Artificial Intelligence 193 (2012) 217–250

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network

Roberto Navigli *, Simone Paolo Ponzetto

Dipartimento di Informatica, Sapienza University of Rome, Italy

ARTICLE INFO

Article history:
Received 30 August 2011
Received in revised form 26 April 2012
Accepted 1 July 2012
Available online 1 August 2012

Keywords:
Knowledge acquisition
Word sense disambiguation
Graph algorithms
Semantic networks

ABSTRACT

We present an automatic approach to the construction of BabelNet, a very large, wide-coverage multilingual semantic network. Key to our approach is the integration of lexicographic and encyclopedic knowledge from WordNet and Wikipedia. In addition, Machine Translation is applied to enrich the resource with lexical information for all languages. We first conduct in vitro experiments on new and existing gold-standard datasets to show the high quality and coverage of BabelNet. We then show that our lexical resource can be used successfully to perform both monolingual and cross-lingual Word Sense Disambiguation: thanks to its wide lexical coverage and novel semantic relations, we are able to achieve state-of the-art results on three different SemEval evaluation tasks.

1. Introduction

In the information society, knowledge – i.e., the information and expertise needed to understand any subject of interest – is a key skill for understanding and decoding an ever-changing world. Much information is conveyed by means of linguistic communication (either oral or written), therefore it is critical to know how words are used to express meaning, i.e., we need lexical knowledge. The typical example is that of a foreign language learner, who needs as much lexical knowledge as possible in order to understand communications expressed in the foreign language. However, lexical knowledge is an essential component not only for human understanding of text, but also for performing language-oriented automatic tasks effectively. Most, if not all, areas of Natural Language Processing (NLP) have been shown to benefit from the availability of lexical knowledge at different levels. These include, among others, text summarization [82], named entity disambiguation [16], Question Answering [48,66], text categorization [40,130,88], coreference resolution [107,111], sentiment analysis [125,129] and plagiarism detection [10]. There is clear evidence in the literature that the amount and quality of knowledge heavily impacts even difficult tasks such as Word Sense Disambiguation [26,28,87,89]: richer knowledge sources can be of great benefit to both knowledge-rich systems [89,106] and supervised classifiers [98,137]. Finally, wide-coverage structured lexical knowledge is expected to be beneficial for areas other than text processing, e.g., grounded applications such as Geographic Information Systems and situated robots [11].

Lexical knowledge is available in many different forms, ranging from unstructured terminologies (i.e., lists of terms), to glossaries (e.g., Web-derived domain glossaries [35]), thesauri (e.g., Roget's Thesaurus [116]), machine-readable dictionaries (e.g., LDOCE [110]) and full-fledged computational lexicons and ontologies, such as WordNet [77,37] and Cyc [63]. However, building such resources manually is an onerous task. It requires dozens of years and has to be repeated from scratch for each new language. Then to this has to be added the cost of interlinking the resources across languages and domains. Manual efforts of this kind include EuroWordNet [128], MultiWordNet [104], BalkaNet [127], and many others. However, resources

* Corresponding author.
E-mail addresses: navigli@di.uniroma1.it (R. Navigli), ponzetto@di.uniroma1.it (S.P. Ponzetto).

0004-3702/$ – see front matter © 2012 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.artint.2012.07.001

for non-English languages often have much poorer coverage. As a result, an obvious bias exists towards conducting research in resource-rich languages such as English.

Recently, the increasing availability of online collaborative resources has attracted the attention of many researchers in the Artificial Intelligence community [69].[1] Such resources contain semi-structured information, mainly in textual, possibly hyperlinked, form. Wikipedia[2] is a case in point here, being the largest and most popular collaborative and multilingual resource of world and linguistic knowledge. Much work in the literature has been devoted to the extraction of structured information from Wikipedia, including extracting lexical and semantic relations between concepts [117,123], factual information [135], and transforming the Web encyclopedia into a full-fledged semantic network [84,72,81]. One major feature of Wikipedia is its richness of explicit and implicit semantic knowledge, mostly about named entities (e.g., Apple as a company). However, its encyclopedic nature is also a major limit, in that it lacks full coverage for the lexicographic senses of a given lemma (e.g., the apple fruit and tree senses are merged into one single meaning). Such a lexical coverage, instead, can be provided by a highly-structured computational lexicon such as WordNet.

In this paper, we take a major step towards realizing the vision of a wide-coverage multilingual knowledge resource. We present a novel[3] integration and enrichment methodology that produces a very large multilingual semantic network: BabelNet. This resource is created by linking the largest multilingual Web encyclopedia – i.e., Wikipedia – to the most popular computational lexicon – i.e., WordNet. The integration is performed via an automatic mapping and by filling in lexical gaps in resource-poor languages with the aid of Machine Translation. The result is an "encyclopedic dictionary" that provides concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations. The contribution of this work is threefold:

1. We present a lightweight methodology to automatically map encyclopedic entries to a computational lexicon. At its core, the proposed approach estimates mapping probabilities using a variety of methods, including ones based on simple bag-of-words and more advanced graph representations. We show that our method is able to map tens of thousands of Wikipedia pages to their corresponding WordNet senses, with a performance near 78% F_1 measure.
2. Thanks to the automatic mapping of Wikipedia to WordNet we are able to build an integrated multilingual semantic network where millions of concepts are lexicalized in 6 different languages. We start by harvesting human-edited translations provided by Wikipedia with its inter-language links and, to fill translation gaps (i.e., missing translations for resource-poor languages), we apply a state-of-the-art statistical Machine Translation (MT) system to millions of sense-tagged sentences from Wikipedia and SemCor [79]. As a result we are able to cover a substantial part of existing wordnets, as well as to provide many novel lexicalizations.
3. We use the knowledge encoded in BabelNet to perform knowledge-rich, graph-based Word Sense Disambiguation in both a monolingual and multilingual setting. The results indicate that the explicit semantic relations from WordNet and the topical associative ones from Wikipedia can complement each other and enable us to achieve state-of-the-art performance when they are combined within a wide-coverage semantic network.

The remainder of this article is organized as follows: in Section 2 we introduce the two resources which are used to build BabelNet, i.e., WordNet and Wikipedia. Section 3 presents an overview of our resource and its construction methodology. We perform an intrinsic evaluation of BabelNet in Section 4, and provide statistics for its current version in Section 5. Section 6, instead, presents a set of extrinsic evaluations aimed at benchmarking the performance of BabelNet on monolingual and multilingual Word Sense Disambiguation using standard datasets from the SemEval competitions. We present related work in Section 7 and conclude with final remarks and future work directions in Section 8.

2. Knowledge resources

BabelNet aims at providing an "encyclopedic dictionary" by merging WordNet and Wikipedia. Accordingly in the following we provide a brief overview of these two resources.

2.1. WordNet

WordNet [77,37] is by far the most popular lexical knowledge resource in the field of NLP. It is a computational lexicon of the English language based on psycholinguistic principles. A concept in WordNet is represented as a synonym set (called *synset*), i.e., the set of words that share the same meaning. For instance, the concept of play as a dramatic work is expressed by the following synset[4]:

[1] See also the upcoming special issue of this journal on Artificial Intelligence, Wikipedia and semi-structured resources.

[2] http://www.wikipedia.org.

[3] This article builds upon and expands [91] and [106]. The resource described in this paper is freely available at http://lcl.uniroma1.it/babelnet under a Creative Commons Attribution-Noncommercial-Share Alike License.

[4] In the following we use WordNet version 3.0. Following [87] we denote with w_p^i the i-th sense of a word w with part of speech p (e.g., play_n^1 denotes the first nominal sense of play). We use word senses to unambiguously denote the corresponding synsets (e.g., play_n^1 for $\{\mathrm{play}_n^1, \mathrm{drama}_n^1, \mathrm{dramatic\ play}_n^1\}$). Hereafter, we use *word sense* and *synset* interchangeably.

$$\{\mathsf{play}^1_n, \mathsf{drama}^1_n, \mathsf{dramatic\ play}^1_n\},$$

where each word's subscript and superscript indicate its part of speech (e.g., n stands for noun) and sense number, respectively. Words can be polysemous and therefore the same word, e.g., play, can appear in more than one synset. For instance, WordNet represents the concept of dramatic play with the above synset and the concept of children's play activity with the following synset:

$$\{\mathsf{play}^8_n, \mathsf{child's\ play}^2_n\}.$$

For each synset, WordNet provides a textual definition, or *gloss*. For example, the gloss of the first synset of play_n is: "a dramatic work intended for performance by actors on a stage". Synsets can contain sample sentences to provide examples of their usage, e.g., "he wrote several plays but only one was produced on Broadway" for the dramatic work sense of play. Finally, synsets are related to each other by means of *lexical* and *semantic relations*. The inventory of semantic relations varies among parts of speech, e.g., it includes 12 relations for nouns. For instance, given two nominal synsets, typical semantic relations that can hold between them in WordNet include:

- *is-a* relations such as hypernymy (expressing concept generalization, e.g., play^1_n *is-a* $\mathsf{dramatic\ composition}^1_n$) and hyponymy (expressing concept specialization): the *is-a* relation is by far the most common in WordNet. It structures the concepts expressed by synsets into a lexicalized taxonomy where each concept inherits information from its superordinate concepts.
- *instance-of* relations denoting set membership between a named entity and the class it belongs to (for instance, $\mathsf{Shakespeare}^1_n$ is an instance of $\mathsf{dramatist}^1_n$).[5]
- *part-of* relations expressing the elements of a partition by means of meronymy (e.g., a $\mathsf{stage\ direction}^1_n$ is a meronym of play^1_n) and holonymy (e.g., a play^1_n is a holonym of $\mathsf{stage\ direction}^1_n$).

In addition to the standard WordNet relations, in this paper we also consider *gloss* relations. Given a synset S and its set of disambiguated gloss words $gloss(S) = \{\mathsf{s}_1, \ldots, \mathsf{s}_k\}$,[6] we define a semantic gloss relation between S and each synset S_i containing a sense $s_i \in gloss(S)$, $i = 1, \ldots, k$. For instance, the disambiguated gloss for play^1_n contains, among others, senses like actor^1_n and stage^3_n, so S – i.e., play^1_n – is related to both of the latter synsets via the gloss relation.

2.2. Wikipedia

Our second resource, Wikipedia, is a multilingual Web-based encyclopedia. It is a collaborative open source medium edited by volunteers to provide a very large wide-coverage repository of encyclopedic knowledge. Each article in Wikipedia is represented as a page (henceforth, Wikipage) and presents information about a specific concept (e.g., PLAY (THEATRE)) or named entity (e.g., WILLIAM SHAKESPEARE).[7] The title of a Wikipage (e.g., PLAY (THEATRE)) is composed of the lemma of the concept defined (e.g., play) plus an optional label in parentheses which specifies its meaning if the lemma is ambiguous (e.g., theatre vs. activity).[8]

The text in Wikipedia is partially structured. Apart from Wikipages having tables and infoboxes (a special kind of table which summarizes the most important attributes of the entity referred to by a page, such as the birth date and biographical details of a playwright like WILLIAM SHAKESPEARE), various relations exist between the pages themselves. These include:

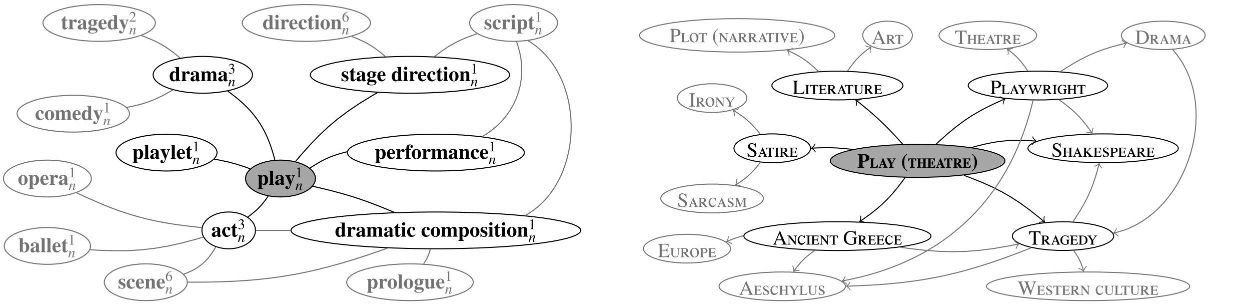
- **Redirect pages:** These pages are used to forward to the Wikipage containing the actual information about a concept of interest. This is used to point alternative expressions for a concept to the same entry, and thus models *synonymy*. For instance, STAGEPLAY and THEATRICAL PLAY both redirect to PLAY (THEATRE).
- **Disambiguation pages:** These pages collect links for a number of possible concepts an arbitrary expression could be referred to. This models *homonymy* and *polysemy*, e.g., PLAY links to both pages PLAY (THEATRE) and PLAY (ACTIVITY).
- **Internal links:** Wikipages typically contain hypertext linked to other Wikipages, which refers to related concepts. For instance, PLAY (THEATRE) links to LITERATURE, PLAYWRIGHT, DIALOGUE, etc., whereas PLAY (ACTIVITY) points to SOCIALIZATION, GAME, RECREATION, and so on.
- **Inter-language links:** Wikipages also provide links to their counterparts (i.e., corresponding concepts) contained within wikipedias in other languages (e.g., the English Wikipage PLAY (THEATRE) links to the Italian DRAMMA and German BÜHNENWERK).
- **Categories:** Wikipages can be assigned to one or more categories, i.e., special pages used to encode topics, e.g., PLAY (THEATRE) is categorized under THEATRE, DRAMA, LITERATURE, etc.

[5] This is a specific form of *is-a* introduced in WordNet 2.1 [78].

[6] Sense disambiguated glosses are distributed by the Princeton WordNet project at http://wordnet.princeton.edu/glosstag.shtml.

[7] Throughout the paper, unless otherwise stated, we use the general term *concept* to denote either a concept or a named entity.

[8] We use the English Wikipedia database dump from November 3, 2009, which includes 3,083,466 Wikipages. Throughout this paper, we use Sans Serif for words, SMALL CAPS for Wikipedia pages and CAPITALS for Wikipedia categories.



(a) Excerpt of the WordNet graph centered on the synset play^1_n.

(b) Excerpt of the Wikipedia graph centered on the Wikipage Play (theatre).

Fig. 1. Excerpts of the WordNet (a) and Wikipedia graphs (b). Both resources can be viewed as graphs by taking synsets (Wikipages, respectively) as nodes and lexical and semantic relations between synsets (hyperlinks between pages) as edges.

Both WordNet and Wikipedia can be viewed as graphs. In the case of WordNet, nodes are synsets and edges lexical and semantic relations between synsets[9] whereas, in the case of Wikipedia, nodes are Wikipages and edges the hyperlinks between them (i.e., the above-mentioned *internal* links). An excerpt of the WordNet and Wikipedia graphs centered on the synset play^1_n and Wikipage Play (theatre) is given in Fig. 1(a) and (b), respectively.[10] The two graphs highlight the *complementarity* of these two resources: while there are corresponding nodes in the two graphs (e.g., $\mathsf{tragedy}^2_n$ and Tragedy), each resource also contains knowledge which is missing in the other: this includes missing concepts (for instance, no Wikipage corresponding to $\mathsf{direction}^6_n$), named entities (such as Ancient Greece missing in WordNet), as well as relations (e.g., the topical relation between Shakespeare and Tragedy).

3. BabelNet

BabelNet encodes knowledge as a labeled directed graph $G = (V, E)$ where V is the set of *nodes* – i.e., *concepts* such as play and *named entities* such as Shakespeare – and $E \subseteq V \times R \times V$ is the set of *edges* connecting pairs of concepts (e.g., play *is-a* dramatic composition). Each edge is labeled with a *semantic relation* from R, i.e., $\{is\text{-}a, part\text{-}of, \ldots, \epsilon\}$, where ϵ denotes an unspecified semantic relation. Importantly, each node $v \in V$ contains a set of lexicalizations of the concept for different languages, e.g., $\{\mathsf{play}_{\mathrm{EN}}$, $\mathsf{Theaterstück}_{\mathrm{DE}}$, $\mathsf{dramma}_{\mathrm{IT}}$, $\mathsf{obra}_{\mathrm{ES}}$, ..., $\mathsf{pièce\ de\ théâtre}_{\mathrm{FR}}\}$. We call such multilingually lexicalized concepts *Babel synsets*. Concepts and relations in BabelNet are harvested from the largest available semantic lexicon of English, WordNet, and a wide-coverage collaboratively-edited encyclopedia, Wikipedia (introduced in Section 2). In order to build the BabelNet graph, we collect at different stages:

a. From WordNet, all available word senses (as *concepts*) and all the lexical and semantic pointers between synsets (as *relations*);
b. From Wikipedia, all encyclopedic entries (i.e., Wikipages, as *concepts*) and semantically unspecified *relations* from hyperlinked text.

An overview of BabelNet is given in Fig. 2. The excerpt highlights that WordNet and Wikipedia can overlap both in terms of concepts and relations: accordingly, in order to provide a *unified resource*, we merge the intersection of these two knowledge sources. Next, to enable multilinguality, we collect the lexical realizations of the available concepts in different languages. Finally, we connect the multilingual Babel synsets by establishing semantic relations between them. Thus, our methodology consists of three main steps:

1. We **combine WordNet and Wikipedia** by automatically acquiring a mapping between WordNet senses and Wikipages (Section 3.1). This avoids duplicate concepts and allows their inventories of concepts to complement each other.
2. We **harvest multilingual lexicalizations** of the available concepts (i.e., Babel synsets) by using (a) the human-generated translations provided by Wikipedia (the so-called *inter-language* links), as well as (b) a machine translation system to translate occurrences of the concepts within sense-tagged corpora (Section 3.2).

[9] Lexical relations link senses (e.g., dental^1_a pertains-to tooth^1_n). However, relations between senses can easily be extended to the synsets which contain them, thus making all the relations connect synsets.

[10] We represent the WordNet fragment as an unlabeled undirected graph for the sake of compactness. Note that for our purposes this has no impact, since our graph-based disambiguation methods do not distinguish between different kinds of relations in the lexical knowledge base and WordNet relations such as hypernymy and hyponymy are paired so as to be symmetric.

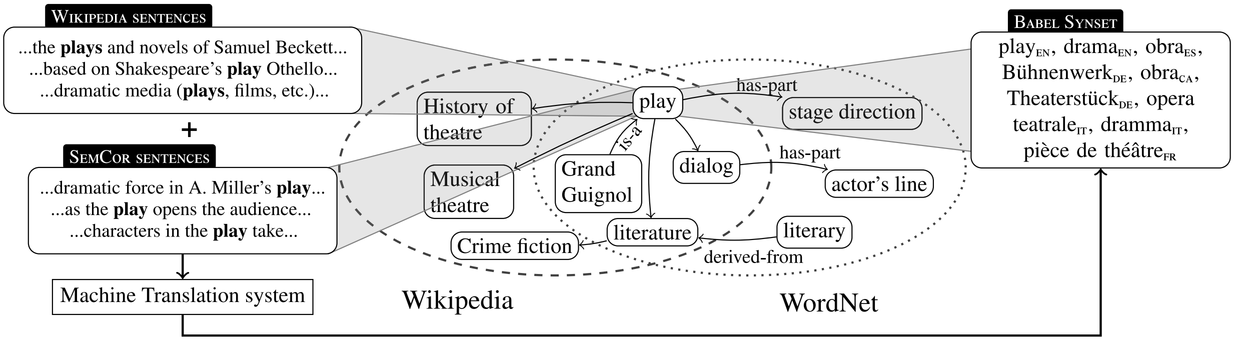

Wikipedia sentences
...the **plays** and novels of Samuel Beckett...
...based on Shakespeare's **play** Othello...
...dramatic media (**plays**, films, etc.)...
+
SemCor sentences
...dramatic force in A. Miller's **play**...
...as the **play** opens the audience...
...characters in the **play** take...
Machine Translation system
History of theatre
Musical theatre
Crime fiction
play
is-a
Grand Guignol
literature
dialog
has-part
stage direction
has-part
actor's line
literary
derived-from
Wikipedia
WordNet
Babel Synset
play_EN, drama_EN, obra_ES, Bühnenwerk_DE, obra_CA, Theaterstück_DE, opera teatrale_IT, dramma_IT, pièce de théâtre_FR

Fig. 2. An illustrative overview of BabelNet (we label nodes with English lexicalizations only): unlabeled edges are obtained from links in the Wikipages (e.g., Play (theatre) links to Musical theatre), whereas labeled ones from WordNet (e.g., play^1_n *has-part* $\text{stage direction}^1_n$).

