in: The Semantic Web – ISWC 2013 – Proceedings of the 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21–25, 2013, Part I, 2013, pp. 347–362.
[16] 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, COLING '06, Sydney, Australia, 17–21 July 2006, 2006, pp. 793–800.
[17] S. Soderland, B. Mandhani, Moving from textual relations to ontologized relations, in: AAAI Spring Symposium: Machine Reading, AAAI, 2007, pp. 85–90.
[18] G. Sutcliffe, M. Suda, A. Teyssandier, N. Dellis, G. de Melo, Progress towards effective automated reasoning with world knowledge, in: FLAIRS Conference, 2010, pp. 110–115.
[19] M. Pasca, S. Harabagiu, The informative role of Wordnet in open-domain question answering, in: Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources, 2001, pp. 138–143.
[20] R. Navigli, Word Sense Disambiguation: a survey, ACM Comput. Surv. 41 (2) (2009) 1–69.
[21] R. Navigli, A quick tour of Word Sense Disambiguation, induction and related approaches, in: M. Bieliková, G. Friedrich, G. Gottlob, S. Katzenbeisser, G. Turán (Eds.), SOFSEM 2012: Theory and Practice of Computer Science, in: Lecture Notes in Computer Science, vol. 7147, Springer, Heidelberg, 2012, pp. 115–129.
[22] O. Medelyan, D. Milne, C. Legg, I.H. Witten, Mining meaning from Wikipedia, Int. J. Hum.-Comput. Stud. 67 (9) (2009) 716–754.
[23] E.H. Hovy, R. Navigli, S.P. Ponzetto, Collaboratively built semi-structured content and Artificial Intelligence: the story so far, Artif. Intell. 194 (2013) 2–27.
[24] 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, IJCAI '07, Hyderabad, India, 6–12 January 2007, 2007, pp. 2670–2676.
[25] A. Fader, S. Soderland, O. Etzioni, Identifying relations for open information extraction, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, Edinburgh, UK, 2011, pp. 1535–1545.
[26] A. Moro, R. Navigli, Integrating syntactic and semantic analysis into the open Information extraction paradigm, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI '13, Beijing, China, 2013, pp. 2148–2154.
[27] A. Gómez-Pérez, D. Manzano-Macho, et al., A survey of ontology learning methods and techniques, OntoWeb Deliverable D 1 (5), 2003.
[28] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proceedings of the 25th International Conference on Computational Linguistics, COLING '92, Nantes, France, 1992, pp. 539–545.
[29] S.P. Ponzetto, M. Strube, Deriving a large scale taxonomy from Wikipedia, in: Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence, AAAI '07, Vancouver, B.C., Canada, 22–26 July 2007, 2007, pp. 1440–1445.
[30] J. Hoffart, F.M. Suchanek, K. Berberich, G. Weikum, YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia, Artif. Intell. 194 (2013) 28–61.
[31] V. Nastase, M. Strube, Transforming Wikipedia into a large scale multilingual concept network, Artif. Intell. 194 (2013) 62–85.
[32] G. de Melo, G. Weikum, MENTA: inducing multilingual taxonomies from Wikipedia, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, New York, NY, USA, 2010, pp. 1099–1108.
[33] T. Kliegr, V. Zeman, M. Dojchinovski, Linked hypernyms dataset-generation framework and use cases, in: 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing, 2014, p. 82.
[34] T. Flati, D. Vannella, T. Pasini, R. Navigli, Two is bigger (and better) than one: the Wikipedia bitaxonomy project, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 945–955.
[35] R. Navigli, P. Velardi, Learning word-class lattices for definition and hypernym extraction, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 1318–1327.
[36] R. Navigli, S.P. Ponzetto, BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artif. Intell. 193 (2012) 217–250.
[37] N. Calzolari, L. Pecchia, A. Zampolli, Working on the Italian machine dictionary: a semantic approach, in: Proceedings of the 5th Conference on Computational Linguistics, COLING '73, Pisa, Italy, 1973, pp. 49–70.
[38] R.A. Amsler, A taxonomy for English nouns and verbs, in: Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics, ACL '81, Stanford, California, USA, 1981, pp. 133–138.
[39] N. Calzolari, Towards the organization of lexical definitions on a database structure, in: Proceedings of the 9th Conference on Computational Linguistics, COLING '82, Prague, Czechoslovakia, 1982, pp. 61–64.
