Entity Linking meets Word Sense Disambiguation: a Unified Approach

Andrea Moro, Alessandro Raganato, Roberto Navigli
Dipartimento di Informatica,
Sapienza Università di Roma,
Viale Regina Elena 295, 00161 Roma, Italy
{moro,navigli}@di.uniroma1.it
ale.raganato@gmail.com

Abstract

Entity Linking (EL) and Word Sense Disambiguation (WSD) both address the lexical ambiguity of language. But while the two tasks are pretty similar, they differ in a fundamental respect: in EL the textual mention can be linked to a named entity which may or may not contain the exact mention, while in WSD there is a perfect match between the word form (better, its lemma) and a suitable word sense.

In this paper we present Babelfy, a unified graph-based approach to EL and WSD based on a loose identification of candidate meanings coupled with a densest subgraph heuristic which selects high-coherence semantic interpretations. Our experiments show state-of-the-art performances on both tasks on 6 different datasets, including a multilingual setting. Babelfy is online at http://babelfy.org

1 Introduction

The automatic understanding of the meaning of text has been a major goal of research in computational linguistics and related areas for several decades, with ambitious challenges, such as Machine Reading (Etzioni et al., 2006) and the quest for knowledge (Schubert, 2006). Word Sense Disambiguation (WSD) (Navigli, 2009; Navigli, 2012) is a historical task aimed at assigning meanings to single-word and multi-word occurrences within text, a task which is more alive than ever in the research community.

Recently, the collaborative creation of large semi-structured resources, such as Wikipedia, and knowledge resources built from them (Hovy et al., 2013), such as BabelNet (Navigli and Ponzetto, 2012a), DBpedia (Auer et al., 2007) and YAGO2 (Hoffart et al., 2013), has favoured the emergence of new tasks, such as Entity Linking (EL) (Rao et al., 2013), and opened up new possibilities for tasks such as Named Entity Disambiguation (NED) and Wikification. The aim of EL is to discover mentions of entities within a text and to link them to the most suitable entry in a reference knowledge base. However, in contrast to WSD, a mention may be partial while still being unambiguous thanks to the context. For instance, consider the following sentence:

(1) Thomas and Mario are strikers playing in Munich.

This example makes it clear how intertwined the two tasks of WSD and EL are. In fact, on the one hand, *striker* and *play* are polysemous words which can be disambiguated by selecting the game/soccer playing senses of the two words in a dictionary; on the other hand, *Thomas* and *Mario* are partial mentions which have to be linked to the appropriate entries of a knowledge base, that is, *Thomas Müller* and *Mario Gomez*, two well-known soccer players.

The two main differences between WSD and EL lie, on the one hand, in the kind of inventory used, i.e., dictionary vs. encyclopedia, and, on the other hand, in the assumption that the mention is complete or potentially partial. Notwithstanding these differences, the tasks are similar in nature, in that they both involve the disambiguation of textual fragments according to a reference inventory. However, the research community has so far tackled the two tasks separately, often duplicating efforts and solutions.

In contrast to this trend, research in knowledge acquisition is now heading towards the seamless in-

tegration of encyclopedic and lexicographic knowledge into structured language resources (Hovy et al., 2013), and the main representative of this new direction is undoubtedly BabelNet (Navigli and Ponzetto, 2012a). Given such structured language resources it seems natural to suppose that they might provide a common ground for the two tasks of WSD and EL.

More precisely, in this paper we explore the hypothesis that the lexicographic knowledge used in WSD is also useful for tackling the EL task, and, vice versa, that the encyclopedic information utilized in EL helps disambiguate nominal mentions in a WSD setting. We propose Babelfy, a novel, unified graph-based approach to WSD and EL, which performs two main steps: i) it exploits random walks with restart, and triangles as a support for reweighting the edges of a large semantic network; ii) it uses a densest subgraph heuristic on the available semantic interpretations of the input text to perform a joint disambiguation with both concepts and named entities. Our experiments show the benefits of our synergistic approach on six gold-standard datasets.

2 Related Work

2.1 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is the task of choosing the right sense for a word within a given context. Typical approaches for this task can be classified as i) supervised, ii) knowledge-based, and iii) unsupervised. However, supervised approaches require huge amounts of annotated data (Zhong and Ng, 2010; Shen et al., 2013; Pilehvar and Navigli, 2014), an effort which cannot easily be repeated for new domains and languages, while unsupervised ones suffer from data sparsity and an intrinsic difficulty in their evaluation (Agirre et al., 2006; Brody and Lapata, 2009; Manandhar et al., 2010; Van de Cruys and Apidianaki, 2011; Di Marco and Navigli, 2013). On the other hand, knowledge-based approaches are able to obtain good performance using readily-available structured knowledge (Agirre et al., 2010; Guo and Diab, 2010; Ponzetto and Navigli, 2010; Miller et al., 2012; Agirre et al., 2014). Some of these approaches marginally take into account the structural properties of the knowledge base (Mihalcea, 2005). Other approaches, instead, leverage the structural properties of the knowledge base by exploiting centrality and connectivity measures (Sinha and Mihalcea, 2007; Tsatsaronis et al., 2007; Agirre and Soroa, 2009; Navigli and Lapata, 2010).