3. We **establish relations between Babel synsets** by collecting all relations found in WordNet, as well as all wikipedias in the languages of interest (Section 3.3): in order to encode the strength of association between synsets, we compute their degree of correlation using a measure of relatedness based on the Dice coefficient.

Throughout the section, we illustrate our approach by means of an example focused on the Wikipage Play (theatre) and the WordNet senses of play.

3.1. Mapping Wikipedia to WordNet

The first phase of our methodology aims at establishing links between Wikipages and WordNet senses. Formally, given the entire set of pages $Senses_{\text{Wiki}}$ and WordNet senses $Senses_{\text{WN}}$, we acquire a mapping:

$$\mu : Senses_{\text{Wiki}} \to Senses_{\text{WN}} \cup \{\epsilon\},$$

such that, for each Wikipage $w \in Senses_{\text{Wiki}}$, we have:

$$\mu(w) = \begin{cases} s \in Senses_{\text{WN}}(w) & \text{if a link can be established,} \\ \epsilon & \text{otherwise,} \end{cases}$$

where $Senses_{\text{WN}}(w)$ is the set of senses of the *lemma* of w in WordNet. Given a Wikipage w, its corresponding lemma is given by either its title (tragedy for Tragedy) or the main token of a sense-labeled title (play for Play (theatre)). For example, if our mapping methodology linked Play (theatre) to the corresponding WordNet sense play^1_n, we would have $\mu(\text{Play (theatre)}) = \text{play}^1_n$. Our method works as follows:

1. We first develop a *mapping algorithm* (Section 3.1.1) that:
   (a) leverages resource-specific properties of our source and target resources, namely monosemous senses and redirections;
   (b) given a Wikipage, finds the WordNet sense that maximizes the probability of the sense providing an adequate corresponding concept for the page.
2. We then view resource mapping as a disambiguation problem, and associate a so-called *disambiguation context* with both WordNet senses and Wikipages (Section 3.1.2).
3. Finally, we define two strategies to *estimate the conditional probability* of a WordNet sense given a Wikipage, both based on disambiguation contexts (Section 3.1.3). These estimators either
   (a) make use of a simple bag-of-words (BoW) approach, or
   (b) leverage the graph structure of the target resource, i.e., WordNet's in our case.

3.1.1. Mapping algorithm

In order to link each Wikipage to a WordNet sense, we make use of the mapping algorithm whose pseudocode is presented in Algorithm 1. The following steps are performed:

- Initially (lines 1–2), our mapping μ is empty, i.e., it links each Wikipage w to ϵ.
- For each Wikipage w whose lemma is monosemous both in Wikipedia and WordNet (i.e., $|Senses_{\text{Wiki}}(w)| = |Senses_{\text{WN}}(w)| = 1$) we map w to its only WordNet sense w^1_n (lines 3–5).
- Finally, for each remaining Wikipage w for which no mapping was previously found (i.e., $\mu(w) = \epsilon$, line 7), we do the following:

Algorithm 1 The mapping algorithm.

Input: $Senses_{\text{Wiki}}$, $Senses_{\text{WN}}$
Output: a mapping $\mu : Senses_{\text{Wiki}} \to Senses_{\text{WN}} \cup \{\epsilon\}$

```
1:  for each $w \in Senses_{\text{Wiki}}$
2:      $\mu(w) := \epsilon$
3:  for each $w \in Senses_{\text{Wiki}}$
4:      if $|Senses_{\text{Wiki}}(w)| = |Senses_{\text{WN}}(w)| = 1$ then
5:          $\mu(w) := w_n^1$
6:  for each $w \in Senses_{\text{Wiki}}$
7:      if $\mu(w) = \epsilon$ then
8:          for each $d \in Senses_{\text{Wiki}}$ s.t. $d$ redirects to $w$
9:              if $\mu(d) \neq \epsilon$ and $\mu(d)$ is in a synset of $w$ then
10:                 $\mu(w) :=$ sense of $w$ in synset of $\mu(d)$; break
11: for each $w \in Senses_{\text{Wiki}}$
12:     if $\mu(w) = \epsilon$ then
13:         if no tie occurs then
14:             $\mu(w) := \operatorname*{argmax}_{s \in Senses_{\text{WN}}(w)} p(s|w)$
15: return $\mu$
```

- lines 8–10: for each Wikipage d which is a redirection to w, for which a mapping was previously found (i.e., $\mu(d) \neq \epsilon$, that is, d is monosemous in both Wikipedia and WordNet) and such that it maps to a sense $\mu(d)$ in a synset S that also contains a sense of w, we map w to the corresponding sense in S;
- lines 11–14: if a Wikipage w has not been linked yet, we assign the most likely sense to w based on the maximization of the conditional probabilities $p(s|w)$ over the senses $s \in Senses_{\text{WN}}(w)$ (no mapping is established if a tie occurs, line 13).

As a result of the execution of the algorithm, the mapping μ is returned (line 15). The gist of the mapping algorithm is the calculation of the conditional probability $p(s|w)$ of selecting the WordNet sense s given the Wikipage w. The sense s which maximizes this probability is obtained as follows:

$$\begin{aligned}\mu(w) = \operatorname*{argmax}_{s \in Senses_{\text{WN}}(w)} p(s|w) &= \operatorname*{argmax}_{s} \frac{p(s, w)}{p(w)} \\ &= \operatorname*{argmax}_{s} p(s, w). \end{aligned} \tag{1}$$

This last formula is obtained by observing that $p(w)$ does not influence our maximization, as it is a constant independent of s. As a result, the most appropriate sense s is determined by maximizing the joint probability $p(s, w)$ of sense s and page w.

3.1.2. Disambiguation contexts

In order to estimate the joint probability of a WordNet sense and Wikipage, we start with the same technique as that adopted in Word Sense Disambiguation [87], and define a *disambiguation context* for each of the two concepts. Basically, given a concept, i.e., a page or sense, this disambiguation context is a set of words obtained from the corresponding resource, whose senses are associated with the input concept by means of some semantic relation and which provide evidence for a potential link in our mapping μ.

3.1.2.1. Disambiguation context of a Wikipage. Given a Wikipage w, we use the following information as disambiguation context:

- **Sense labels:** e.g., given the page Play (theatre), the word theatre is added to the disambiguation context.
- **Links:** the titles' lemmas of the pages linked from the Wikipage w (i.e., outgoing links). For instance, the links in the Wikipage Play (theatre) include literature, comedy, etc.
- **Redirections:** the titles' lemmas of the pages which are redirecting to w, e.g., Playlet redirects to Play (theatre), so we include playlet in the context.
- **Categories:** Wikipages are typically classified according to one or more categories. For example, the Wikipage Play (theatre) is categorized as PLAYS, DRAMA, THEATRE, etc. While many categories are very specific and do not appear in WordNet (e.g., THEATRE CHARACTERS), we consider their syntactic heads [108] for inclusion in the disambiguation context (i.e., character).

The disambiguation context $Ctx(w)$ of a Wikipage w is then defined as the set of words obtained from all of the four sources above. For example, Ctx(Play (theatre)) = {theatre, literature, comedy, ..., playlet, drama, ..., character}.

3.1.2.2. Disambiguation context of a WordNet sense. Given a WordNet sense s and its synset S, we instead use the following sources as disambiguation context:

- **Synonymy:** all synonyms of s in synset S. For instance, given the synset of play_n^1, all its synonyms are included in the context (that is, drama and dramatic play).
- **Hypernymy/Hyponymy:** all synonyms in the synsets H such that H is either a hypernym (i.e., a generalization) or a hyponym (i.e., a specialization) of S. For example, given play_n^1, we include its hypernym dramatic composition.
- **Gloss:** the set of lemmas of the content words occurring within the gloss of s. For instance, given $s = \text{play}_n^1$, defined as "a dramatic work intended for performance by actors on a stage", we add to the disambiguation context of s the following lemmas: work, dramatic work, intend, performance, actor, stage.

Given a WordNet sense s, we define its disambiguation context $Ctx(s)$ as the set of words obtained from some or all of the four sources above. For example, $Ctx(\text{play}_n^1) = \{$drama, dramatic play, composition, work, intend, ..., actor, stage$\}$.

3.1.3. Probability estimation

Given the disambiguation contexts, we can compute the probability of a WordNet sense and Wikipage referring to the same concept, i.e., the joint probability defined in Eq. (1). We estimate $p(s, w)$ as:

$$p(s,w) = \frac{score(s,w)}{\sum_{\substack{s' \in Senses_{\text{WN}}(w),\\ w' \in Senses_{\text{Wiki}}(w)}} score(s',w')}. \tag{2}$$

We define two different ways of computing the $score(s, w)$ function:

- **Bag-of-words method:** computes $score(s,w) = |Ctx(s) \cap Ctx(w)| + 1$ (we add 1 as a smoothing factor). This is a simple method already proposed in [91], that determines the best sense s by computing the intersection of the disambiguation contexts of s and w, and thus does not exploit the structural information available in WordNet or Wikipedia.
- **Graph-based method:** starts with the flat disambiguation context of the Wikipage $Ctx(w)$ and transforms it into the structured representation of a graph, which is then used to score the different senses of w in WordNet. A labeled directed graph $G = (V, E)$ is built following the same procedure outlined in [89] which connects possible senses of w's lemma with the senses of the words found in $Ctx(w)$. Specifically:
  1. We first define the set of nodes of G to be made up of all WordNet senses for the lemma of Wikipage w and for the words in $Ctx(w)$, i.e., $V := Senses_{\text{WN}}(w) \cup \bigcup_{cw \in Ctx(w)} Senses_{\text{WN}}(cw)$. Initially, the set of edges of G is empty, i.e., $E := \emptyset$.
  2. Next, we connect the nodes in V based on the paths found between them in WordNet. Formally, for each vertex $v \in V$, we perform a depth-first search along the WordNet graph and every time we find a node $v' \in V$ $(v \neq v')$ along a simple path $v, v_1, \ldots, v_k, v'$ of maximal length L, we add all intermediate nodes and edges of such a path to G, i.e., $V := V \cup \{v_1, \ldots, v_k\}$, $E := E \cup \{(v, v_1), \ldots, (v_k, v')\}$.

  The result of this procedure is a subgraph of WordNet containing (1) the senses of the words in context, (2) all edges and intermediate senses found in WordNet along all paths of maximal length L that connect them. To compute $score(s, w)$ given a disambiguation graph G, we define a scoring function of the paths starting from s and ending in any of the senses of the context words $Ctx(w)$:

  $$score(s,w) = \sum_{cw \in Ctx(w)} \sum_{s' \in Senses_{\text{WN}}(cw)} \sum_{p \in paths_{\text{WN}}(s,s')} e^{-(length(p)-1)} \tag{3}$$

  where $paths_{\text{WN}}(s, s')$ is the set of all paths between s and s' in WordNet, and $length(p)$ the length of path p in terms of its number of edges.

We illustrate the execution of our mapping algorithm by way of an example. Let us focus on the Wikipage PLAY (THEATRE). The word is polysemous both in Wikipedia and WordNet, thus lines 3–5 of the algorithm do not concern this Wikipage. In the main part of our algorithm (lines 6–14) we aim to find a mapping μ(PLAY (THEATRE)) to an appropriate WordNet sense of the word play. To this end, we first check whether a redirection exists to PLAY (THEATRE) that was previously disambiguated (lines 8–10). Next, we construct the disambiguation context for the Wikipage by including words from its label, links, redirections and categories (cf. Section 3.1.2). The context thus includes, among others, the following lemmas: drama, comedy, performing art, literature, tragedy and performance (cf. also Fig. 1). We now construct the disambiguation contexts for two of the WordNet senses of play, namely the 'theatre' (#1) and the 'activity' (#8) senses. To do so, we include lemmas from their synsets, hypernyms, hyponyms, and glosses. The context for play_n^1 includes: performance, drama, act, playlet. The context for play_n^8 contains among others: house, doctor, fireman, diversion and imagination. When mapping using the *bag-of-words method* we simply compute the size of the intersection between the disambiguation context of the Wikipage and each of the WordNet senses of interest: the sense with the largest intersection is #1, so the following mapping is established: μ(PLAY (THEATRE))= play_n^1. In the case of the *graph-based method*, instead, we construct a disambiguation graph

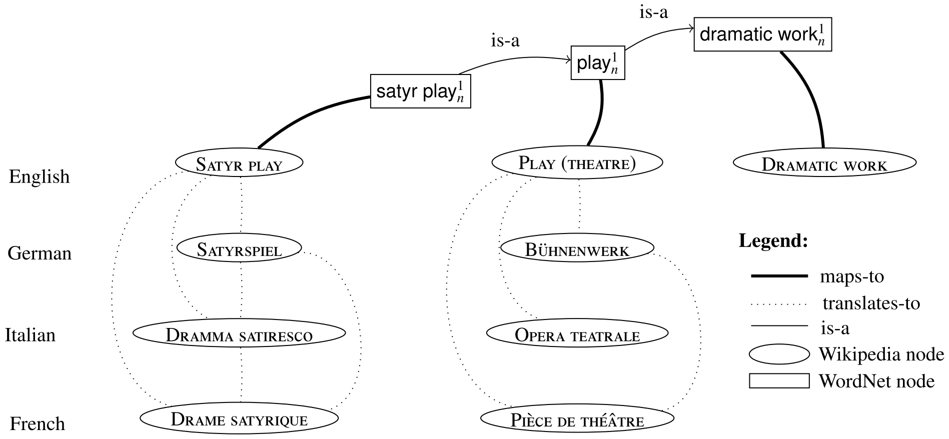


Fig. 3. Translating Babel synsets based on Wikipedia inter-language links. After mapping (bold links in the figure), Babel synsets integrating WordNet synsets and Wikipages are straightforwardly translated by collecting the manual translations provided by editors as hyperlinks to wikipedias in languages other than English (dotted links in the figure). Note that, in this specific example, WordNet and Wikipedia are highly complementary to each other, since WordNet provides relations missing from Wikipedia, while Wikipedia adds the multilingual dimension.

for the Wikipage context and select the highest scoring WordNet sense based on Eq. 3. In our example, the result is the same as the bag-of-words method: however, note that by structuring the Wikipage context using WordNet relations we are now able to capture broader semantic relations between context terms such as the ones contained in the following paths:

$\text{play}_n^1 - \text{drama}_n^3 - \text{tragedy}_n^2$,
$\text{play}_n^1 - \text{drama}_n^3 - \text{comedy}_n^1$,
$\text{play}_n^1 - \text{dramatist}_n^1 - \text{Shakespeare}_n^1$,
$\text{play}_n^1 - \text{dramatist}_n^1 - \text{Eliot}_n^1 - \text{literature}_n^1$,
$\text{play}_n^1 - \text{act}_v^{10} - \text{roleplaying}_n^1 - \text{acting}_n^1 - \text{character}_n^4$.

As a result, the graph-based method is expected to establish mappings even in cases when the intersection of the WordNet and Wikipedia disambiguation context is empty, thus having more recall. As an example, let us consider the Wikipage Play (activity), which refers to the 'playing games' sense of play. Here, the disambiguation context for the Wikipage contains, among others, words such as game, toy, childhood and recreation. But while the BoW method is not able to induce any mapping, since the intersection with the disambiguation context of any of play_n^1 and play_n^8 is empty, using the graph-based method we are able to output the mapping $\mu(\text{Play (activity)}) = \text{play}_n^8$ by finding paths such as:

$\text{play}_n^8 - \text{diversion}_n^1 - \text{game}_n^3$,
$\text{play}_n^8 - \text{diversion}_n^1 - \text{activity}_n^1$,
$\text{play}_n^8 - \text{child}_n^1 - \text{childhood}_n^2$.

3.2. Translating Babel synsets

So far we have linked English Wikipages to WordNet senses. Given a Wikipage w, and provided it is mapped to a sense s (i.e., $\mu(w) = s$), we create a Babel synset $S \cup W$, where S is the WordNet synset to which sense s belongs, and W includes: (i) w; (ii) the set of redirections to w; (iii) all the pages linked via its inter-language links (that is, translations of the Wikipage into other languages); (iv) the redirections to the inter-language links found in the Wikipedia of the target language. For instance, given that $\mu(\text{Play (theatre)}) = \text{play}_n^1$, the corresponding Babel synset is $\{\text{play}_{\text{EN}}$, $\text{Bühnenwerk}_{\text{DE}}$, $\text{pièce de théâtre}_{\text{FR}}$, … , $\text{opera teatrale}_{\text{IT}}\}$ (Fig. 3). Including redirections additionally enlarges the Babel synset with $\{\text{Theaterstück}_{\text{DE}}$, $\text{texte dramatique}_{\text{FR}}\}$. However, two issues arise: first, a concept might be covered only in one of the two resources (either WordNet or Wikipedia), meaning that no link can be established (e.g., Musical theatre or actor's line_n^1 in Fig. 2); second, even if covered in both resources, the Wikipage for the concept might not provide any translation for the language of interest (e.g., the Spanish and Catalan inter-language links for Play (theatre) are missing in Wikipedia).

In order to address the above issues, and thus guarantee high coverage for all languages, we developed a methodology for translating senses in the Babel synset into missing languages. Given a WordNet word sense in our Babel synset of interest (e.g., play_n^1) we collect its occurrences in SemCor [79], a corpus of more than 200,000 words annotated with WordNet senses. We do the same for Wikipages by retrieving sentences in Wikipedia with links to the Wikipage of interest (e.g., Play (theatre)). By repeating this step for each English lexicalization in a Babel synset, we obtain a collection of sentences

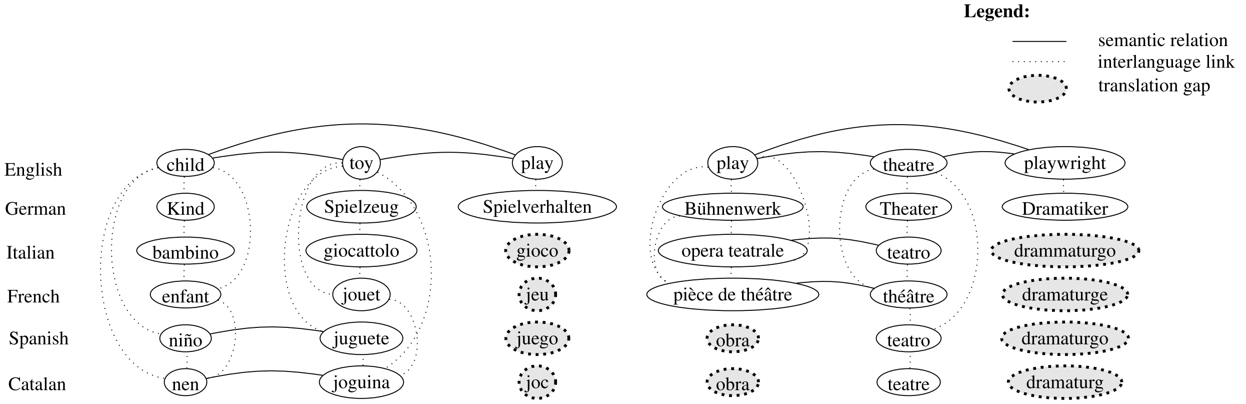


Fig. 4. Translating Babel synsets based on a Machine Translation system. In order to fill lexical gaps (i.e., missing translations, typically for resource-poor languages), sense-annotated data are collected from SemCor and Wikipedia, and their most frequent translations are included as additional lexicalizations in the network.

for the synset (see left part of Fig. 2). Next, we apply a state-of-the-art Machine Translation system[11] and translate the set of sentences into all languages of interest. Given a specific term in the initial Babel synset, we collect the set of its translations. We then identify the most frequent translation in each language and add it to the Babel synset (in the case of ties, we collect all top-scored translations). Note that translations are sense-specific, as the context in which a sense-annotated term occurs is provided to the translation system. For instance, in order to collect missing translations for PLAY (THEATRE) and its corresponding WordNet sense play_n^1, we collect from Wikipedia occurrences of hyperlinks to the Wikipage and translate sentences such as the following:

(a) Best known for his **[[Play (theatre)|play]]** Ubu Roi, which is often cited as a forerunner to the surrealist theatre of the 1920s and 1930s, Jarry wrote in a variety of genres and styles.