[40] N. Ide, J. Véronis, Extracting knowledge bases from machine-readable dictionaries: have we wasted our time?, in: Proceedings of the Workshop on Knowledge Bases and Knowledge Structures, KB&KS '93, Tokyo, Japan, 1993, pp. 257–266.
[41] Z. Kozareva, E.H. Hovy, A semi-supervised method to learn and construct taxonomies using the Web, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '10, Seattle, WA, USA, 2010, pp. 1110–1118.

[42] R. Navigli, P. Velardi, S. Faralli, A graph-based algorithm for inducing lexical taxonomies from scratch, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, 2011, pp. 1872–1877.
[43] P. Velardi, S. Faralli, R. Navigli, OntoLearn reloaded: a graph-based algorithm for taxonomy induction, Comput. Linguist. 39 (3) (2013) 665–707.
[44] D. Klein, C.D. Manning, Fast exact inference with a factored model for natural language parsing, in: Advances in Neural Information Processing Systems, vol. 15, NIPS, Vancouver, British Columbia, Canada, 2003, pp. 3–10.
[45] H. Saggion, Identifying definitions in text collections for question answering, in: Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC '04, Lisbon, Portugal, 26–28 May 2004, European Language Resources Association, 2004, pp. 1927–1930.
[46] 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.
[47] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates, Web-scale information extraction in KnowItAll, in: Proceedings of the 13th International Conference on World Wide Web, WWW '04, 2004, pp. 100–110.
[48] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates, Unsupervised named-entity extraction from the web: an experimental study, Artif. Intell. 165 (1) (2005) 91–134.
[49] S. Blohm, Using the web to reduce data sparseness in pattern-based information extraction, in: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD, Springer, Warsaw, Poland, 2007, pp. 18–29.
[50] 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, NAACL HLT '13, Boston, Massachusetts, 2–7 May 2004, 2004, pp. 321–328.
[51] R. Snow, D. Jurafsky, A. 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, COLING-ACL '06, 2006, pp. 801–808.
[52] S.P. Ponzetto, M. Strube, Taxonomy induction based on a collaboratively built knowledge repository, Artif. Intell. 175 (9–10) (2011) 1737–1756.
[53] F.M. Suchanek, G. Kasneci, G. Weikum, YAGO: a large ontology from Wikipedia and WordNet, J. Web Semant. 6 (3) (2008) 203–217.
[54] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ive, DBpedia: a nucleus for a web of open data, in: Proceedings of 6th International Semantic Web Conference Joint with 2nd Asian Semantic Web Conference, ISWC+ASWC 2007, Busan, Korea, 2007, pp. 722–735.
[55] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the International Conference on Management of Data, SIGMOD '08, New York, NY, USA, 2008, pp. 1247–1250.
[56] V. Nastase, M. Strube, B. Boerschinger, C. Zirn, A. Elghafari, WikiNet: a very large scale multi-lingual concept network, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10, Valletta, Malta, 2010, pp. 1015–1022.
[57] A. Sumida, N. Yoshinaga, K. Torisawa, Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia, in: LREC, European Language Resources Association, 2008, pp. 2462–2469.
[58] M.T. Pilehvar, R. Navigli, A robust approach to aligning heterogeneous lexical resources, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 468–478.
[59] E. Niemann, I. Gurevych, The people's web meets linguistic knowledge: automatic sense alignment of Wikipedia and Wordnet, in: Proceedings of the Ninth International Conference on Computational Semantics, IWCS '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 205–214.
[60] I. Gurevych, J. Eckle-Kohler, S. Hartmann, M. Matuschek, C.M. Meyer, C. Wirth, UBY: a large-scale unified lexical-semantic resource based on LMF, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, Stroudsburg, PA, USA, 2012, pp. 580–590.
[61] M.M. Christian, G. Iryna, To exhibit is not to loiter: a multilingual, sense-disambiguated wiktionary for measuring verb similarity, in: Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012, vol. 4, 2012, pp. 1763–1780.
[62] J. Camacho-Collados, M.T. Pilehvar, R. Navigli, Nasari: a novel approach to a semantically-aware representation of items, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 567–577.
[63] J. Camacho-Collados, M.T. Pilehvar, R. Navigli, Nasari: integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities, Artif. Intell. 240 (2016) 36–64.