One of the key steps of many knowledge-based WSD algorithms is the creation of a graph representing the semantic interpretations of the input text. Two main strategies to build this graph have been proposed: i) exploiting the direct connections, i.e., edges, between the considered sense candidates; ii) populating the graph according to (shortest) paths between them. In our approach we manage to unify these two strategies by automatically creating edges between sense candidates performing Random Walk with Restart (Tong et al., 2006).

The recent upsurge of interest in multilinguality has led to the development of cross-lingual and multilingual approaches to WSD (Lefever and Hoste, 2010; Lefever and Hoste, 2013; Navigli et al., 2013). Multilinguality has been exploited in different ways, e.g., by using parallel corpora to build multilingual contexts (Guo and Diab, 2010; Banea and Mihalcea, 2011; Lefever et al., 2011) or by means of ensemble methods which exploit complementary sense evidence from translations in different languages (Navigli and Ponzetto, 2012b). In this work, we present a novel exploitation of the structural properties of a multilingual semantic network.

2.2 Entity Linking

Entity Linking (Erbs et al., 2011; Rao et al., 2013; Cornolti et al., 2013) encompasses a set of similar tasks, which include Named Entity Disambiguation (NED), that is the task of linking entity mentions in a text to a knowledge base (Bunescu and Pasca, 2006; Cucerzan, 2007), and Wikification, i.e., the automatic annotation of text by linking its relevant fragments of text to the appropriate Wikipedia articles. Mihalcea and Csomai (2007) were the first to tackle the Wikification task. In their approach they disambiguate each word in a sentence independently by exploiting the context in which it occurs. However, this approach is local in that it lacks a collective notion of coherence between the selected Wikipedia pages. To overcome this problem, Cucerzan (2007) introduced a global approach based on the simultaneous disambiguation of all the terms in a text and the use of lexical context to disambiguate the mentions. To maximize the semantic agreement Milne

and Witten (2008) introduced the analysis of the semantic relations between the candidate senses and the unambiguous context, i.e., words with a single sense candidate. However, the performance of this algorithm depends heavily on the number of links incident to the target senses and on the availability of unambiguous words within the input text. To overcome this issue a novel class of approaches have been proposed (Kulkarni et al., 2009; Ratinov et al., 2011; Hoffart et al., 2011) that exploit global and local features. However, these systems either rely on a difficult NP-hard formalization of the problem which is infeasible for long text, or exploit popularity measures which are domain-dependent. In contrast, we show that the semantic network structure can be leveraged to obtain state-of-the-art performance by synergistically disambiguating both word senses and named entities at the same time.

Recently, the explosion of on-line social networking services, such as Twitter and Facebook, have contributed to the development of new methods for the efficient disambiguation of short texts (Ferragina and Scaiella, 2010; Hoffart et al., 2012; Böhm et al., 2012). Thanks to a loose candidate identification technique coupled with a densest subgraph heuristic, we show that our approach is particularly suited for short and highly ambiguous text disambiguation.

2.3 The Best of Two Worlds

Our main goal is to bring together the two worlds of WSD and EL. On the one hand, this implies relaxing the constraint of a perfect association between mentions and meanings, which is, instead, assumed in WSD. On the other hand, this relaxation leads to the inherent difficulty of encoding a full-fledged sense inventory for EL. Our solution to this problem is to keep the set of candidate meanings for a given mention as open as possible (see Section 6), so as to enable high recall in linking partial mentions, while providing an effective method for handling this high ambiguity (see Section 7).

A key assumption of our work is that the lexicographic knowledge used in WSD is also useful for tackling the EL task, and vice versa the encyclopedic information utilized in EL helps disambiguate nominal mentions in a WSD setting. We enable the joint treatment of concepts and named entities by enforcing high coherence in our semantic interpretations.

3 WSD and Entity Linking Together

Task. Our task is to disambiguate and link all nominal and named entity mentions occurring within a text. The linking task is performed by associating each mention with the most suitable entry of a given knowledge base.[1]

We point out that our definition is unconstrained in terms of what to link, i.e., unlike Wikification and WSD, we can link overlapping fragments of text. For instance, given the text fragment *Major League Soccer*, we identify and disambiguate several different nominal and entity mentions: *Major League Soccer*, *major league*, *league* and *soccer*. In contrast to EL, we link not only named entity mentions, such as *Major League Soccer*, but also nominal mentions, e.g., *major league*, to their corresponding meanings in the knowledge base.

Babelfy. We provide a unified approach to WSD and entity linking in three steps:

1. Given a lexicalized semantic network, we associate with each vertex, i.e., either concept or named entity, a semantic signature, that is, a set of related vertices (Section 5). This is a preliminary step which needs to be performed only once, independently of the input text.

2. Given a text, we extract all the linkable fragments from this text and, for each of them, list the possible meanings according to the semantic network (Section 6).