Similarly, from SemCor we collect and automatically translate, among others, the following sentence:

(b) The situation in which we find ourselves is brought out with dramatic force in Arthur Miller's $\textbf{play}_n^1$ The Crucible, which deals with the Salem witch trials.

As a result, we can enrich the initial Babel synset with the following words: drame_{FR}, $\text{dramma}_{\text{IT}}$, obra_{CA}, obra_{ES} (Fig. 4, additional lexicalizations are shown in dotted ellipses). Note that not only do we obtain translations for Catalan and Spanish which were unavailable in the first phase (e.g., because no inter-language link was provided), but we also obtain more lexicalizations for the French and Italian languages (for which, instead, inter-language links were available). To ensure precision, we translate only WordNet and Wikipedia senses that occur in at least 3 different sentences. To harvest all translations of interest in a reasonable amount of time,[12] we collect at most 10 sentences for each sense. Moreover, in the case of Wikipedia, we translate only titles of Wikipages which do not refer to named entities. To do this, we first identify the named entities based on a simple heuristic: we assume that Wikipage titles which contain at least two tokens starting with an uppercase letter (e.g., WILLIAM SHAKESPEARE) are proper names that refer to named entities. While more complex heuristics could be used (e.g., the one proposed in [16]), we found this very simple approach worked surprisingly well – i.e., achieving an accuracy of 94% on a validation sample of 100 pages. As a result of this procedure we are able to reduce the number of Wikipedia concepts to translate from over 3 million to 324,137. For these named entities we do not collect any sentence and assume that they remain the same across languages (that is, given that BabelNet currently contains European languages only, we do not address any issue related to proper name transliteration). Similarly, we perform a contextless translation for those words in WordNet which are monosemous: in this case, in fact, we simply include in BabelNet the translations returned by Google Translate.

As a result of our translation procedure we produce a very large sense-labeled corpus, that we call BabelCor, and which is freely available together with BabelNet (see Section 5 for details).

[11] We use the Google Translate API (http://research.google.com/university/translate). An initial prototype used a statistical machine translation system based on Moses [57] and trained on Europarl [56]. However, we found this system unable to cope with many technical terms in the domains of sciences, literature, history, etc.

[12] Google Translate requires us to keep the traffic below 1 query per second (at most 86,400 translation queries per day).

3.3. Harvesting semantic relations

The final step of our methodology consists of establishing semantic relations between our multilingual Babel synsets. This is achieved by: (i) collecting the relations directly from the two knowledge sources which are used to build BabelNet, namely WordNet and Wikipedia; (ii) weighting them using a relatedness measure based on the Dice coefficient. We first collect all lexical and semantic relations from WordNet (including the gloss relations introduced in Section 2.1). For instance, given the Babel synset for play_n^1, we connect it to the Babel synsets of $\mathsf{playlet}_n^1$, act_n^3, etc. (cf. Fig. 1(a)). We then include all relations from Wikipedia, making use of its internal hyperlink structure: for each Wikipage, we collect all links occurring within it and establish an unspecified semantic relation ϵ between their corresponding Babel synsets (cf. the semantic relations for Play (theatre) in Fig. 1(b)). To harvest as many relevant relations as possible, we make use of *all* wikipedias in the available languages: that is, relations from wikipedias in languages other than English are also included. For instance, while the page Play (theatre) does not link directly to a highly related concept such as Acting, by pivoting on German (based on the interlanguage links) we find that Bühnenwerk links to Schauspiel, so a link can be established between the two Babel synsets that contain these English and German senses.

Edges in BabelNet are weighted to quantify the strength of association between Babel synsets. We use different strategies to leverage WordNet's and Wikipedia's distinctive properties – i.e., the availability of high-quality definitions from WordNet, and large amounts of hyperlinked text from Wikipedia – both based on the Dice coefficient. Given a semantic relation between two WordNet synsets s and s', we compute its corresponding weight using a method similar to the Extended Gloss Overlap measure for computing semantic relatedness [8]. We start by collecting (a) synonyms and (b) all gloss words from s and s', as well as their directly linked synsets, into two bags of words S and S'. We remove stopwords and lemmatize the remaining words. We then compute the degree of association between the two synsets by computing the Dice coefficient as the number of words the two bags have in common normalized by the total number of words in the bags: $\frac{2\times|S\cap S'|}{|S|+|S'|}$. For instance, given the following bags for play_n^1 and act_n^3:

play$_n^1$ {drama, **dramatic play**, **work**, **performance**, **dramatic work**, genre, **dramatic composition**, **television**, **actor**, **stage**, act, **subdivision**, **opera**, **ballet**, dramatic, **perform**, theater, morality play, **movie**, allegorical, satyr play, chorus, burlesque, role, stage direction, horrific, nature, macabre, playwright, dramatist, playlet};

act$_n^3$ {**subdivision**, play, **opera**, **ballet**, concert dance, music, story, representation, theatrical, **perform**, trained, dancer, overture, sing, interlude, **dramatic play**, **work**, **performance**, **dramatic work**, **actor**, intend, **stage**, scene, **dramatic composition**, **television**, **movie**},

the Dice coefficient gives $\frac{2\times 13}{31+26}=0.46$ as strength of correlation (the two bags contain 31 and 26 terms, respectively, and have 13 terms in common). In the case of edges corresponding to semantic relations between Wikipedia pages, instead, we quantify the degree of correlation between the two pages by using a co-occurrence based method, previously used for large-scale thesaurus extraction [54,138], which draws on large amounts of hyperlinked text.[13] Given two Wikipages w and w', we compute the frequency of occurrence of each individual page (f_w and $f_{w'}$) as the number of hyperlinks found in Wikipedia which point to it, and the co-occurrence frequency of w and w' ($f_{w,w'}$) as the number of times these links occur together within a context (i.e., a sliding window of 40 words in our case). The strength of association between w and w' is then given by applying the Dice coefficient formula to these frequency counts, namely: $\frac{2\times f_{w,w'}}{f_w+f_{w'}}$. For example, given the Wikipages Play (theatre) and Satire, we find in Wikipedia that they occur as a link 1,560 and 2,568 times, respectively, and co-occur 9 times within the same context. As a result, the Dice coefficient for these two pages is 0.0044.

4. In vitro evaluation

We perform two *in vitro* evaluations to assess the quality of BabelNet, namely: an evaluation of the mapping between Wikipedia and WordNet (Section 4.1) and an evaluation of the translations of Babel synsets (Section 4.2).

4.1. Mapping evaluation

In this section we describe our evaluation of the quality of mapping from Wikipedia pages to WordNet senses (Section 4.1.1). To corroborate our results and show the generality of our mapping method, we report further experiments on linking Wikipedia categories to WordNet (Section 4.1.2).

4.1.1. Mapping Wikipedia pages to WordNet

4.1.1.1. Experimental setting. To perform an experimental evaluation of the quality of our mappings from Wikipages to WordNet senses, we created a gold standard consisting of manually labeled ground-truth mappings. The gold standard is created

[13] During prototyping we tried to compute the correlation between WordNet synsets in a similar way by using sense labeled data from the SemCor corpus [79]. However, this produced a low-quality output with most of the synset pairs having a null Dice score, due to sparse counts resulting from SemCor's small size.

Table 1
Performance on mapping Wikipedia pages to WordNet synsets. Underlined results are those using the best value for the maximum search depth, found by maximizing the F_1 measure on the dataset from [99] used as development data. The best results for each metric are in bold: the best overall results (in terms of balanced F_1-measure and accuracy) are obtained by building disambiguation graphs using all WordNet relations and limiting the maximum depth of the depth-first search to 2. Using more relations yields improvements in recall, but also, as a trade-off, decreases in precision.

Mapping method			P	R	F_1	A
BoW		taxonomic	**89.7**	47.8	62.3	72.6
		gloss	87.6	51.8	65.1	74.0
		taxonomic + gloss	87.5	65.6	75.0	80.9
Graph	taxonomic relations					
	max depth	@ 2	87.2	60.8	71.6	77.9
		@ 3	81.6	65.0	72.4	78.7
		@ 4	78.3	69.5	73.6	79.4
	gloss relations					
	max depth	@ 2	80.5	60.6	69.1	77.0
		@ 3	77.5	65.2	70.9	78.2
		@ 4	72.4	67.1	69.6	78.0
	taxonomic + gloss relations					
	max depth	@ 2	81.2	74.6	**77.7**	**82.7**
		@ 3	72.8	**77.4**	75.1	80.1
		@ 4	64.3	76.2	69.8	75.0
		MFS baseline	25.4	49.2	33.5	25.4
		Random baseline	24.2	46.9	31.9	24.2

from a dataset which includes all lemmas whose senses are contained both in WordNet and Wikipedia. The dataset contains 80,295 lemmas, which correspond to 105,797 WordNet senses and 199,735 Wikipedia pages. The average polysemy is 1.3 and 2.5 for WordNet senses and Wikipages, respectively (2.9 and 4.7 when excluding monosemous words). From this dataset, we selected a random sample of 1,000 Wikipages and asked an annotator with previous experience in lexicographic annotation to provide the correct WordNet sense for each page (an empty sense label was given, if no correct mapping was possible). The gold-standard includes 505 non-empty mappings, i.e., Wikipages with a corresponding WordNet sense. In order to quantify the quality of the annotations and the difficulty of the task, a second annotator sense tagged a subset of 200 pages from the original sample. Our annotators achieved a κ inter annotator agreement [17] of 0.9, which indicates almost perfect agreement and is comparable with similar annotation efforts [99].

We evaluate our mapping methodology (cf. Section 3.1) using the BoW and the graph-based methods to estimate mapping probabilities (Section 3.1.3). Prior to applying our mapping algorithm in any of the two settings, we remove from the WordNet-Wikipedia intersection those Wikipages whose sense label is among the 100 most frequent ones (the value is experimentally set using the dataset from [99] as held-out development data), which helps us avoid mapping WordNet senses to pages belonging to domains which are typically found in Wikipedia only (bands, movies, etc.). We explore different disambiguation contexts for the WordNet senses (cf. Section 3.1.2): these include contexts based on synonymy, hypernymy and hyponymy (i.e., a taxonomic setting), glosses, and their union. For the disambiguation context of a Wikipage, instead, we use all the information that is available, i.e., sense labels, links and categories (cf. Section 3.1.2).[14] Additionally, for the graph-based method, we vary the maximum depth of the depth-first search to test the effect of exploring increasingly bigger portions of WordNet when building the disambiguation graphs.

4.1.1.2. Parameter tuning. The graph-based estimates of the mapping probabilities (Section 3.1.3) depend heavily on the maximum depth of the depth-first search. In order to find the best value for the search depth for each disambiguation context, we optimize the F_1 measure using the dataset from [99] as development data: while we cannot use these data as test set for evaluation, due to mismatches in the sense inventory,[15] they nevertheless provide a well-balanced dataset for estimating the optimal value of our search parameter.

4.1.1.3. Results and discussion. Evaluation is performed in terms of standard measures of precision (the ratio of correct sense labels to the non-empty labels output by the mapping algorithm), recall (the ratio of correct sense labels to the total of non-empty labels in the gold standard) and F_1-measure (a harmonic mean of precision and recall calculated as $\frac{2PR}{P+R}$). In addition

[14] We leave out the evaluation for different contexts of a Wikipage for the sake of brevity. However, during prototyping we found that the best results were obtained by using the largest context available, as reported in Table 1.

[15] In fact, the procedure for building the sense inventory of [99] collects as Wikipedia senses of an input lemma *all* Wikipages where the word occurs as anchor text of an internal link. For instance, given an occurrence of a link to RADIO PERSONALITY with anchor text host, the former is assumed to be a potential sense of the latter (a frequency threshold of 3 different articles for each hyperlink is employed to reduce the amount of noise from such free-form sense annotations).

we calculate accuracy, which also takes into account empty sense labels (that is, calculated on all 1,000 test instances). As baseline we use the most frequent WordNet sense (MFS), as well as a random sense assignment.

The results obtained from applying our mapping algorithms to the test data are reported in Table 1. Both in the case of accuracy and F_1-measure, the best results for the graph-based method are achieved using the optimal value of the search depth chosen during our tuning phase, which indicates that our findings and the behavior of this method generalize well across different datasets.

Our mapping methods achieve up to almost 78% F_1 and improve over the baselines by a large margin. Higher performance can be obtained by using more disambiguation information. That is, using a richer disambiguation context helps to better estimate the conditional probability of a WordNet sense given a Wikipage. In the case of the BoW method, the combination of taxonomic and gloss information attains a slight variation in terms of precision (-2.2% and -0.1% compared to taxonomic and gloss relations, respectively), but a significantly high increase in recall ($+17.8\%$ and $+13.8\%$). This implies that the information provided by different disambiguation contexts only partially overlap and, when used separately, each produces different mappings with a similar level of precision. When comparing the BoW with the graph-based method, the results support our intuitions from Section 3.1.3: by building disambiguation graphs, we are able to relate candidate senses of the Wikipedia page of interest with senses of the words in context by means of broad semantic relations which are not necessarily found in their immediate neighbors (as given in the bags of words): as a result, the graph-based method generally achieves a higher recall than the BoW approach (up to $+21.7\%$, $+15.3\%$ and $+11.8\%$ when using taxonomic relations, gloss-derived ones, and both, respectively), thus yielding an improvement of the F_1 measure of up to 2.7 points when using both taxonomic and gloss relations to explore the WordNet graph with maximum depth set to 2. More specifically, when analyzing the contribution of different WordNet relations to the construction of the disambiguation graph, we note that the F_1 performance when using only taxonomic relations increases with the maximum depth of the search. However, when using gloss-derived relations, we observe decreases in F_1-measure after a maximum depth of 3. This is because gloss-derived relations, in contrast to taxonomic ones, can be relatively noisy (for instance, act^3_n is related to $\mathsf{subdivision}^2_n$) and, while this still helps boost recall, it also leads to a considerable decrease in precision. This effect is amplified when combining both kinds of WordNet relations: the “recall up vs. precision down” effect needs to be counter-balanced by limiting the depth of the search. With maximum depth of 2 we are thus able to achieve the best result of 77.7% F_1-measure. A similar discussion concerns the accuracy of the different settings, thus showing that our mapping methodology is consistently robust with both empty and non-empty sense assignments.

As for the baselines, the most frequent sense is just 1.6% and 1.2% above the random baseline in terms of F_1 and accuracy, respectively. In fact, a χ^2 test on accuracy reveals no statistically significant difference at $p < 0.05$. This is a consequence of the random distribution of senses in our dataset and the “unbiased” Wikipedia coverage of WordNet senses (we hypothesize this coverage will not necessarily correlate with sense frequency information in WordNet). So selecting the first WordNet sense, rather than any other sense, for each target page represents a choice that is as arbitrary as picking a sense at random.

4.1.2. Mapping Wikipedia categories to WordNet

Our mapping methodology is general in nature. Thus, to corroborate the robustness of our approach, we report further experiments on linking Wikipedia categories to WordNet. To perform such mapping we apply exactly the same algorithm as the one used in Section 3.1.1, only defining a different category-specific disambiguation context as the set of words obtained from the following sources:

- **Sub/Super-category:** the lemmas[16] of the categories immediately dominating and dominated by the target Wikipage category within the category tree. For instance, the category PLAYS has super-categories DRAMA, THEATRE, LITERATURE BY GENRE and WORKS BY MEDIUM, as well as, among others, sub-categories PLAYS ADAPTED INTO FILMS and MASQUES, so we add to the disambiguation context the following words: drama, theatre, literature, work and masque.
- **Page titles:** the titles' lemmas of the pages contained in the category, together with their redirections. E.g., PLAYS contains the page PLAY (THEATRE), to which PLAYLET, STAGEPLAY, PLAYGOER and THEATRICAL PLAY all redirect. These pages thus contribute the following lemmas to the disambiguation context: playlet, stageplay, playgoer and play.
- **Page categories:** the lemmas of the categories found within the pages of the target Wikipage category. For instance, pages contained in the category PLAYS are categorized as DRAMA, THEATRE, PERFORMING ARTS and LITERATURE, so we add drama, theatre, performing art and literature to the disambiguation context.

For instance, $Ctx(\text{PLAYS}) = \{\mathsf{drama}, \mathsf{theatre}, \mathsf{literature}, \ldots, \mathsf{masque}, \mathsf{playlet}, \mathsf{stageplay}, \ldots, \mathsf{performing\ art}\}$.

4.1.2.1. Experimental setting. To evaluate the performance of our mapping algorithms when applied to Wikipedia categories, as well as to compare it with other competing proposals from the literature, we opted for the publicly available dataset presented in [126]. The dataset contains 200 nouns which are polysemous in WordNet and whose senses are mapped to 207 categories. Among the 207 categories, 37 (17.9%) correspond to more than one WordNet sense – e.g., the category COMMUNISTS corresponds to both WordNet senses of communist: this is due to the WordNet senses being usually finer-

[16] We use the method presented in [108, Section 3.2.1] to find lemmas of Wikipedia categories in WordNet.

Table 2
Performance on mapping Wikipedia categories to WordNet synsets. Systems marked with * were evaluated by 10-fold cross-validation. Underlined results are those using the best value for the maximum search depth, found using the dataset from [105] as development data.

Mapping method			Accuracy	
			w/o MFS	+ MFS
BoW		taxonomic	59.4	73.4
		gloss	43.5	70.0
		taxonomic + gloss	64.3	76.8
Graph	taxonomic relations			
	max depth	@ 2	66.2	72.5
		@ 3	69.1	71.5
		@ 4	71.0	72.5
	gloss relations			
	max depth	@ 2	59.9	69.6
		@ 3	60.9	66.2
		@ 4	60.9	63.8
	taxonomic + gloss relations			
	max depth	@ 2	**78.7**	**80.7**
		@ 3	74.9	75.4
		@ 4	68.6	69.1
		MFS baseline	64.7	
		Ponzetto & Navigli [105]	73.9	
Toral et al. [126]		best unsupervised	64.7	
		best supervised*	77.7	
		best voting	68.0	
		best unsupervised combination	65.7	
		best supervised combination*	77.2	
		oracle (upper bound)	84.5	

grained than the Wikipedia ones (cf. also the coarse-grained Word Sense Disambiguation evaluation in Section 6.2). Another 16 categories (7.7% of the dataset) have no corresponding sense in WordNet (for instance, CHIEF EXECUTIVES). Given that the dataset contains almost exclusively non-empty mappings, we also evaluated the setting in which our mapping algorithm always outputs a WordNet sense for an input category. That is, in case a WordNet sense cannot be assigned to a Wikipedia category (i.e., no sense is triggered by the disambiguation context or a tie occurs), we output the category's lemma most frequent sense (MFS) from SemCor.[17] To quantify the method's performance, we follow [126] and use accuracy as evaluation metric, as computed using the official Senseval scorer.[18] We compare our mapping algorithms against our previous category mapping approach [105], based on structural overlaps, as well as a variety of unsupervised and supervised approaches presented in [126], the latter ones evaluated using 10-fold cross-validation.

4.1.2.2. Parameter tuning. As in the case of mapping Wikipages, a fair evaluation of the graph-based method includes establishing the optimal value for the search depth on held-out data. To this end, we use the dataset from our previous work [105] as development data to find the value of the maximum search depth that yields the best accuracy. For tuning purposes only, to avoid noise deriving from the incorrect identification of the categories' syntactic heads, we use only instances containing single-word category labels (88 in total) – i.e., we do not consider categories with complex noun phrase labels such as THEATRE PRODUCTION COMPANIES or WORKS BASED ON PLAYS.

4.1.2.3. Results and discussion. The results, shown in Table 2, support our previous findings on mapping Wikipedia pages to WordNet senses. First, when compared to the BoW method, the graph-based one achieves higher performance by generating more accurate mappings (up to +17.4% when using gloss relations). All the graph-based settings beat the MFS baseline – a very difficult competitor for this task [126] – except for those that use only gloss-derived relations without the most frequent sense back-off. Using the MFS back-off, in turn, is always beneficial and represents a sensible choice for this particular task given that most of the Wikipedia categories have a corresponding WordNet sense. Moreover, we note that taxonomic and gloss-derived relations are complementary to each other: building disambiguation graphs which contain both kinds of semantic relation we are able to achieve the best performance, i.e., above 80% accuracy, which is only 4 points below the oracle system combination (i.e., a system that outputs the sense found in the gold standard, if found by any member of

[17] In a typical mapping task scenario, the MFS could not be used as a sensible fallback strategy when no mapping between Wikipedia categories and WordNet synsets can be established. We use it in the present context since, as mentioned above, the dataset we use is mostly targeted at evaluating the system's accuracy in establishing mappings, rather than in recognizing them.

[18] http://www.senseval.org/senseval3/scoring.

Table 3
Average number of distinct translations per English source word.

Corpus	Catalan	French	German	Italian	Spanish
SemCor	2.86	2.61	3.25	2.45	2.67
Wikipedia	1.80	1.73	1.74	1.67	1.71
SemCor ∪ Wikipedia	1.82	1.75	1.77	1.68	1.73

Table 4
Number of WordNet synsets and senses translated using SemCor and/or Wikipedia, as monosemous words, and their union.

	SemCor	Wikipedia	SemCor ∪ Wikipedia	monosemous	all
# synsets	3,901	31,308	33,359	62,259	68,554
# senses	6,852	35,372	40,504	101,853	115,606

an ensemble [126]). The additional richness provided by graphs with more relations at higher depth can, however, lead to incorrect sense assignments and a degradation of the overall results. The best results are therefore achieved when using taxonomic and gloss relations and limiting the depth of the search to 2: in this setting, we are able to beat any of the approaches presented in [126] by a large margin. As in the case of mapping Wikipages, the best results on the test data are achieved using the optimal value of the search depth chosen during our tuning phase, thus corroborating the generality of our findings across datasets.

4.2. Translation evaluation

So far we have focused on the quality of the mapping between WordNet and Wikipedia. We next concentrate on analyzing the other major component of our approach, namely the use of a Machine Translation system. We first present two descriptive analyses, in order to characterize the contribution made by the use of an MT system. We analyze the kind of output the MT system returns by quantifying its richness in terms of output diversification (Section 4.2.1), and the amount of WordNet synsets which are translated using different sources such as SemCor or our Wikipedia corpus (Section 4.2.2). We then move on to perform a second set of experiments, aimed at quantifying BabelNet's coverage against gold-standard resources (Section 4.2.3), i.e., manually assembled wordnets in languages other than English. Finally, we also perform an additional manual evaluation of the extra coverage provided by BabelNet (Section 4.2.4).

4.2.1. Degree of translation output diversification

In order to test whether the lexical knowledge contained in BabelNet is an artifact of translating Babel synsets, we first quantify how rich the output of the MT system is in terms of the average number of translations that it outputs for each English source word. The results are presented in Table 3, where we compute the statistics for the translations obtained separately from SemCor and Wikipedia, as well as both corpora together. Our results indicate that the number of distinct translations per word is much higher for SemCor than Wikipedia. A closer look at the output reveals, in fact, that the sense inventory of Wikipedia contains many specialized senses, such as Broadway theater, or Roman amphitheatre, which can have only one or two translations at most. This is in contrast to the more general vocabulary provided by senses annotated in SemCor, e.g., including different senses for play, which allow for a far higher variety in the translation output. Finally, combining both corpora attenuates the effects of SemCor's translation variety, essentially due to its smaller size in comparison to the Wikipedia corpus.

4.2.2. Analysis of the translated WordNet

During the construction of BabelNet we translate WordNet synsets using heterogeneous sources, namely SemCor, BabelCor (our sense-annotated Wikipedia corpus), as well as the translation of monosemous words. Thus, in order to give a better idea of the different contributions of these translation sources in helping us build the 'core' of BabelNet, i.e., its region where lexicographic (WordNet's) and encyclopedic (Wikipedia's) knowledge meet, we present in Table 4 statistics of how many WordNet synsets are translated using SemCor and/or Wikipedia, or monosemous word translations, or all three combined. The figures show that, due to its limited size and our additional requirement of a minimum of 3 annotated sentences per sense (cf. Section 3.2), SemCor is simply too small to provide a substantial number of translations for BabelNet. By complementing SemCor at different times with sense translations from Wikipedia, as well as monosemous word translations, we are able to translate a substantial portion of WordNet, consisting of 83.4% and 79.0% of its overall 82,115 and 146,312 nominal synsets and senses, respectively. Note that the large contribution of translations from monosemous words is due to the content of WordNet itself, where 101,863 out of 117,798 nominal lemmas are, in fact, monosemous.

4.2.3. Automatic evaluation of translations

4.2.3.1. Datasets. To compute the coverage of BabelNet against gold-standard wordnets, we use the following manually assembled lexical knowledge bases, whose size in terms of number of synsets and word senses is reported in Table 5:

Table 5
Size of the gold-standard wordnets.

	Catalan	French	German	Italian	Spanish
Word senses	64,171	44,265	15,762	57,255	83,114
Synsets	40,466	31,742	9,877	32,156	55,365

- **Catalan and Spanish:** the Multilingual Central Repository [7];
- **French:** WOrdnet Libre du Français [118, WOLF];
- **German:** the subset of GermaNet [62] included in EuroWordNet for German;
- **Italian**: MultiWordNet [104].

All wordnets are linked to the English WordNet, which in turn is contained in BabelNet: this allows us to quantify their degree of overlap with BabelNet without the need to (automatically or manually) map across resources – i.e., the synsets of the English WordNet act as a pivot.

4.2.3.2. Evaluation measures. Let $\mathcal{B}$ be BabelNet, $\mathcal{F}$ our gold-standard non-English wordnet (e.g., GermaNet), and let $\mathcal{E}$ be the English WordNet. Given a synset $S_\mathcal{F} \in \mathcal{F}$, we denote its corresponding Babel synset as $S_\mathcal{B}$ and its synset in the English WordNet as $S_\mathcal{E}$. We then quantify the coverage of BabelNet against our gold-standard wordnets both in terms of synsets and word senses. For synsets, we calculate coverage as follows:

$$\mathrm{SynsetCov}(\mathcal{B},\mathcal{F}) = \frac{\sum_{S_\mathcal{F}\in\mathcal{F}} \delta(S_\mathcal{B}, S_\mathcal{F})}{|\{S_\mathcal{F}\in\mathcal{F}\}|}, \tag{4}$$

where the function $\delta(S_\mathcal{B}, S_\mathcal{F})$ is 1 if the two synsets $S_\mathcal{B}$ and $S_\mathcal{F}$ have a synonym in common, 0 otherwise. That is, synset coverage is determined as the percentage of synsets of $\mathcal{F}$ that share a term with the corresponding Babel synsets. For word senses we calculate a similar measure of coverage:

$$\mathrm{WordCov}(\mathcal{B},\mathcal{F}) = \frac{\sum_{S_\mathcal{F}\in\mathcal{F}} \sum_{s_\mathcal{F}\in S_\mathcal{F}} \delta'(S_\mathcal{B}, s_\mathcal{F})}{|\{s_\mathcal{F}\in S_\mathcal{F} : S_\mathcal{F}\in\mathcal{F}\}|}, \tag{5}$$

where $s_\mathcal{F}$ is a word sense in synset $S_\mathcal{F}$ and $\delta'(S_\mathcal{B}, s_\mathcal{F}) = 1$ if $s_\mathcal{F} \in S_\mathcal{B}$, 0 otherwise. That is, we calculate the ratio of word senses in our gold-standard resource $\mathcal{F}$ that also occur in the corresponding synset $S_\mathcal{B}$ to the overall number of senses in $\mathcal{F}$.

Computing coverage provides only part of the picture. In fact, while our gold-standard wordnets cover only a portion of the English WordNet, the overall coverage of BabelNet is much higher. We thus calculate *extra coverage* for synsets as the proportion of WordNet synsets which are covered by BabelNet but not by the reference resource $\mathcal{F}$:

$$\mathrm{SynsetExtraCov}(\mathcal{B},\mathcal{F}) = \frac{\sum_{S_\mathcal{E}\in\mathcal{E}\setminus\mathcal{F}} \delta(S_\mathcal{B}, S_\mathcal{E})}{|\{S_\mathcal{F}\in\mathcal{F}\}|}. \tag{6}$$

Similarly, we calculate extra coverage for word senses found in BabelNet and contained within WordNet synsets which are not covered by the reference resource $\mathcal{F}$:

$$\mathrm{WordExtraCov}(\mathcal{B},\mathcal{F}) = \frac{\sum_{S_\mathcal{E}\in\mathcal{E}\setminus\mathcal{F}} \sum_{s_\mathcal{E}\in S_\mathcal{E}} \delta'(S_\mathcal{B}, s_\mathcal{E})}{|\{s_\mathcal{F}\in S_\mathcal{F} : S_\mathcal{F}\in\mathcal{F}\}|}. \tag{7}$$

4.2.3.3. Results and discussion. In order to evaluate the different contributions of Wikipedia's inter-language links and our approach to filling translation gaps by means of a Machine Translation system, we evaluate coverage and extra coverage of word senses and synsets at different stages:

(a) using only the inter-language links from Wikipedia (Wiki Links);
(b) using only the automatic translations of the sentences from Wikipedia (Wiki Transl.);
(c) using only the automatic translations of the sentences from SemCor (WordNet Transl.);
(d) using all available translations, i.e., BabelNet.

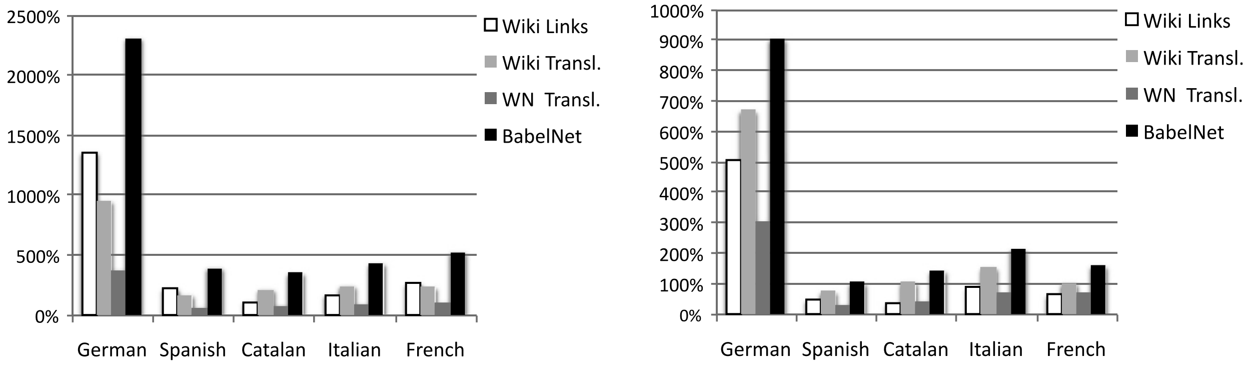
We report coverage results in Table 6. The percentage of word senses covered by BabelNet ranges from 52.9% (Italian) to 66.4 (Spanish) and 86.0% (French). Synset coverage ranges from 73.3% (Catalan) to 76.6% (Spanish) and 92.9% (French). Synset coverage is higher because a synset in the reference resource is considered to be covered if it shares at least one word with the corresponding Babel synset. Details on extra coverage – which quantifies the amount of word senses and synsets in the English WordNet for which BabelNet, but not the non-English gold-standard resources, is able to provide a translation – are given in Table 7 and Fig. 5. The results show that we provide for all languages a high extra coverage both at the word sense level – ranging from 340% (Catalan) to 2,298% (German) – and at the synset level – ranging from 102%

Table 6
Coverage against gold-standard wordnets (percentages).

	WordCov (Senses)				SynsetCov (Synsets)			
Resource	Wiki		WordNet	BabelNet	Wiki		WordNet	BabelNet
Method	Links	Transl.	Transl.	All	Links	Transl.	Transl.	All
Catalan	20.3	46.9	25.0	**64.0**	25.2	54.1	29.6	**73.3**
French	70.0	69.6	16.3	**86.0**	72.4	79.6	19.4	**92.9**
German	39.6	42.6	21.0	**57.6**	50.7	58.2	28.6	**73.4**
Italian	28.1	39.9	19.7	**52.9**	40.0	58.0	28.7	**73.7**
Spanish	34.4	47.9	25.2	**66.4**	40.7	56.1	30.0	**76.6**

Table 7
Extra coverage against gold-standard wordnets (percentages).

	WordExtraCov (Senses)				SynsetExtraCov (Synsets)			
Resource	Wiki		WordNet	BabelNet	Wiki		WordNet	D..BabelNet
Method	Links	Transl.	Transl.	All	Links	Transl.	Transl.	All
Catalan	100	204	71	**340**	35	105	42	**142**
French	255	223	92	**514**	63	102	67	**159**
German	1349	940	367	**2298**	506	668	303	**902**
Italian	160	234	83	**419**	87	153	68	**213**
Spanish	214	158	56	**384**	48	74	30	**102**



(a) word senses

(b) synsets

Fig. 5. Extra coverage against gold-standard wordnets: word senses (a) and synsets (b).

(Spanish) to 902% (German). Cases of novel translations not found in the non-English wordnets include, for instance, autore teatrale$_{\mathrm{IT}}$, théâtre de rue$_{\mathrm{FR}}$ or Theatersaison$_{\mathrm{DE}}$.

Tables 6 and 7 show that the best results are obtained when combining *all* available translations, i.e., both from Wikipedia and the machine translation system. The performance figures suffer from the errors of the mapping phase (see Section 4.1). Nonetheless, the results are generally high, with a peak for French, since WOLF was created semi-automatically by combining several resources, *including* Wikipedia. The relatively lower word sense coverage for Italian (52.9%) is, instead, due to the lack of many common words in the gold-standard synsets. Examples include playwright$_{\mathrm{EN}}$ translated only as drammaturgo$_{\mathrm{IT}}$ but not as the equally common autore teatrale$_{\mathrm{IT}}$, theatrical production$_{\mathrm{EN}}$ translated as allestimento$_{\mathrm{IT}}$ but not as produzione teatrale$_{\mathrm{IT}}$ or messa in scena$_{\mathrm{IT}}$, etc.

4.2.4. Manual evaluation of translations

4.2.4.1. Experimental setup. The previous evaluation quantified the extent to which the non-English wordnets are covered by BabelNet. However, that evaluation does not say anything about the precision of the *additional* lexicalizations that BabelNet provides. Given that BabelNet shows a remarkably high 'added value' in terms of extra coverage – ranging from 340% to 2,298% of the national wordnets (see Fig. 5) – we need to perform a manual evaluation to assess the quality of these novel translations. In order to build a dataset of translations to be validated by human annotators, we selected for each of the five languages other than English a random set of 600 Babel synsets composed as follows: 200 synsets whose senses exist in WordNet only, 200 synsets in the intersection between WordNet and Wikipedia (i.e., those mapped with our method illustrated in Section 3.1), 200 synsets whose lexicalizations exist in Wikipedia only. Our validation dataset thus includes $600 \times 5 = 3{,}000$ Babel synsets, none of which is covered by any of the five non-English gold-standard wordnets. The Babel synsets were manually validated by expert annotators who decided which senses (i.e., lexicalizations) were appropriate

Table 8
Precision of BabelNet on synonyms in WordNet (WN), Wikipedia (Wiki) and their intersection (WN ∩ Wiki): percentage and total number of words (in parentheses) are reported.

Resource	Catalan	French	German	Italian	Spanish
WordNet	75.58 (258)	67.16 (268)	73.76 (282)	72.32 (271)	69.45 (275)
Wiki	92.71 (398)	96.44 (758)	97.74 (709)	99.09 (552)	92.46 (703)
WordNet ∩ Wiki	82.98 (517)	77.43 (709)	78.37 (777)	80.83 (574)	78.53 (643)

Table 9
Number of lemmas, synsets and word senses in the 6 languages currently covered by BabelNet.

Language	Lemmas	Synsets	Word senses
English	5,938,324	3,032,406	6,550,579
Catalan	3,518,079	2,214,781	3,777,700
French	3,754,079	2,285,458	4,091,456
German	3,602,447	2,270,159	3,910,485
Italian	3,498,948	2,268,188	3,773,384
Spanish	3,623,734	2,252,632	3,941,039
Total	23,935,611	3,032,406	26,044,643

given the corresponding WordNet gloss and/or Wikipage. Note that the synsets that intersect with Wikipedia may have translations from Wikipedia links and Wikipedia translated sentences, whereas the synsets that intersect with WordNet may have translations from SemCor sentences (cf. Section 4.2.3).

4.2.4.2. Results and discussion. In Table 8 we report precision (i.e., the percentage of lexicalizations deemed correct) and, in parentheses, the total number of translations evaluated for each language (columns) in one of the three regions of BabelNet (rows). The results show that different regions of BabelNet contain translations of different quality: WordNet-only synsets have a precision of around 72%, which considerably increases by using translations from Wikipedia links and sense-labeled data (around 80% in the intersection and 95% with Wikipedia-only translations). The absolute numbers in parentheses indicate that the number of translations from Wikipedia is higher than that of WordNet: this is due to our method collecting many translations from the redirections found in the wikipedias of the target languages (Section 3.2), as well as to the paucity of examples in SemCor. In addition, some of the synsets in WordNet with no Wikipedia counterpart are very difficult to translate. Examples include terms like yodeling, crape fern, peri, and many others for which we could not find translations in major editions of bilingual dictionaries. In contrast, good translations were produced using our machine translation method when enough sentences were available. Examples include, among others, Laientheater$_{\mathrm{DE}}$ for amateur theatre$_{\mathrm{EN}}$, attore cinematografico$_{\mathrm{IT}}$ for film actor$_{\mathrm{EN}}$, etc.

5. Anatomy of BabelNet

In this section we provide statistics for the current version of BabelNet, obtained by applying the construction methodology described in Section 3, and evaluated in Section 4.

5.1. WordNet-Wikipedia mapping configuration

The version of BabelNet that we describe in this section is based on the best performing mapping technique among those evaluated in Section 4.1.1 (in terms of accuracy and balanced F-measure) – i.e., a graph-based method combining taxonomic and gloss relations at maximum depth of 2. The overall mapping contains 89,226 pairs of Wikipages and word senses they map to, covers 52% of the noun senses in WordNet, and has a reported accuracy of more than 82% (cf. Table 1).

The WordNet–Wikipedia mapping contains 72,572 lemmas, 10,031 and 26,398 of which are polysemous in WordNet and Wikipedia, respectively. Our mapping thus covers at least one sense for 62.9% of WordNet's polysemous nouns (10,031 out of 15,935): these polysemous nouns can refer to 44,449 and 71,918 different senses in WordNet and Wikipedia, respectively, 13,241 and 16,233 of which are also found in the mapping.

5.2. Lexicon

BabelNet currently covers 6 languages, namely: English, Catalan, French, German, Italian and Spanish. Its lexicon includes lemmas which denote both concepts (e.g., dramatic play) and named entities (e.g., Shakespeare). The second column of Table 9 shows the number of lemmas for each language. The lexicons have the same order of magnitude for the 5 non-English languages, whereas English shows larger numbers due to the lack of inter-language links and annotated sentences for many terms, which prevents our construction approach from providing translations.

Table 10
Number of monosemous and polysemous words by part of speech (verbs, adjectives and adverbs are the same as in WordNet 3.0).

POS	Monosemous words	Polysemous words
Noun	22,763,265	1,134,857
Verb	6,277	5,252
Adjective	1,503	4,976
Adverb	3,748	733
Total	22,789,793	1,145,818

Table 11
Composition of Babel synsets: number of synonyms from the English WordNet, Wikipedia pages and translations, as well as translations of WordNet's monosemous words and SemCor's sense annotations.

		English	Catalan	French	German	Italian	Spanish	Total
English WordNet		206,978	–	–	–	–	–	206,978
Wikipedia	pages	2,955,552	123,101	524,897	506,892	404,153	349,375	4,863,970
	redirections	3,388,049	105,147	617,379	456,977	217,963	404,009	5,189,524
	translations	–	3,445,273	2,844,645	2,841,914	3,046,323	3,083,365	15,261,520
WordNet	monosemous	–	97,327	97,680	97,852	98,089	97,435	488,383
	SemCor	–	6,852	6,855	6,850	6,856	6,855	34,268
Total		6,550,579	3,777,700	4,091,456	3,910,485	3,773,384	3,941,039	26,044,643

Table 12
Number of lexico-semantic relations harvested from WordNet, WordNet glosses and the 6 wikipedias.

	English	Catalan	French	German	Italian	Spanish	Total
WordNet	364,552	–	–	–	–	–	364,552
WordNet glosses	617,785	–	–	–	–	–	617,785
Wikipedia	50,104,884	978,006	5,613,873	5,940,612	3,602,395	3,411,612	69,651,382
Total	51,087,221	978,006	5,613,873	5,940,612	3,602,395	3,411,612	70,633,719

In Table 10 we report the number of monosemous and polysemous words divided by part of speech. Given that we work with nominal synsets only, the numbers for verbs, adjectives and adverbs are the same as in WordNet 3.0. As for nouns, we observe a very large number of monosemous terms (almost 23 million), but also a large number of polysemous terms (more than 1 million). Both numbers are considerably larger than in WordNet, because – as remarked above – terms here denote both concepts (mainly from WordNet) and named entities (mainly from Wikipedia).

5.3. Concepts

BabelNet contains more than 3 million concepts, i.e., Babel synsets, and more than 26 million word senses (regardless of their language). In Table 9 we report the number of synsets covered for each language (third column) and the number of word senses lexicalized in each language (fourth column). 72.3% of the Babel synsets contain lexicalizations in all 6 languages and the overall number of word senses in English is much higher than those in the other languages (owing to the high number of synonyms available in the English WordNet synsets). Each Babel synset contains 8.6 synonyms, i.e., word senses, on average, in any language. The number of synonyms per synset for each language individually ranges from a maximum 2.2 for English to a minimum 1.7 for Italian, with an average of 1.8 synonyms per language.

In Table 11 we show for each language the number of word senses obtained directly from WordNet, Wikipedia pages and redirections, as well as Wikipedia and WordNet translations (as a result of the translation process described in Section 3.2).

5.4. Relations

We now turn to relations in BabelNet. Relations come either from Wikipedia hyperlinks (in any of the covered languages) or WordNet. All our relations are semantic, in that they connect Babel synsets (rather than senses), however the relations obtained from Wikipedia are unlabeled.[19] In Table 12 we show the number of lexico-semantic relations from WordNet, WordNet glosses and the 6 wikipedias used in our work. We can see that the major contribution comes from the English Wikipedia (50 million relations) and wikipedias in other languages (a few million relations, depending on their size in terms of number of articles and links therein).

[19] In a future release of the resource we plan to perform an automatic labeling based on work in the literature. See [81] for recent work on the topic.

Table 13
Glosses for the Babel synset referring to the concept of play as ‘dramatic work’.

English	WordNet	A dramatic work intended for performance by actors on a stage.
	Wikipedia	A play is a form of literature written by a playwright, usually consisting of scripted dialogue between characters, intended for theatrical performance rather than just reading.
Catalan		El drama en termes generals és una obra literària o una situació de la vida real que resulta complexa i difícil però amb un final favorable o feliç.
French		Le drame (du latin drama, emprunté au grec ancien δρᾶμα/*drâma*, qui signifie action (théâtrale), pièce de théâtre) désigne étymologiquement toute action scénique.
German		Drama (altgriechisch δρᾶμα dráma ‘Handlung’) ist ein Oberbegriff für Texte mit verteilten Rollen.
Italian		Un dramma, dal greco “drama” (azione, storia; da δραν, *fare*), è una forma letteraria che include parti scritte per essere interpretate da attori.
Spanish		Drama (del griego δρᾶμα, hacer o actuar) es la forma de presentación de acciones a través de su representación por actores.

5.5. Glosses

Each Babel synset naturally comes with one or more glosses (possibly available in many languages). In fact, WordNet provides a textual definition for each English synset, while in Wikipedia a textual definition can be reliably obtained from the first sentence of each Wikipage.[20] Overall, BabelNet includes 4,683,031 glosses (2,985,243 of which are in English). In Table 13 we show the glosses for the Babel synset which refers to the concept of play as ‘dramatic work’.

5.6. Sense-tagged corpus

BabelNet also includes a sense-tagged corpus containing the sentences input to the Machine Translation system (cf. Section 3.2). The corpus, called BabelCor, is built by collecting from SemCor and Wikipedia those sentences which contain an occurrence of a polysemous word labeled with a WordNet sense (in SemCor) or hyperlinked to a Wikipage (in Wikipedia). A frequency threshold of at least 3 sentences per sense is used in order to make sure that meaningful statistics are computed from the MT system’s output, thus ensuring precision. As a result, BabelCor contains almost 2 million sentences (1,986,557 in total, of which 46,155 from SemCor and 1,940,402 from Wikipedia), which provide sense-annotated data for 330,993 senses contained in BabelNet (6,856 from WordNet and 324,137 from Wikipedia).

6. Extrinsic evaluation

In this section we present a set of three *extrinsic* evaluations quantifying the impact of BabelNet against a variety of benchmarking datasets. Crucially, the next three subsections aim to show that state-of-the-art performance can be achieved and surpassed when BabelNet is used as the component providing the knowledge needed to perform lexical disambiguation both at the monolingual and multilingual level. Specifically, we tested BabelNet on three SemEval tasks: evaluation of wide-coverage knowledge resources (Section 6.1), coarse-grained English all-words Word Sense Disambiguation (Section 6.2) and cross-lingual Word Sense Disambiguation (Section 6.3).

6.1. Evaluation of wide-coverage knowledge resources

We first evaluate BabelNet using the SemEval-2007 task 16 on evaluating wide-coverage knowledge resources [27] (KBEval henceforth). In this task, a variety of knowledge bases are assessed by first generating so-called topic signatures [65] and then using these to perform monolingual Word Sense Disambiguation (WSD) on standard datasets from previous Senseval [74] and SemEval [109] competitions.

KBEval provides a unified framework for comparing different knowledge bases while being as neutral as possible as regards the specific properties of each resource. First, given a concept, a topic signature – i.e., a weighted word vector – is generated: the vector elements represent words which are related to the initial concept, together with their strength of association. These topic signatures are then used to perform WSD: given a test sentence, namely a target word in context, we consider the topic signatures for each of the target word’s senses. Then, we compute a simple score based on word overlap between each of these topic signatures and the test sentence. The word sense with the highest overlap is selected. Note that this represents, in practice, a simple unsupervised WSD algorithm which aims at maximizing the lexical overlap between the target word’s context and the vectors assigned to each concept in the knowledge resource.

[20] “The article should begin with a short declarative sentence, answering two questions for the nonspecialist reader: *What (or who) is the subject?* and *Why is this subject notable?*”, extracted from http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles. This simple, albeit powerful, heuristic has been previously used successfully to construct a corpus of definitional sentences [97] and learn a definition and hypernym extraction model [95].

Table 14
Excerpt of topic signatures for different senses of future using BabelNet-1 and BabelNet-2.

(a) BabelNet-1

future^1_n	futuristic:1.0, time to come:1.0, past:0.35, tomorrow:0.12, foretelling:0.06, prognostication:0.06
future^2_n	tense:1.0, perfect:0.44, grammar:0.29, grammatical aspect:0.14, linguistics:0.14, verb:0.13
future^3_n	finance:1.0, commodities trade:0.07, goods economics:0.08, price:0.04, buy:0.03, purchase:0.03

(b) BabelNet-2

future^1_n	futurity:1.0, yesteryear:0.68, eternalism:0.42, fiction:0.4, timeline:0.33, oracle:0.31, doomsday:0.17
future^2_n	tense:1.0, perfect:0.44 modality:0.42, auxiliary:0.41, active voice:0.21, grammatical tense:0.20
future^3_n	finance:1.0, dollar:0.30, turnover:0.29, nominative:0.27, economics:0.17, service:0.17, law:0.14

Table 15
Results on the SemEval-2007 task 16: Evaluation of wide coverage knowledge resources.

(a) Senseval-3 English Lexical Sample task:

Knowledge base	P	R	F_1	Avg. size
TRAIN	65.1	65.1	65.1	450
TRAIN-MFS	54.5	54.5	54.5	–
WN-MFS	53.0	53.0	53.0	–
SEMCOR-MFS	49.0	49.1	49.0	–
TSSEM	52.5	52.4	52.4	103
BabelNet-1	44.3	44.3	44.3	119
BabelNet-2	35.0	35.0	35.0	2,128
KnowNet-20	44.1	44.1	44.1	610
RANDOM	19.1	19.1	19.1	–

(b) SemEval-2007 English Lexical Sample (task 17):

Knowledge base	P	R	F_1	Avg. size
TRAIN	87.6	87.6	87.6	450
TRAIN-MFS	81.2	81.2	81.2	–
WN-MFS	66.2	59.9	62.9	–
SEMCOR-MFS	42.4	38.4	40.3	–
WN + XWN + KN-20	53.0	53.0	53.0	627
BabelNet-1	52.2	46.3	49.1	130
BabelNet-2	56.9	53.1	54.9	2, 352
KnowNet-20	49.5	46.1	47.7	561
RANDOM	19.1	19.1	19.1	–

In the case of BabelNet, given an input word, we first collect all Babel synsets where the word occurs as a WordNet synonym. For each Babel synset, we then generate a topic signature for the synset by adopting the same method used by the SemEval task organizers for other semantic networks [27]: that is, we collect all Babel synsets that can be reached from the initial synset at distance 1 (‘BabelNet-1’) or 2 (‘BabelNet-2’) and then output all their English lexicalizations. In Table 14 we show an excerpt of the topic signatures for the senses of the word future$_n$ using BabelNet.

Results for BabelNet, together with other competing knowledge resources, are presented in Table 15. Evaluation is performed using standard metrics of precision, recall and F_1-measure on the datasets from the Senseval-3 and SemEval-2007 English lexical sample tasks. Following the original task evaluation report [27], we also report the average size of the topic signature per word sense. Since words in these datasets are all annotated with WordNet senses, which thus provides the sense inventory, for each instance of a target word in the test set we output its WordNet sense found in the highest-scoring Babel synset. For each dataset we compare with a variety of standard baselines for the task:

- **RANDOM**, which performs a random sense assignment (lower bound).
- **SEMCOR-MFS**, which assigns the most frequent sense of a word found in the SemCor sense-tagged corpus.
- **WN-MFS**, which selects the first sense of the target word, as found in WordNet 1.6.
- **TRAIN-MFS**, which assigns the most frequent sense of the target word, as found in the training data of each dataset.
- **TRAIN**, which builds a topic signature for each word sense directly from the training data based on a TF$*$IDF scoring. As pointed out in [27], whereas in a standard WSD scenario this would represent a somewhat simple supervised approach, in the context of this evaluation framework it provides, instead, an upper bound, since no better topic signatures can be created for a sense than those derived from its own annotated corpus.

In addition, we also compare the performance of BabelNet with KnowNet-20 [28] – to date the best performing resource not making use of sense-annotated data – and the best knowledge bases for each of the Senseval and SemEval datasets, namely:

- **TSSEM**, which provides topic signatures based on SemCor. For each word sense found in SemCor, a word vector is built by collecting all co-occurring words from its sentences, which are then weighted using a TF$*$IDF scoring. This is the best performing knowledge base for the Senseval-3 English lexical sample dataset.
- **WN + XWN + KN-20**, a combination of WordNet, eXtended WordNet [75] and KnowNet-20, which was the best performing resource on the SemEval-2007 English lexical sample task.

The results show that BabelNet is highly competitive in this task by achieving the best performance among all knowledge resources not making use of sense-annotated data on the Senseval-3 dataset, as well as the best overall results on the SemEval-2007 data. In the case of the Senseval-3 data we perform better than KnowNet-20 ($+0.2$ P/R/F_1) and also provide topic signatures of smaller average size (491 words less per topic on average). Similarly to the findings from [28] for all the knowledge resources that they consider, the behavior of BabelNet on the SemEval-2007 data is quite different from its behavior on the Senseval-3 dataset: on the SemEval dataset, an increasingly better performance is achieved by generating larger topic signatures for all words by collecting concept lexicalizations at a distance of 2. In this setting, in fact, thanks to a high recall, we are able to achieve the best overall performance in terms of balanced F-measure on the SemEval-2007 dataset, thus also improving on an integrated resource (i.e., $+1.9$ F_1 improvement vs. WN + XWN + KN-20) which combines high-quality manually curated knowledge (from WordNet and eXtended WordNet) with large amounts of automatically acquired semantic relations (from KnowNet-20). To better understand these performance trends we evaluated BabelNet in another experimental setting where we used only WordNet or Wikipedia relations at different times. On both datasets we observed consistent behavior in that both subsets of relations yielded lower results when used separately (e.g., 49.9 and 43.3 F_1 on SemEval-2007 for WordNet and Wikipedia-only relations at distance 2, respectively), and the best overall results were obtained by their combination (shown in Table 15). This indicates the complementarity of the relations found in the two resources (mostly taxonomic relations from WordNet and topically associative ones from Wikipedia) and their mutual benefits for relation expansion in the given lexico-semantic task. Finally, the different results on the Senseval-3 and SemEval-2007 data are primarily due to the granularity of the sense inventory – i.e., fined-grained WordNet senses for Senseval vs. coarse-grained ones from OntoNotes [51] for SemEval. The availability of a coarse-grained sense inventory such as that of OntoNotes leads, in fact, to higher performance figures [140]: behavior which we now move to investigate more closely.

6.2. Coarse-grained English all-words Word Sense Disambiguation

We now extrinsically evaluate the impact of BabelNet on the SemEval-2007 coarse-grained all-words WSD task [90]. Again, we perform an extrinsic evaluation in a lexical disambiguation task, since WSD has been shown in the literature to offer a solid benchmark for knowledge-rich approaches to lexical semantics [87], mostly due to the fact that knowledge is indeed an essential requirement for robust lexical processing. But while knowledge in the form of labeled examples for training supervised models is known to be time-consuming to produce and to scale poorly [31], the information contained in lexical resources such as WordNet has also been found to be insufficient for high-performance WSD [26,89]. Thus, we explore in our experiments the potential of a highly-interconnected semantic network such as BabelNet for robust knowledge-based WSD: BabelNet embeds WordNet and extends it with millions of topical, semantic associative relations and, accordingly, it represents a natural choice for testing whether large amounts of high-quality knowledge can provide the basis of a WSD system which is able to compete with state-of-the-art supervised approaches.

Performing experiments in a coarse-grained setting is a natural choice for several reasons: first, it has been argued that the fine granularity of the WordNet sense inventory is one of the main obstacles to accurate WSD (cf. the discussion in [86,87]); second, the meanings of Wikipages are intuitively coarser than those in WordNet.[21] For instance, mapping PLAY (THEATRE) to the first or third sense in WordNet is an arbitrary choice, as the Wikipage refers to both senses. Finally, given their different nature, WordNet and Wikipedia do not fully overlap. WordNet-only Babel synsets, namely those made up entirely of WordNet senses for which no corresponding Wikipage could be found, do not benefit from the richness provided by additional relations from Wikipedia and typically suffer from poor connectivity. This is to say, semantic relations from Wikipedia can heavily skew the distribution of outgoing links for the different senses of a word and penalize those fine-grained senses for which no mapping could be found. For instance, mapping PLAY (THEATRE) and PLAY (ACTIVITY) to play_n^1 and play_n^8 respectively, implies that the outdegree of their corresponding Babel synsets will be higher compared to unmapped WordNet senses and, vice versa, the probability of selecting other senses of play_n will decrease.

[21] Note that our polysemy rates from Section 4.1 also include Wikipages whose lemma is contained in WordNet, but which have out-of-domain meanings, i.e., encyclopedic entries referring to specialized named entities such as e.g., PLAY (TV SERIES) or ACT (BAND). We computed the polysemy rate for a random sample of 20 polysemous words by manually removing these NEs and found that Wikipedia's polysemy rate is indeed lower than that of WordNet – i.e., average polysemy of 2.1 vs. 2.8.

Similarly to the experimental setting of [27], we use WSD as a benchmarking extrinsic assessment of the quality of the information contained in BabelNet. To achieve this, we define a general framework for transforming an input context, consisting of a set of target words in context to be disambiguated, into a graph, based on the lexicon and semantic relations contained in different knowledge resources, i.e., WordNet and BabelNet. We then use this *context graph* to apply a variety of algorithms for graph-based lexico-semantic disambiguation and take the results to indicate the quality of the underlying knowledge resource used to construct the graph.

6.2.1. Context graph construction

We follow [89] and construct a directed graph $G = (V, E)$ for an input word sequence $\sigma = (w_1, \ldots, w_n)$ containing a set of words to be disambiguated,[22] based on the lexical and semantic relations found in a given knowledge resource *KB* – i.e., WordNet or BabelNet, in our case. In order to build the graph, we follow the same procedure used to create graphs for estimating mapping probabilities (Section 3.1.3). The result of this procedure is a subgraph of *KB* which contains all senses of the words found in σ, as well as all edges and intermediate senses found in *KB* along the paths of maximal length L that connect them.

6.2.2. Edge filtering

In order to ensure that our context graphs are noise-free, that is, that they do not contain weak semantic links which potentially degrade the performance of the WSD algorithms that use them, we define a set of filters to constrain the set of paths that are used to build them:

- **Minimum edge weight:** remove all edges from the graph whose weight is below a certain threshold;
- **Sense shifts:** filter out all paths connecting different senses of the same word. This is to avoid the risk that senses of the same word might reinforce each other, thus reducing the empirical evidence provided by the senses of the other words in context (cf. also [5] and [89]).

6.2.3. Disambiguation algorithms

Given the above graph representation $G = (V, E)$ of a word sequence σ, we view WSD as a ranking problem. For each word $w_i \in \sigma$, we want to rank its senses $Senses_{KB}(w_i) \subseteq V$ based on some connectivity measure applied to G, and output the most appropriate meaning for w_i, i.e.:

$$\hat{s}_i = \operatorname*{argmax}_{s \in Senses_{KB}(w_i)} score(s).$$

In this paper we explore four different measures which are expected to capture different aspects of the network's topology:

- **Degree centrality (Degree):** The first algorithm relies on the notion of vertex degree and ranks the senses of a given word in the context graph based on the number of their outgoing edges:

  $$score(s) = \left|\left\{(s, v) \in E : v \in V\right\}\right|. \tag{8}$$

  In practice, this connectivity measure weights a sense as more appropriate if it has a high degree, and for each word in context the sense with the highest vertex degree is selected. While this is a relatively simple connectivity measure which does not fully take into account the connecting paths found in the context graph, it has been shown to yield competitive performance [89,106].
- **Inverse path length sum (PLength):** The second scoring measure aims at ranking the senses of the target word by looking at the *full* connecting paths found in the context graph, instead of considering only the incident edges as in the case of Degree. This measure in fact scores a sense by summing up the scores of all paths which connect it to other senses in the context graph:

  $$score(s) = \sum_{p \in paths(s)} \frac{1}{e^{length(p)-1}}, \tag{9}$$

  where $paths(s)$ is the set of paths connecting s to other senses of context words, $length(p)$ is the number of edges in the path p and each path is scored with the exponential inverse decay of the path length.
- **Path probability sum (SProbability):** This measure also computes the score of a sense as the sum of the scores of its outgoing paths. However, it uses an alternative measure for scoring paths in the graph, which is sensitive to the weights of each single edge. In fact, this measure scores each path by computing its probability, under the assumption that edges are independent of each other:

  $$score(s) = \sum_{p \in paths(s)} \prod_{(u,v) \in p} w(u, v), \tag{10}$$

[22] In our experiments we always take σ to be a single sentence – thus our algorithms always operate on a sentence-by-sentence basis.

Table 16
Performance on the SemEval-2007 task 07: coarse-grained English all-words WSD. Best results for each of the adopted measures – recall (R), precision (P) and balanced F-measure (F_1) – are in bold. Thanks to additional semantic relations from Wikipedia, BabelNet is able to outperform WordNet both when evaluating its performance on disambiguating nouns only and across all parts of speech (all differences in recall between BabelNet and WordNet are statistically significant according to a χ^2 test, $p < 0.01$).

Resource	Algorithm	Nouns only			All words		
		P	R	F_1	P	R	F_1
WordNet	Degree	81.1	67.3	73.6	**79.6**	61.0	69.1
	PLength	81.7	67.9	74.2	78.9	60.6	68.5
	SProbability	79.1	65.7	71.8	77.7	59.6	67.4
	PageRank	80.5	66.5	72.9	79.1	56.2	65.7
BabelNet	Degree	**83.3**	**81.7**	**82.5**	79.4	74.8	77.1
	PLength	82.8	81.1	82.0	77.8	73.3	75.5
	SProbability	82.0	80.3	81.1	77.6	73.2	75.3
	PageRank	81.6	79.9	80.7	78.5	67.6	72.6
	MFS BL	77.4	77.4	77.4	78.9	**78.9**	**78.9**
	Random BL	63.5	63.5	63.5	62.7	62.7	62.7

where the weight $w(u, v)$ is given by the weight of edge (u, v) in the knowledge base (see Section 3.3 for details on the edge weighting process) normalized as follows. We first consider for each node v its set of outgoing edges corresponding to semantic relations from WordNet and Wikipedia separately:

$$E_{out}(v) = E_{out}^{\text{Wiki}}(v) \cup E_{out}^{\text{WN}}(v). \tag{11}$$

The set of edges from each single resource is taken to define a probability distribution over the possible adjacent nodes that can be reached from v in WordNet or Wikipedia. Thus, the weights of the edges of each resource are appropriately normalized as:

$$\sum_{e \in E_{out}^{\text{Wiki}}} w(e) = 1, \quad \text{and} \quad \sum_{e \in E_{out}^{\text{WN}}} w(e) = 1. \tag{12}$$

Finally, we interpolate the contribution from WordNet and Wikipedia so as to determine the degree of importance of each resource in determining the final weight. The two models are combined by linear interpolation, where each weight is rescaled by a factor λ such that:

$$\lambda \sum_{e \in E_{out}^{\text{Wiki}}} w(e) + (1 - \lambda) \sum_{e \in E_{out}^{\text{WN}}} w(e) = 1. \tag{13}$$

- **PageRank:** The fourth method we use scores the various word senses with a global algorithm based on the topology of the context graph in its entirety. We follow previous work [76,73] and apply PageRank to induce the ranking of the senses. Given our context graph connecting the senses of the words in the input sequence, we first apply traditional PageRank [14] over the graph and then, for each word in context, output its sense with the highest PageRank score.

6.2.4. Parameter tuning

The construction phase of our context graphs includes a series of free parameters which can affect the performance of our algorithms. These include: (1) the maximum length of a path connecting senses of different words in context; (2) the minimum weight of an edge connecting two senses in the context graph; (3) the value of λ used in Eq. (13) to interpolate the weights of the semantic relations from WordNet and Wikipedia. To ensure generality, we use the SemCor corpus [79] as held-out development set. To tune the parameters, we performed experiments with WordNet and BabelNet to maximize the F_1 measure on the dataset and found the best results to be given by limiting the search to a maximum depth of 3, retaining only those edges with a minimum weight of 0.01, and setting λ to 0.9.

6.2.5. Results and discussion

We report our results in terms of precision, recall and F_1-measure on the SemEval-2007 coarse-grained all-words dataset [90]. In Table 16 we show the results when evaluating WordNet and BabelNet on the nouns-only subset (1,108) and the full dataset with all words (2,269 instances).[23] In line with common practice, we compare with random sense assignment and the most frequent sense (MFS) from SemCor as baselines. BabelNet consistently performs better than WordNet both on the nouns-only dataset and across all parts of speech (all differences are statistically significant based on a χ^2 test,

[23] We leave out the evaluation using only BabelNet's subset of Wikipedia relations since we are primarily interested in evaluating the resource 'as is', namely in its entirety. All experiments using Wikipedia-only relations revealed a performance lower than that of using BabelNet, thus indicating, again, the benefits of complementing taxonomic relations (WordNet) with associative relations (Wikipedia), and supporting the performance trends observed in Section 6.1, as well as the previous analysis of BabelNet's English-only subset [106].

Table 17
Performance on SemEval-2007 coarse-grained all-words WSD with MFS as a back-off strategy when no sense assignment is attempted. The differences between the results in bold in each column of the table are not statistically significant at $p < 0.05$ based on a χ^2 test.

Resource	Algorithm	Nouns only P/R/F$_1$	All words P/R/F$_1$
WordNet	Degree	80.1	79.7
	PLength	80.3	79.8
	SProbability	79.5	79.3
	PageRank	79.7	79.4
BabelNet	Degree	**84.7**	**82.3**
	PLength	**85.4**	**82.7**
	SProbability	**84.6**	**82.1**
	PageRank	82.1	80.1
	SUSSX-FR	81.1	77.0
	TreeMatch	N/A	73.6
	NUS-PT	82.3	**82.5**
	SSI	**84.1**	**83.2**
	MFS BL	77.4	78.9
	Random BL	63.5	62.7

$p < 0.01$). Improvements are given by a higher recall, thanks to the enriched structure of BabelNet: exploiting encyclopedic relations from Wikipedia and complementing them with those from WordNet yields, in fact, an improvement in recall of up to +14.6% for nouns (SProbability) and +13.8% for all parts of speech (Degree): this improvement, coupled with small variations in the precision rate, yields an overall improvement on the F$_1$-measure of up to +9.3% for nouns (SProbability) and +8.0% for all parts of speech (Degree). Results for different methods using the same knowledge resource are not statistically significantly different, thus highlighting the fact that considerable improvements in knowledge-based WSD can be achieved by means of enriching existing semantic networks with high-quality relations, regardless of the method used to obtain these relations – i.e., from a complementary perspective, graph-based methods will not necessarily achieve a competitive performance unless they are fed with high-quality and wide-coverage structured knowledge.

The improvements given by BabelNet are smaller when evaluating on the entire dataset, rather than its noun-only subset: this is because, when comparing with WordNet, performance on verbs and adverbs tends be lower, due to the enriched network amplifying the bias of the connectivity measures towards verb senses which are (directly or indirectly) connected to the noun hierarchy. Nevertheless, in this case, too, BabelNet outperforms WordNet by a statistically significant margin, again thanks to improvements in recall with small decreases in the precision rate as a trade-off. Finally, using BabelNet enables us to beat the MFS baseline on nouns, which is a notably difficult competitor for unsupervised and knowledge-rich systems: we are not able to achieve the same result when evaluating on all words, due to the much lower recall deriving from the limited connectivity of parts of speech other than nouns in WordNet.

To further investigate the performance of BabelNet, we follow [89] and run our algorithms in a weakly supervised setting where the WSD system attempts no sense assignment if the highest score among those assigned to the senses of a target word is below a certain (empirically estimated) threshold. In this setting, in order to disambiguate all instances, we use the MFS as a back-off strategy: that is, the system falls back to assigning to the target word in context its most frequent sense from SemCor. Similarly to all other parameters, the optimal value for this threshold is estimated by maximizing the F$_1$ of each measure on our development set: given the scores for all instances, these are normalized in the $[0, 1]$ interval and the optimal value is found by evaluating performance at each step by incrementing the threshold value by 0.01. Finally, in order to benchmark BabelNet not only against WordNet, but also against other state-of-the-art systems, we compare it with the best unsupervised [59] (SUSSX-FR) and supervised [18] (NUS-PT) systems participating in the SemEval-2007 coarse-grained all-words task. We also compare with Structural Semantic Interconnections [94] (SSI) – a knowledge-based system that participated out of competition – and the unsupervised proposal from [19] (TreeMatch).

Table 17 shows the results of our algorithms in the weakly-supervised setting. BabelNet achieves a competitive performance both on the entire dataset and its subset containing only nominal target instances. On this latter subset, its performance is comparable with SSI and significantly better than the best supervised and unsupervised systems (+3.1% and +4.3% F$_1$ against NUS-PT and SUSSX-FR). On the entire dataset, it outperforms SUSSX-FR and TreeMatch (+5.7% and +9.1%) and its recall is not statistically different from that of SSI and NUS-PT. This result is particularly interesting, given that BabelNet is extended only with relations between nominals, and, in contrast to SSI, it does not rely on a costly human effort to validate the set of semantic relations. Last but not least, we achieve state-of-the-art performance with a battery of simpler algorithms that are based on the notion of connectivity in the context graph.[24]

[24] Small performance differences with WordNet++ [106] (e.g., −0.1% R/P/F$_1$ on the nouns-only dataset) are due to a different weighting scheme and are not statistically significant. WordNet++ uses in fact a filter to rule out weak semantic relations from Wikipedia by computing, for a given pair of Wikipages, their semantic strength as the degree of overlap between the two bags of words built from the labels of their Wikipedia categories. BabelNet, instead, relies

6.3. Cross-lingual Word Sense Disambiguation

BabelNet is a multilingual knowledge resource, in that it provides a semantic network where related concepts are connected within a graph structure and whose lexicalizations are available for different languages. Given these distinguishing features, BabelNet is a powerful resource for performing knowledge-based lexical disambiguation in a multilingual setting. Accordingly, we performed a last batch of experiments on the SemEval-2010 cross-lingual Word Sense Disambiguation task [60] (CL-WSD henceforth). In CL-WSD, lexical disambiguation is operationalized as a word translation task. Given a word in context (i.e., an English polysemous noun), the system disambiguates the target word by translating it into a different language: the translation is considered to be correct if it preserves the meaning of the word in context in the source language. By combining Machine Translation and WSD into a hybrid task, CL-WSD is able to overcome the granularity issue affecting predefined sense inventories [87] by assuming that all sense distinctions are given by the translations available in a multilingual corpus.

To disambiguate an English word in context we use the same framework as that used in monolingual WSD (cf. Section 6.2). Given a target word, we first build a disambiguation graph by connecting all Babel synsets corresponding to the different senses of the word to all the senses of other context words. Once again, the disambiguation graph is built by exploring the BabelNet graph using a Depth-First Search up to a maximum depth of 3. Next, in order to rank the different senses, and similarly to our monolingual setting, we apply the same set of measures based on paths (PLength, SProbability), node connectivity (Degree) and global graph topology (PageRank). Once the most likely sense of the target word has been established, we proceed to output its translation. To do so we collect all lexicalizations found in the corresponding Babel synset and select those which are also found in the CL-WSD sense inventory: this consists of a mapping from English words to their translations, as found in the word alignments obtained by applying GIZA++ [101] to the 1–1 sentence alignments from the Europarl corpus [56].

6.3.1. Experimental setting

We evaluate on the SemEval-2010 CL-WSD dataset [60], which consists of 1,000 test instances (50 sentences for each of the 20 target words). The evaluation scheme is based on the SemEval-2007 English lexical substitution task [68] and consists of an adaptation of the standard metrics of precision and recall for the translation setting. Given a test instance, the ground truth consists of a list of translations provided by different human judges, each weighted by the number of raters who judged it to be correct. For instance, given an occurrence of the sense of the English word coach as trainer in the following sentence:

(c) Strangely, the national **coach** of the Irish teams down the years has had little direct contact with the four provincial coaches

the gold-standard translations in the different languages are the following ones (the number beside each word counts the translators who selected that word as an appropriate translation):

Dutch: coach 3; speler-trainer 1; trainer 3; voetbaltrainer 1;
French: capitaine 1; entraîneur 3;
German: Coach 1; Fußballtrainer 1; Nationaltrainer 2; Trainer 3;
Italian: allenatore 3;
Spanish: entrenador 3.

Precision then computes the number of correct system translations over the total number of translations returned by the system, each weighted by the number of annotators who selected the translation as the correct one. Recall instead computes the number of correct translations given by the system over the total number of items in the test set, where each item is again weighted by the number of raters who deemed it correct.

Given that a single Babel synset can contain multiple translations for the same English word in a target language, we explore for this task an unsupervised setting where we return for each test instance only the most frequent translation found in the BabelNet sense inventory, as ordered by frequency of alignment in Europarl. To provide an answer for all instances, we return this most frequent translation even when no sense assignment is attempted – i.e., no sense of the target word is connected to any other sense of the context words – or a tie occurs. Our knowledge-based approach to the CL-WSD task involves two major steps: first, given an English target word in context, we disambiguate it to the highest-ranking Babel synset; next, given the multilingual translations found in the selected Babel synset, we return the appropriate lexicalization(s) in the language of interest. Thus, it could happen that our two-stage method selects a wrong concept for the target word but still outputs a correct translation or, vice versa, that it returns an incorrect or unforeseen translation, even if the correct Babel synset was found. Consequently, to gain a better insight into the performance of BabelNet when applied to the CL-WSD task, we benchmark it in two different settings:

on a weighting scheme based on the Dice coefficient (Section 3.3): this represents a more general solution that does not require a knowledge resource to contain categorized concepts, and thus can be applied to both WordNet and Wikipedia relations.

Table 18
Results on the SemEval-2010 task 3: Cross-lingual Word Sense Disambiguation.

	French		German		Italian		Spanish	
	P	R	P	R	P	R	P	R
Baseline	21.25	21.25	13.16	13.16	15.18	15.18	19.74	19.74
UvT-v	–	–	–	–	–	–	23.39	23.39
UvT-g	–	–	–	–	–	–	19.83	19.64
T3-COLEUR	21.97	21.75	13.18	13.05	14.82	14.67	19.83	19.64
Degree	22.94	22.94	17.15	17.15	18.03	18.03	22.48	22.48
+ ORACLE TRANSLATIONS	25.82	25.82	20.16	20.16	21.13	21.13	25.26	25.26
PLength	23.42	23.42	17.72	17.72	18.19	18.19	22.76	22.76
+ ORACLE TRANSLATIONS	25.87	25.87	20.42	20.42	21.47	21.47	25.76	25.76
SProbability	23.27	23.27	17.61	17.61	18.14	18.14	22.69	22.69
+ ORACLE TRANSLATIONS	25.85	25.85	20.50	20.50	21.74	21.74	25.48	25.48
PageRank	22.62	22.62	16.98	16.98	16.76	16.76	21.11	21.11
+ ORACLE TRANSLATIONS	26.00	26.00	20.85	20.85	21.71	21.71	26.19	26.19
BabelNet upper bound	30.21	30.21	25.39	25.39	27.67	27.67	30.73	30.73
Task upper bound	39.44	100.00	34.36	100.00	40.00	100.00	39.54	100.00

- **Standard setting:** we return the most frequent translation from the highest-scoring synset. In the case that no sense assignment is attempted (i.e., no Babel synset of the target word is connected to any other sense of the context words, or a tie occurs), the system returns the most frequent word alignment found in the Europarl corpus.
- +**ORACLE TRANSLATIONS:** we start with the output of the standard setting and, for each instance, we remove from the set of translations of the highest scoring Babel synset all those which are not found in the gold-standard annotation. We then return the most frequent translation from this set of 'clean' lexicalizations and back-off to the most frequent Europarl alignment if no such translation is available as a result of this filtering process.

In addition, we computed the following two upper bounds:

- **BabelNet upper bound:** we output for each test instance the largest subset of gold-standard translations found among the Babel synsets containing the test instance. Once again, we fill the missing sense assignment by backing off to the Europarl most frequent translations. This upper bound quantifies how well we can aim at performing knowledge-based CL-WSD using BabelNet.
- **Task upper bound:** since the CL-WSD evaluation metrics do not represent percentages (due to the variability of less credit being given to those items where annotators express differences), we compute the task upper bound by providing as answer for each test item the most frequent translation among those chosen by the human annotators.

We compare the performance of BabelNet with the best unsupervised [45] (T3-COLEUR) and supervised [44] (UvT-v and UvT-g) proposals which participated in the SemEval CL-WSD competition. In our experiments performance is evaluated using a modified version of the official scorer, which includes bug fixes and computes precision and recall on the *entire* dataset, rather than calculating the average across all target words (in line with the scoring criteria for other SemEval WSD evaluations, including the original lexical substitution task). To make the comparison with other systems fair, we re-evaluated the other systems' output using our modified scorer.[25]

6.3.2. Results and discussion

Table 18 presents the results of our approach on the CL-WSD task, where we evaluate our systems on the French, German, Italian and Spanish translations.[26] The results indicate that using BabelNet 'as-is' already outperforms the baseline by a large margin, as well as both unsupervised (T3-COLEUR) and supervised (UvT-g) systems. While all algorithms again exhibit comparable performance, similarly to the monolingual WSD scenario, PLength yields the best overall results, followed by SProbability, Degree and PageRank in turn. As a result of this we perform better than any other system that participated in the competition, except for the UvT-v system from [44], i.e., an ensemble architecture (available only for Spanish and Dutch) which combines different supervised classifiers using local and global context features within a voting architecture.

In the 'ORACLE TRANSLATIONS' setting, filtering the output of our system to retain only the gold-standard translations additionally improves the results by removing wrong translations found in BabelNet: this setting evaluates the performance of the disambiguation component of the CL-WSD system and shows that further improvements can be achieved by im-

[25] This modified scorer is available at http://lcl.uniroma1.it/babelnet/clwsd-scorer.pl. Bug fixes were promptly submitted to and endorsed by the task organizers.

[26] We leave out the CL-WSD evaluation on Dutch, since this language is not covered by the current version of BabelNet.

proving the quality of the lexicalizations found within the Babel synsets. Finally, the BabelNet upper bound configuration indicates that an additional performance boost can be obtained by improving the disambiguation component and selecting better translations as output: we take this to be good news, since it implies that knowledge-based CL-WSD is, indeed, a framework capable of achieving very high performance.

7. Related work

In this section we review and contextualize our work within the existing body of literature. Our contribution lies in three main areas, namely, multilingual knowledge acquisition, information extraction from collaborative knowledge resources, and monolingual and multilingual Word Sense Disambiguation. We accordingly present an overview of related contributions in each of these areas.

7.1. Multilingual knowledge acquisition

Pioneering efforts to provide machine readable lexical knowledge for English such as WordNet [37], Cyc [63] and SUMO [100] first concentrated on manually building semantic networks. Similarly, initial attempts to build multilingual resources were manual, and led to the creation of a multitude of wordnets[27] such as EuroWordNet [128], MultiWordNet [104], BalkaNet [127], Arabic WordNet [13], the Multilingual Central Repository [7], bilingual electronic dictionaries such as EDR [139], and full-fledged frameworks for the development of multilingual lexicons [64]. Community efforts in this direction have led, over recent years, to the creation of institutions such as the Global WordNet Association[28] and the Global WordNet Grid [102], which aim at building a collection of concepts shared among different wordnets, as well as other projects aimed at collecting multilingual textual resources on a very large scale [1]. As is often the case with manually assembled resources, all these lexical knowledge repositories have a well-defined structure: in addition, they encode high-quality expert knowledge such as fine-grained sense distinctions (WordNet) or logical axioms like rules and other assertions (Cyc). Thus, these resources concentrate on deep encoding of semantic information, a task which is arguably beyond the scope and kind of expertise found in collaborative resources: however, as a downside, they are typically hindered by high development costs and insufficient coverage. This barrier has led to proposals that acquire multilingual lexicons from parallel text [42,39,70, *inter alia*], monolingual corpora [58,119,46] or machine readable dictionaries (MRDs) [25]. Other proposals include the creation of a semantic network based on the graph-based disambiguation of glosses from a bilingual MRD [38], as well as a method based on the combination of a pre-existing lexical knowledge base with bilingual mappings from a MRD [36]. Recent proposals like TransGraph [33] and PanDictionary [67] present graph-based methods for producing massive multilingual translation dictionaries from Web resources such as online lexicons and wiktionaries, and have been shown to be robust enough to improve search applications [22]. However, while providing lexical coverage on a very large scale for hundreds of thousands of language pairs, these resources do not encode semantic relations between concepts denoted by their lexical entries. In this paper, we focus instead on creating a wide-coverage semantic network where millions of concepts lexicalized in different languages are connected by a multitude of semantic relations.

7.2. Information extraction and integration from collaborative knowledge resources

Over the last few decades a large body of work has been published concerning the development of methods for automatically harvesting knowledge and enriching existing resources such as WordNet. These include proposals to extract semantic information from dictionaries [21,38,115, *inter alia*], approaches using lexico-syntactic patterns [49,23,43,96], heuristic methods based on lexical and semantic regularities [47], taxonomy-based ontologization [96,103,122]. Other work relies on the disambiguation of collocations, either obtained from specialized learner's dictionaries [85] or from topic signatures from the Web [28]. State-of-the-art proposals such as KnowItAll [32], KnowNet [28], TextRunnner [9,30] and TaxoLearn [96] all aim at acquiring repositories of knowledge from the Web with minimal or no supervision. However, as a trade-off, they either rely on a limited set of predefined semantic relations (KnowItAll and TaxoLearn) or they do not provide a fully ontologized resource as output (KnowItAll and TextRunner), or they contain few named entities (KnowNet).

In the attempt to stake out a middle ground between entirely manual and entirely automatic approaches, the last few years have seen a growing interest in using collaborative contributions from volunteers for knowledge acquisition [113,20, 6]. In particular, many research efforts concentrated on extracting large-scale repositories of knowledge from Wikipedia, which, due to its low entrance barrier and vast user base, provides large amounts of information at practically no cost. Previous work aimed at transforming its content into a knowledge base includes open-domain relation extraction [81,134, 136], the acquisition of taxonomic [135,108] and other semantic relations [83,34], as well as full-fledged ontologies [124, 12]. Applications using the knowledge contained in Wikipedia include, among others, text categorization [40], computing semantic similarity of texts [40,107,80], coreference resolution [107], multi-document summarization [82], text generation [120] and simplification [132], and clustering web search queries [133]. However, little attention has so far been given to

[27] Cf. the list available at http://www.globalwordnet.org/gwa/wordnet_table.htm.

[28] http://www.globalwordnet.org.

Table 19
Comparison of the lexical knowledge contained in BabelNet with WikiNet and UWN/MENTA.

Resource		Lemmas	Concepts	Word senses
UWN		822,212	117,659	1,595,763
MENTA	upper-level	837,627	82,115	845,210
	full	–	5,379,832	–
WikiNet		11,721,594	3,707,718	14,200,945
BabelNet		23,935,611	3,032,406	26,044,643

exploiting the multilingual properties of Wikipedia, notable exceptions being its use for the automatic induction of parallel corpora [2], named entity extraction [131] and recognition [114], cross-language information retrieval [24] and multilingual information extraction [3]. Overall, extracting and using knowledge from Wikipedia has been so successful as to be used to boost, along with other knowledge sources, commercial projects such as Freebase,[29] a very large collaborative knowledge base, and Trueknowledge,[30] a semantic answer engine.

BabelNet integrates semi-structured information from Wikipedia with the relational structure of WordNet into a unified, wide-coverage, multilingual semantic network. We achieve this by developing a mapping algorithm that can be applied to both Wikipedia pages and categories with high accuracy. Previous attempts at linking WordNet with Wikipedia include a manual mapping of Wikipedia entries to WordNet concepts [12], a model based on vector spaces [117], automatically linking Wikipedia categories to WordNet based on structural information [105], as well as associating them with the most frequent WordNet sense heuristic [124]. Finally, UWN/MENTA [71,72] uses this heuristic as a feature of a supervised linker: using a supervised method for resource mapping has the advantage of yielding competitive results at the cost of not being applicable to arbitrary resources where no manual mapping is available. Our proposal, in contrast, can be applied to other resources for additional resource integration and enrichment (see, e.g., our experiments on mapping WordNet to Wikipedia categories presented in Section 4.1.2).

7.2.1. Comparison with WikiNet and UWN/MENTA

The research closest to ours is WikiNet [84] and UWN/MENTA [71,72]. Nastase et al. [84] present WikiNet, a multilingual semantic network built from Wikipedia and including semantic relations between Wikipedia entities which are collected from the category network, infoboxes and article bodies. De Melo and Weikum [71] develop a Universal WordNet (UWN) by automatically acquiring a semantic network for languages other than English. UWN is bootstrapped from WordNet and is built by collecting evidence extracted from existing wordnets, translation dictionaries, and parallel corpora. The result is a graph containing more than 800,000 words from over 200 languages in a hierarchically structured semantic network with over 1.5 million links from words to word senses. The same authors later present in [72] a methodology for building MENTA, a multilingual taxonomy containing 5.4 million entities, which is also built from WordNet and Wikipedia. Both UWN/MENTA and WikiNet have been developed in parallel with BabelNet and offer complementary sources of structured information – i.e., large amounts of facts about entities (UWN/MENTA, based on its integration with YAGO2 [50]), and explicit semantic relations between concepts denoted by Wikipedia categories (WikiNet). Since these two contributions are strongly related to our work, we present in the following a comparative analysis of all three resources.

Table 19 compares all resources in terms of the amount of lexical and conceptual knowledge they contain – i.e., their number of lemmas, concepts and senses. For WikiNet we consider only the subset of languages also contained in BabelNet, whereas in the case of MENTA we report figures for both the overall resource and its multilingually lexicalized upper-level (we include only publicly available figures from de Melo and Weikum [71,72]). The figures indicate that the lexical knowledge contained in our resource compares favorably in size with both WikiNet and UWN/MENTA. Both WikiNet and MENTA have a higher number of concepts: however, this is due either to their using a more recent Wikipedia version (January 2011 vs. November 2009) and also including concepts and lexicalizations from Wikipedia's category system (WikiNet), or to their collecting entities and concepts from wikipedias other than the English one (MENTA). However, the lexicalizations found in these two resources are simply taken directly from Wikipedia inter-language links and no effort is made to address the issue of translation gaps, as we, in contrast, do: as a result, their number of lemmas and word senses is far lower than ours.[31]

Next, we present in Table 20 a more in-depth comparative analysis between BabelNet and WikiNet, since the latter is freely available for download. To this end, we quantify the size of their intersection, i.e., the set of lemmas shared by both resources, and also compute the measures of coverage and extra coverage (introduced in Section 4.2.3) between them, this time at the level of lemmas. Thus, in Table 20, "coverage" quantifies the proportion of lemmas in the intersection which are also found in the reference resource (i.e., WikiNet in our case), and "extra coverage" is, instead, the ratio of lemmas found only in BabelNet to WikiNet's lemmas. The results show that BabelNet covers on average more than 70% of the terms

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

[30] http://www.trueknowledge.com.

[31] Once again we leave out the analysis of the full set of 5.4 million entities of MENTA since no statistics are available for the lexical knowledge they contain, namely their number of lexicalizations and senses.

Table 20
Comparison of BabelNet with WikiNet: size of the intersection, term coverage and extra coverage.

Language	|WikiNet ∩ BabelNet|	Coverage	ExtraCoverage
English	5,314,006	72.8%	8.9%
Catalan	135,378	66.4%	1641.0%
French	848,765	77.2%	243.5%
German	874,272	78.5%	236.6%
Italian	595,266	75.7%	359.3%
Spanish	358,422	70.1%	623.2%

from WikiNet: the fact that we do not achieve full coverage is due to using different Wikipedia versions,[32] as well as to WikiNet including Wikipedia categories, whose labels typically consist of complex phrases (e.g., SCIENTISTS WHO COMMITTED SUICIDE), which are difficult for us to cover. In contrast, the results indicate that we are able to achieve a very high extra coverage, thanks to our additional lexicalizations obtained from the output of the MT system. The smaller the set of terms found in WikiNet for a language (e.g., for Catalan and Spanish), the more beneficial our approach and the higher the extra coverage: as a result, we are able to provide comparable coverage across *all* languages, with no substantial differences between resource-rich and resource-poor languages.

Similarly to WikiNet and UWN/MENTA, our proposal brings together lexicographic and encyclopedic knowledge in many languages, thus providing a multilingual encyclopedic dictionary. However, our focus is not only on providing *high-quality conceptual knowledge* (contained in the backbone provided by WordNet and Wikipedia), but also on integrating it with *high-coverage lexical knowledge for all languages*. Crucially, we argue that in order to achieve this one cannot simply rely on Wikipedia inter-language links, since these have limited coverage – especially in the case of resource-poor languages (e.g., there is no translation for PLAY (THEATRE) in Basque or Hungarian). BabelNet provides a solution to this problem by filling translation gaps (i.e., missing translations) by means of Statistical Machine Translation and providing, as a result, a large sense-annotated corpus for hundreds of thousands of concepts. Further advancements can be achieved by integrating all resources and thus bringing together all their strengths, including large amounts of labeled semantic relations (WikiNet), theoretically sound taxonomy induction algorithms (UWN/MENTA), and vast amounts of lexical knowledge (BabelNet). However, in this paper we go one step beyond other proposals in terms of methodology. Indeed, this paper presents the first contribution showing that a very large multilingual lexical knowledge base can be used to achieve state-of-the-art performance on several lexical disambiguation tasks in both monolingual and cross-lingual settings, as opposed to the small-scale manual intrinsic evaluations used to benchmark WikiNet and UWN/MENTA.[33]

7.3. Monolingual and multilingual Word Sense Disambiguation

Lexical knowledge also lies at the core of Word Sense Disambiguation (WSD), the task of computationally identifying the meanings of words in context [87]. In order to achieve high performance supervised approaches to WSD require large training sets where instances (i.e., target words in context) are manually annotated with the most appropriate word senses. Producing this kind of knowledge is extremely costly: at a throughput of one sense annotation per minute [31] and tagging one thousand examples per word, dozens of person-years would be required for a supervised classifier to disambiguate all the words in the English lexicon with high accuracy. In contrast, knowledge-based approaches exploit the information contained in wide-coverage lexical resources, such as WordNet. However, it has been demonstrated that the amount of lexical and semantic information contained in such resources is typically insufficient for high-performance WSD [26,89]. In other words, WSD systems have to face the well-known issue of the knowledge acquisition bottleneck. Several methods have been proposed for automatically extending existing resources and it has been shown that highly-interconnected semantic networks have a great impact on WSD [89,94]. However, to date, the real potential of knowledge-rich WSD systems has been demonstrated only in combination with either a large semi-supervised extension of WordNet [85] or sophisticated algorithms [4]. In contrast, we have shown how simple knowledge-based algorithms can equal and surpass current state-of-the-art performance in monolingual WSD when they are provided with a rich set of concepts and semantic relations, such as those in BabelNet.

The knowledge acquisition bottleneck problem mentioned above makes it difficult to perform accurate WSD in non-English languages, due to the lack of rich and wide-coverage knowledge resources for most languages. A solution to this issue is to use bilingual corpora for the creation of sense inventories. The idea underlying this approach is that the plausible translations of a word in context restrict its possible senses to a manageable subset of meanings [112]. Parallel corpora have been used for the automatic creation of a sense-tagged dataset for supervised WSD [41]. Other approaches include the use of a coherence index for identifying the tendency to lexicalize senses differently across languages [52] and the clustering

[32] This effect is, in fact, larger for smaller wikipedias such as the Catalan one, since growth is known to be quasi-exponential, cf. http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth.

[33] UWN [71] is indeed extrinsically evaluated on computing semantic relatedness in German and cross-lingual text classification. However, these two evaluations look more like case studies, since they either use small datasets (semantic relatedness) or are not compared to other state-of-the-art approaches from the literature.

of source words which translate into the same target word, then used to perform WSD using a similarity measure [29]. A historical approach [15] consists of the use of bilingual corpora to perform unsupervised word alignment and determine the most appropriate translation for a target word according to the most informative feature from a set of contextual features. This approach has recently been revamped by proposing the use of monolingual local context and bilingual information from aligned translations as features for a supervised word translator in a cross-lingual WSD setting [61].

All the above approaches to multilingual or cross-lingual WSD rely on bilingual corpora, including those which exploit existing multilingual WordNet-like resources [53] or use automatically induced multilingual co-occurrence graphs [121]. However, yet again, this requirement is often very hard to satisfy, especially if we need wide coverage, as expected in real-world applications. Our work on BabelNet effectively attacks the knowledge acquisition bottleneck by providing an unprecedented lexical coverage of non-English languages (cf. Section 4.2). As a result, state-of-the-art WSD is achieved in a cross-lingual setting. As shown in Section 6.3, BabelNet is built following the same design principles as other multilingual lexical resources like EuroWordNet [128] and the Multilingual Central Repository [7], which also keep their (language-independent) conceptual core separated from the multilingual lexicalizations of their concepts. As a result, BabelNet makes the very same rich semantic network available for all those languages whose lexicalizations are found in the Babel synsets. Thus we would argue that the availability of such high-quality knowledge for all of its languages opens up the possibility of high-performing systems for non-English monolingual WSD as well (we leave this for future research).

8. Conclusions

In this paper we have presented BabelNet, a wide-coverage multilingual knowledge resource obtained by means of a novel automatic construction methodology. Key to our approach is a two-tier methodology, namely: a) a high performing method to produce a mapping between a multilingual encyclopedic knowledge repository (Wikipedia) and a computational lexicon of English (WordNet); b) the use of a state-of-the-art machine translation system to collect a very large amount of multilingual concept lexicalizations, and complement Wikipedia's manually-edited translations. In order to robustly map WordNet with Wikipedia, we investigated different methods for estimating the likelihood of links between these two resources, namely a bag-of-words and a graph-based mapping algorithm. To achieve the best translation performance, we were happy to rely on recent advances in machine translation by using Google's online translation system.[34] Each of these two steps has several advantages. Firstly, the integration process allows the two knowledge resources to contribute different kinds of lexical knowledge, one concerned mostly with named entities, the other with concepts. BabelNet brings together the strengths of WordNet – i.e., its being highly structured and providing labeled lexico-semantic relations – with those of Wikipedia – i.e., providing large amounts of semantic relations, multilinguality and continuous collaborative updates. Thus, even when they overlap, the two resources provide complementary information about the same named entities or concepts. Second, automatically translating a large corpus of sense occurrences from Wikipedia and SemCor enables us to complement the high-quality human translations provided by Wikipedia with automatically generated ones. This way we are able to collect missing translations and automatically fill in the gap between resource-rich languages – such as English – and resource-poor ones. As a result, BabelNet is able to achieve a wide coverage, that is, our Babel synsets contain lexicalizations for most of the covered languages.

Our experiments show that our fully-automated approach produces a large-scale lexical resource with high accuracy. We evaluated the mapping of both Wikipedia pages and categories with manually labeled gold standards. The better results achieved by the graph-based algorithm permits us to establish that exploiting the structure of the target resource boosts the performance on the mapping task. Mapping Wikipedia categories provides, in turn, an indication of the wider applicability of our algorithm. The resource we obtain as a result of the application of our methodology includes millions of semantic relations, mainly from Wikipedia (however, WordNet relations are labeled), and contains more than 3 million concepts (8.6 labels per concept on average). As pointed out in Section 4.2, such coverage is much wider than that of existing wordnets in non-English languages. While BabelNet currently includes 6 languages, links to freely-available wordnets[35] can immediately be established by utilizing the English WordNet as an interlanguage index. Indeed, BabelNet can be extended to virtually any language of interest, provided that language is covered by a Machine Translation system. A thorough extrinsic evaluation of BabelNet shows that it enables state-of-the-art performance in monolingual and cross-lingual Word Sense Disambiguation, allowing even simple knowledge-based algorithms to compete with (and often outperform) supervised systems. All SemEval tasks we use for extrinsic evaluation are very competitive benchmarks, so achieving a state-of-the-art performance on three of them (as we consistently do throughout Section 6) indicates that we have, indeed, produced a wide-coverage resource containing large amounts of high-quality knowledge. These results strongly corroborate previous studies on the importance of high-quality, largely populated knowledge resources [28,89].

As future work we plan to apply our method to other languages, including Eastern European, Arabic, and Asian languages. We also intend to collect additional knowledge by exploring promising directions, namely linking missing concepts in WordNet by establishing their most likely hypernyms – e.g., along the lines of Snow et al. [122] and Navigli and Velardi

[34] This robust translation performance is counter-balanced by our MT system being a black box, due to its commercial nature. See [55] for a different, yet related discussion on using search engine counts for NLP tasks.

[35] http://www.globalwordnet.org.

[95] – and typing the topical, semantically unspecified relations from Wikipedia with explicit semantic relations (cf. previous work on the category system by [83] and [108] and recent work on relation synsets [81]). Finally, we plan to exploit the wide-coverage knowledge contained in BabelNet to enable robust multilingual processing for a variety of complex NLP tasks, such as cross-lingual summarization, question answering and information retrieval. In this light, we hope that our state-of-the-art results on monolingual and multilingual WSD represent just the point of departure, to be appropriately taken to the next level in the near future by enabling high-end, real-world multilingual applications.

Downloads

BabelNet and BabelCor are freely available at http://lcl.uniroma1.it/babelnet under a Creative Commons Attribution-Noncommercial-Share Alike License. A Java API [93] for programmatic access and multilingual WSD is available on the same Web site. BabelNetXplorer [92], a Web browser for BabelNet, is also available at http://lcl.uniroma1.it/bnxplorer.

Acknowledgements

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234. Thanks go to Google for access to the University Research Program for Google Translate. We thank Jim McManus for validating linguistic data and three anonymous AIJ reviewers for their detailed reviews and useful feedback.

References

[1] S. Abney, S. Bird, The human language project: Building a universal corpus of the world's languages, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 88–97.

[2] S.F. Adafre, M. de Rijke, Finding similar sentences across multiple languages in Wikipedia, in: Proceedings of the EACL-06 Workshop on New Text – Wikis and Blogs and Other Dynamic Text Sources, Trento, Italy, 4 April 2006.

[3] E. Adar, M. Skinner, D.S. Weld, Information arbitrage across multi-lingual Wikipedia, in: Proceedings of the Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain, 9–12 February 2009, pp. 94–103.

[4] E. Agirre, O.L. de Lacalle, A. Soroa, Knowledge-based WSD on specific domains: Performing better than generic supervised WSD, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, CA, 14–17 July 2009, pp. 1501–1506.

[5] E. Agirre, A. Soroa, Personalizing PageRank for Word Sense Disambiguation, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March–3 April 2009, pp. 33–41.

[6] L. von Ahn, M. Kedia, M. Blum, Verbosity: A game for collecting common-sense facts, in: Proceedings of the Conference on Human Factors in Computing Systems, Montréal, Québec, Canada, 22–27 April 2006, pp. 75–78.

[7] J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, P. Vossen, The MEANING multilingual central repository, in: Proceedings of the 2nd International Global WordNet Conference, Brno, Czech Republic, 20–23 January 2004, pp. 80–210.

[8] S. Banerjee, T. Pedersen, Extended gloss overlap as a measure of semantic relatedness, in: Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, 9–15 August 2003, pp. 805–810.

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

[10] A. Barrón-Cedeño, P. Rosso, E. Agirre, G. Labaka, Plagiarism detection across distant language pairs, in: Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010, pp. 37–45.

[11] J.A. Bateman, J. Hois, R. Ross, T. Tenbrink, A linguistic ontology of space for natural language processing, Artificial Intelligence 174 (2010) 1027–1071.

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

[13] W. Black, S.E.H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, Introducing the Arabic WordNet project, in: Proceedings of the 3rd International Global WordNet Conference, Jeju Island, South Korea, 22–26 January 2006 pp. 295–299.

[14] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems 30 (1998) 107–117.

[15] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, R.L. Mercer, Word-sense disambiguation using statistical methods, in: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, 18–21 June 1991, pp. 264–270.

[16] R. Bunescu, M. Paşca, Using encyclopedic knowledge for named entity disambiguation, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006, pp. 9–16.

[17] J. Carletta, Assessing agreement on classification tasks: The kappa statistic, Computational Linguistics 22 (1996) 249–254.

[18] Y.S. Chan, H.T. Ng, Z. Zhong, NUS-ML: Exploiting parallel texts for word sense disambiguation in the English all-words tasks, in: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007, pp. 253–256.

[19] P. Chen, W. Ding, C. Bowes, D. Brown, A fully unsupervised Word Sense Disambiguation method using dependency knowledge, in: Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, CO, 31 May–5 June 2009, pp. 28–36.

[20] T. Chklovski, R. Mihalcea, Building a sense tagged corpus with Open Mind Word Expert, in: Proceedings of the ACL-02 Workshop on WSD: Recent Successes and Future Directions, Philadelphia, PA, July 2002, pp. 116–122.

[21] M. Chodorow, R. Byrd, G.E. Heidorn, Extracting semantic hierarchies from a large on-line dictionary, in: Proceedings of the 23th Annual Meeting of the Association for Computational Linguistics, Chicago, IL, 8–12 July 1985, pp. 299–304.

[22] J. Christensen Mausam, O. Etzioni, A rose is a roos is a ruusu: Querying translations for web image search, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, 2–7 July 2009, pp. 193–196.

[23] P. Cimiano, S. Handschuh, S. Staab, Towards the self-annotating Web, in: Proceedings of the 13th World Wide Web Conference, New York, NY, 17–22 May 2004, pp. 462–471.

[24] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, S. Staab, Explicit vs. latent concept models for cross-language information retrieval, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, CA, 14–17 July 2009, pp. 1513–1518.

[25] A. Copestake, E.J. Briscoe, P. Vossen, A. Ageno, I. Castellón, R. Francesc, G. Rigau, H. Rodríguez, A. Samiotou, Acquisition of lexical translation relations from MRDs, Machine Translation: Special Issue on the Lexicon 9 (1995) 33–69.

[26] M. Cuadros, G. Rigau, Quality assessment of large scale knowledge resources, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006, pp. 534–541.

[27] M. Cuadros, G. Rigau, Semeval-2007 task 16: Evaluation of wide coverage knowledge resources, in: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007, pp. 81–86.

[28] M. Cuadros, G. Rigau, KnowNet: building a large net of knowledge from the Web, in: Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK, 18–22 August 2008, pp. 161–168.

[29] M. Diab, Word Sense disambiguation within a multilingual framework, PhD thesis, University of Maryland, College Park, Maryland, 2003.

[30] D. Downey, O. Etzioni, S. Soderland, Analysis of a probabilistic model of redundancy in unsupervised information extraction, Artificial Intelligence 174 (2010) 726–748.

[31] P. Edmonds, Designing a task for SENSEVAL-2, Technical report, University of Brighton, UK, 2000.

[32] O. Etzioni, M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, S. Soderland, D. Weld, A. Yates, Unsupervised named-entity extraction from the Web: An experimental study, Artificial Intelligence 165 (2005) 91–134.

[33] O. Etzioni, K. Reiter, S. Soderland, M. Sammer, Lexical translation with application to image search on the Web, in: Proceedings of Machine Translation Summit XI.

[34] A. Fader, S. Soderland, O. Etzioni, Identifying relations for Open Information Extraction, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), 27–31 July 2011, Edinburgh, UK, pp. 1535–1545.

[35] S. Faralli, R. Navigli, A new minimally-supervised framework for domain Word Sense Disambiguation, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju, South Korea, 2012, pp. 1411–1422.

[36] J. Farreres, K. Gibert, H. Rodríguez, C. Pluempitiwiriyawe, Inference of lexical ontologies. The LeOnI methodology, Artificial Intelligence 174 (2010) 1–19.

[37] C. Fellbaum (Ed.), WordNet: An Electronic Database, MIT Press, Cambridge, MA, 1998.

[38] T. Flati, R. Navigli, The CQC algorithm: Cycling in graphs to semantically enrich and enhance a bilingual dictionary, Journal of Artificial Intelligence Research (JAIR) 43 (2012) 135–171.

[39] P. Fung, A pattern matching method for finding noun and proper noun translations from noisy parallel corpora, in: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, 26–30 June 1995, pp. 236–243.

[40] E. Gabrilovich, S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, Journal of Artificial Intelligence Research (JAIR) 34 (2009) 443–498.

[41] W.A. Gale, K. Church, D. Yarowsky, Using bilingual materials to develop Word Sense Disambiguation methods, in: Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada, 25–27 June 1992, pp. 101–112.

[42] W.A. Gale, K.W. Church, A program for aligning sentences in bilingual corpora, Computational Linguistics 19 (1993) 75–102.

[43] R. Girju, A. Badulescu, D. Moldovan, Automatic discovery of part-whole relations, Computational Linguistics 32 (2006) 83–135.

[44] M. van Gompel, UvT-WSD1: A cross-lingual word sense disambiguation system, in: Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010), Uppsala, Sweden, 15–16 July 2010, pp. 238–241.

[45] W. Guo, M. Diab, COLEPL and COLSLM: An unsupervised WSD approach to multilingual lexical substitution, tasks 2 and 3 SemEval 2010, in: Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010), Uppsala, Sweden, 15–16 July 2010, pp. 129–133.

[46] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, D. Klein, Learning bilingual lexicons from monolingual corpora, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, Ohio, 15–20 June 2008, pp. 771–779.

[47] S.M. Harabagiu, G.A. Miller, D.I. Moldovan, WordNet 2 – a morphologically and semantically enhanced resource, in: Proceedings of the SIGLEX99 Workshop on Standardizing Lexical Resources, 1999, pp. 1–8.

[48] S.M. Harabagiu, D. Moldovan, M. Paşca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, P. Morarescu, FALCON: Boosting knowledge for answer engines, in: Proceedings of the Ninth Text REtrieval Conference, Gaithersburg, Maryland, November 15–20, 2000, pp. 479–488.

[49] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proceedings of the 15th International Conference on Computational Linguistics, Nantes, France, 23–28 August 1992, pp. 539–545.

[50] J. Hoffart, F.M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, G. Weikum, YAGO2: Exploring and querying world knowledge in time, space, context, and many languages, in: Proceedings of the 20th World Wide Web Conference, Hyderabad, India, 28 March–25 April 2011, pp. 229–232.

[51] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, R. Weischedel, Ontonotes: The 90% solution, in: Companion Volume to the Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, New York, NY, 4–9 June 2006, pp. 57–60.

[52] N. Ide, Cross-lingual sense determination: Can it work? Computers and the Humanities 34 (2000) 223–234.

[53] N. Ide, T. Erjavec, D. Tufiş, Sense discrimination with parallel corpora, in: Proceedings of the ACL-02 Workshop on WSD: Recent Successes and Future Directions, Philadelphia, PA, July 2002, pp. 54–60.

[54] M. Ito, K. Nakayama, T. Hara, S. Nishio, Association thesaurus construction methods based on link co-occurrence analysis for Wikipedia, in: Proceedings of the Seventeenth ACM Conference on Information and Knowledge Management, Napa Valley, CA, 26–30 October 2008, pp. 817–826.

[55] A. Kilgarriff, Googleology is bad science, Computational Linguistics 33 (2007) 147–151.

[56] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: Proceedings of Machine Translation Summit X, Phuket, Thailand, 2005, pp. 79–86.

[57] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: open source toolkit for statistical machine translation, in: Companion Volume to the Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 23–30 June 2007, pp. 177–180.

[58] P. Koehn, K. Knight, Learning a translation lexicon from monolingual corpora, in: Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition, Philadelphia, PA, July 2002, pp. 9–16.

[59] R. Koeling, D. McCarthy, Sussx: WSD using automatically acquired predominant senses, in: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007 pp. 314–317.

[60] E. Lefever, V. Hoste, SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation, in: Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010), Uppsala, Sweden, 15–16 July 2010, pp. 15–20.

[61] E. Lefever, V. Hoste, M.D. Cock, ParaSense or how to use parallel corpora for Word Sense Disambiguation, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, 19–24 June 2011, pp. 317–322.

[62] L. Lemnitzer, C. Kunze, GermaNet – representation, visualization, application, in: Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain, 29–31 May 2002, pp. 1485–1491.

[63] D.B. Lenat, Cyc: A large-scale investment in knowledge infrastructure, Communications of the ACM 38 (1995).

[64] A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, A. Zampolli, SIMPLE: A general framework for the development of multilingual lexicons, International Journal of Lexicography 13 (2000) 249–263.

[65] C.Y. Lin, E. Hovy, The automated acquisition of topic signatures for text summarization, in: Proceedings of the 18th International Conference on Computational Linguistics, Saarbrücken, Germany, 31 July–4 August 2000, pp. 495–501.

[66] L.V. Lita, W.A. Hunt, E. Nyberg, Resource analysis for question answering, in: Companion Volume to the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004, pp. 162–165.

[67] Mausam, S. Soderland, O. Etzioni, D. Weld, M. Skinner, J. Bilmes, Compiling a massive, multilingual dictionary via probabilistic inference, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, 2–7 July 2009, pp. 262–270.

[68] D. McCarthy, R. Navigli, The English lexical substitution task, Language Resources and Evaluation 43 (2009) 139–159.

[69] O. Medelyan, D. Milne, C. Legg, I.H. Witten, Mining meaning from Wikipedia, International Journal of Human–Computer Studies 67 (2009) 716–754.

[70] D. Melamed, A word-to-word model of translational equivalence, in: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA, 24–27 June 1996, pp. 238–241.

[71] G. de Melo, G. Weikum, Towards a universal wordnet by learning from combined evidence, in: Proceedings of the Eighteenth ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009, pp. 513–522.

[72] G. de Melo, G. Weikum, MENTA: Inducing multilingual taxonomies from Wikipedia, in: Proceedings of the Nineteenth ACM Conference on Information and, Knowledge Management, Toronto, Canada, 26–30 October 2010, pp. 1099–1108.

[73] R. Mihalcea, Unsupervised large-vocabulary Word Sense Disambiguation with graph-based algorithms for sequence data labeling, in: Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005, pp. 411–418.

[74] R. Mihalcea, T. Chklovski, A. Kilgarriff, The Senseval-3 English lexical sample task, in: Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04, Barcelona, Spain, 25–26 July 2004, pp. 25–28.

[75] R. Mihalcea, D. Moldovan, eXtended WordNet: Progress report, in: Proceedings of the NAACL-01 Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA, June 2001, pp. 95–100.

[76] R. Mihalcea, P. Tarau, E. Figa, PageRank on semantic networks, with application to Word Sense Disambiguation, in: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23–27 August 2004, pp. 1126–1132.

[77] G.A. Miller, R. Beckwith, C.D. Fellbaum, D. Gross, K. Miller, WordNet: an online lexical database, International Journal of Lexicography 3 (1990) 235–244.

[78] G.A. Miller, F. Hristea, WordNet nouns: Classes and instances, Computational Linguistics 32 (2006) 1–3.

[79] G.A. Miller, C. Leacock, R. Tengi, R. Bunker, A semantic concordance, in: Proceedings of the 3rd DARPA Workshop on Human Language Technology, Plainsboro, NJ, 1993, pp. 303–308.

[80] D. Milne, I.H. Witten, An effective, low-cost measure of semantic relatedness obtained from Wikipedia links, in: Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08, Chicago, IL, 13 July, 2008, pp. 25–30.

[81] A. Moro, R. Navigli, WiSeNet: Building a Wikipedia-based semantic network with ontologized relations, in: Proceedings of the 21st ACM Conference on Information and Knowledge Management, Maui, Hawaii, 2012.

[82] V. Nastase, Topic-driven multi-document summarization with encyclopedic knowledge and activation spreading, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Waikiki, Honolulu, Hawaii, 25–27 October 2008, pp. 763–772.

[83] V. Nastase, M. Strube, Decoding Wikipedia category names for knowledge acquisition, in: Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence, Chicago, IL, 13–17 July 2008, pp. 1219–1224.

[84] V. Nastase, M. Strube, B. Börschinger, C. Zirn, A. Elghafari, WikiNet: A very large scale multi-lingual concept network, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, 19–21 May 2010.

[85] R. Navigli, Semi-automatic extension of large-scale linguistic knowledge bases, in: Proceedings of the 18th International Florida AI Research Symposium Conference, Clearwater Beach, FL, 15–17 May 2005, pp. 548–553.

[86] R. Navigli, Meaningful clustering of senses helps boost word sense disambiguation performance, in: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–21 July 2006, Sydney, Australia, pp. 105–112.

[87] R. Navigli, Word Sense Disambiguation: A survey, ACM Computing Surveys 41 (2009) 1–69.

[88] R. Navigli, S. Faralli, A. Soroa, O.L. de Lacalle, E. Agirre, Two birds with one stone: Learning semantic models for Text Categorization and Word Sense Disambiguation, in: Proceedings of the Twentieth ACM Conference on Information and Knowledge Management, Glasgow, Scotland, UK, 24–28 October 2011, pp. 2317–2320.

[89] R. Navigli, M. Lapata, An experimental study on graph connectivity for unsupervised Word Sense Disambiguation, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 678–692.

[90] R. Navigli, K.C. Litkowski, O. Hargraves, Semeval-2007 task 07: Coarse-grained English all-words task, in: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007, pp. 30–35.

[91] R. Navigli, S.P. Ponzetto, BabelNet: Building a very large multilingual semantic network, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 216–225.

[92] R. Navigli, S.P. Ponzetto, BabelNetXplorer: a platform for multilingual lexical knowledge base access and exploration, in: Companion Volume to the Proceedings of the 21st World Wide Web Conference, Lyon, France, 16–20 April 2012, pp. 393–396.

[93] R. Navigli, S.P. Ponzetto, Multilingual WSD with just a few lines of code: the BabelNet API, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, South Korea, 8–14 July 2012, pp. 67–72.

[94] R. Navigli, P. Velardi, Structural semantic interconnections: A knowledge-based approach to Word Sense Disambiguation, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1075–1088.

[95] 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, Uppsala, Sweden, 11–16 July 2010, pp. 1318–1327.

[96] 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, 16–22 July 2011, pp. 1872–1877.

[97] R. Navigli, P. Velardi, J.M. Ruiz-Martínez, An annotated dataset for extracting definitions and hypernyms from the web, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, 19–21 May 2010.

[98] H.T. Ng, H.B. Lee, Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach, in: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA, 24–27 June 1996, pp. 40–47.

[99] E. Niemann, I. Gurevych, The people's web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet, in: Proceedings of the 9th International Conference on Computational Semantics, Oxford, UK, pp. 205–214.

[100] I. Niles, A. Pease, Towards a standard upper ontology, in: Proceedings of the 2nd International Conference on Formal Ontology in Information Systems, Ogunquit, Maine, 17–19 October 2001, pp. 2–9.

[101] F.J. Och, H. Ney, A systematic comparison of various statistical alignment models, Computational Linguistics 29 (2003) 19–51.

[102] A. Pease, C. Fellbaum, P. Vossen, Building the global WordNet grid, in: Proceedings of the 18th International Congress of Linguists (CIL18), Seoul, South Korea, 21–26 July 2008.

[103] M. Pennacchiotti, P. Pantel, Ontologizing semantic relations, in: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–21 July 2006, pp. 793–800.

[104] E. Pianta, L. Bentivogli, C. Girardi, MultiWordNet: Developing an aligned multilingual database, in: Proceedings of the 1st International Global WordNet Conference, Mysore, India, 21–25 January 2002, pp. 21–25.

[105] S.P. Ponzetto, R. Navigli, Large-scale taxonomy mapping for restructuring and integrating Wikipedia, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, CA, 14–17 July 2009, pp. 2083–2088.

[106] S.P. Ponzetto, R. Navigli, Knowledge-rich Word Sense Disambiguation rivaling supervised systems, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 1522–1531.

[107] S.P. Ponzetto, M. Strube, Knowledge derived from Wikipedia for computing semantic relatedness, Journal of Artificial Intelligence Research 30 (2007) 181–212.

[108] S.P. Ponzetto, M. Strube, Taxonomy induction based on a collaboratively built knowledge repository, Artificial Intelligence 175 (2011) 1737–1756.

[109] S. Pradhan, E. Loper, D. Dligach, M. Palmer, Semeval-2007 task-17: English lexical sample, SRL and all words, in: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007, pp. 87–92.

[110] P. Proctor (Ed.), Longman Dictionary of Contemporary English, Longman Group, Harlow, UK, 1978.

[111] A. Rahman, V. Ng, Narrowing the modeling gap: A cluster-ranking approach to coreference resolution, Journal of Artificial Intelligence Research 40 (2011) 469–521.

[112] P. Resnik, D. Yarowsky, Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation, Journal of Natural Language Engineering 5 (1999) 113–133.

[113] M. Richardson, P. Domingos, Building large knowledge bases by mass collaboration, in: Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP), Sanibel Island, FL, 23–25 October 2003, pp. 129–137.

[114] A.E. Richman, P. Schone, Mining wiki resources for multilingual named entity recognition, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, Ohio, 15–20 June 2008, pp. 1–9.

[115] G. Rigau, H. Rodríguez, E. Agirre, Building accurate semantic taxonomies from monolingual MRDs, in: Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada, 10–14 August 1998, pp. 1103–1109.

[116] P.M. Roget, Roget's International Thesaurus, 1st edition, Cromwell, New York, USA, 1911.

[117] 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.

[118] B. Sagot, D. Fišer, Building a free French WordNet from multilingual resources, in: Proceedings of the Ontolex 2008 Workshop, Marrakech, Morocco, 31 May 2008.

[119] M. Sammer, S. Soderland, Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons, in: Proceedings of Machine Translation Summit XI, 2007.

[120] C. Sauper, R. Barzilay, Automatically generating Wikipedia articles: A structure-aware approach, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, 2–7 July 2009, pp. 208–216.

[121] C. Silberer, S.P. Ponzetto UHD, Cross-lingual Word Sense Disambiguation using multilingual co-occurrence graphs, in: Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010), Uppsala, Sweden, 15–16 July 2010, pp. 134–137.

[122] 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, Sydney, Australia, 17–21 July 2006, pp. 801–808.

[123] F.M. Suchanek, G. Ifrim, G. Weikum, Combining linguistic and statistical analysis to extract relations from web documents, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, 20–23 August 2006, pp. 712–717.

[124] F.M. Suchanek, G. Kasneci, G. Weikum, Yago: A large ontology from Wikipedia and WordNet, Journal of Web Semantics 6 (2008) 203–217.

[125] M. Taboada, J. Brooke, M. Tofiloski, K.D. Voll, M. Stede, Lexicon-based methods for Sentiment Analysis, Computational Linguistics 37 (2011) 267–307.

[126] A. Toral, O. Ferrández, E. Agirre, R. Muñoz, A study on linking Wikipedia categories to WordNet synsets using text similarity, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria, 14–16 September 2009, pp. 449–454.

[127] D. Tufiş, D. Cristea, S. Stamou, BalkaNet: Aims, methods, results and perspectives. A general overview, Romanian Journal on Science and Technology of Information 7 (2004) 9–43.

[128] P. Vossen (Ed.), EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer, Dordrecht, The Netherlands, 1998.

[129] X. Wan, Bilingual co-training for sentiment classification of Chinese product reviews, Computational Linguistics 37 (2011) 587–616.

[130] P. Wang, C. Domeniconi, Building semantic kernels for text classification using Wikipedia, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, 24–27 August 2008, pp. 713–721.

[131] W. Wentland, J. Knopp, C. Silberer, M. Hartung, Building a multilingual lexical resource for named entity disambiguation, translation and transliteration, in: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May–1 June 2008.

[132] K. Woodsend, M. Lapata, Learning to simplify sentences with quasi-synchronous grammar and integer programming, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 27–29 July 2011, pp. 409–420.

[133] F. Wu, J. Madhavan, A. Halevy, Identifying aspects for web-search queries, Journal of Artificial Intelligence Research 40 (2011) 667–700.

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

[135] F. Wu, D. Weld, Automatically refining the Wikipedia infobox ontology, in: Proceedings of the 17th World Wide Web Conference, Beijing, China, 21–25 April 2008, pp. 635–644.

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

[137] D. Yarowsky, R. Florian, Evaluating sense disambiguation across diverse parameter spaces, Natural Language Engineering 9 (2002) 293–310.

[138] Z. Ye, X. Huang, H. Lin, A graph-based approach to mining multilingual word associations from Wikipedia, in: Proceedings of the 32nd Annual International ACM Conference on Research and Development in Information Retrieval, Boston, MA, 19–23 July 2009, pp. 690–691.

[139] T. Yokoi, The EDR electronic dictionary, Communications of the ACM 38 (1995) 42–44.

[140] Z. Zhong, H.T. Ng, Y.S. Chan, Word Sense Disambiguation using OntoNotes: An empirical study, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Waikiki, Honolulu, Hawaii, 25–27 October, pp. 1002–1010.