3. We create a graph-based semantic interpretation of the whole text by linking the candidate meanings of the extracted fragments using the previously-computed semantic signatures. We then extract a dense subgraph of this representation and select the best candidate meaning for each fragment (Section 7).

4 Semantic Network

Our approach requires the availability of a wide-coverage semantic network which encodes structural and lexical information both of an encyclopedic and of a lexicographic kind. Although in principle any semantic network with these properties

[1] Mentions which are not contained in the reference knowledge base are not taken into account.

could be utilized, in our work we used the BabelNet[2] 1.1.1 semantic network (Navigli and Ponzetto, 2012a) since it is the largest multilingual knowledge base, obtained from the automatic seamless integration of Wikipedia[3] and WordNet (Fellbaum, 1998). We consider BabelNet as a directed multigraph which contains both concepts and named entities as its vertices and a multiset of semantic relations as its edges. We leverage the multilingual lexicalizations of the vertices of BabelNet to identify mentions in the input text. For example, the entity *FC Bayern Munich* can be lexicalized in different languages, e.g., *F.C. Bayern de Múnich* in Spanish, *Die Roten* in English and *Bayern München* in German, among others. As regards semantic relations, the only information we use is that of the end points, i.e., vertices, that these relations connect, while neglecting the relation type.

5 Building Semantic Signatures

One of the major issues affecting both manually-curated and automatically constructed semantic networks is data sparsity. For instance, we calculated that the average number of incident edges is roughly 10 in WordNet, 50 in BabelNet and 80 in YAGO2, to mention a few. Although automatically-built resources typically provide larger amounts of edges, two issues have to be taken into account: concepts which should be related might not be directly connected despite being structurally close within the network, and, vice versa, weakly-related or even unrelated concepts can be erroneously connected by an edge. For instance, in BabelNet we do not have an edge between *playmaker* and *Thomas Müller*, while we have an incorrect edge connecting *FC Bayern Munich* and *Yellow Submarine (song)*. However, this crisp notion of relatedness can be overcome by exploiting the global structure of the semantic network, thereby obtaining a more precise and higher-coverage measure of relatedness. We address this issue in two steps: first, we provide a structural weighting of the network's edges; second, for each vertex we create a set of related vertices using random walks with restart.

[2]http://babelnet.org

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

Structural weighting. Our first objective is to assign higher weights to edges which are involved in more densely connected areas of the directed network. To this end, inspired by the local clustering coefficient measure (Watts and Strogatz, 1998) and its recent success in Word Sense Induction (Di Marco and Navigli, 2013), we use directed triangles, i.e., directed cycles of length 3, and weight each edge (v, v') by the number of directed triangles it occurs in:

$$weight(v, v') := |\{(v, v', v'') : (v, v'), (v', v''), (v'', v) \in E\}| + 1 \quad (1)$$

We add one to each weight to ensure the highest degree of reachability in the network.

Random Walk with Restart. Our goal is to create a *semantic signature* (i.e., a set of highly related vertices) for each concept and named entity of the semantic network. To do this, we perform a Random Walk with Restart (RWR) (Tong et al., 2006), that is, a stochastic process that starts from an initial vertex of the graph[4] and then, for a fixed number n of steps or until convergence, explores the graph by choosing the next vertex within the current neighborhood or by restarting from the initial vertex with a given, fixed *restart probability* α. For each edge (v, v') in the network, we model the conditional probability $P(v'|v)$ as the normalized weight of the edge:

$$P(v'|v) = \frac{weight(v, v')}{\sum_{v'' \in V} weight(v, v'')}$$

where V is the set of vertices of the semantic network and $weight(v, v')$ is the function defined in Equation 1. We then run the RWR from each vertex v of the semantic network for a fixed number n of steps (we show in Algorithm 1 our RWR pseudocode). We keep track of the encountered vertices using the map $counts$, i.e., we increase the counter associated with vertex v' in $counts$ every time we hit v' during a RWR started from v (see line 11). As a result, we obtain a frequency distribution over the whole set of concepts and entities. To eliminate weakly-related vertices we keep only those items that were hit at least η times (see lines 16–18). Finally, we save the remaining vertices in the set $semSign_v$ which is the semantic signature of v (see line 19).

[4]RWR can be used with an initial set of vertices, however in this paper we use a single initial vertex.

Algorithm 1 Random walk with restart.

```
1:  input: v, the starting vertex;
           α, the restart probability;
           n, the number of steps to be executed;
           P, the transition probabilities;
           η, the frequency threshold.
2:  output: semSign_v, set of related vertices for v.
3:  function RWR(v, α, n, P, η)
4:      v' := v
5:      counts := new Map < Synset, Integer >
6:      while n > 0 do
7:          if random() > α then
8:              given the transition probabilities P(·|v')
9:              of v', choose a random neighbor v''
10:             v' := v''
11:             counts[v'] + +
12:         else
13:             restart the walk
14:             v' := v
15:         n := n − 1
16:     for each v' in counts.keys() do
17:         if counts[v'] < η then
18:             remove v' from counts.keys()
19:     return semSign_v = counts.keys()
```

The creation of our set of semantic signatures, one for each vertex in the semantic network, is a preliminary step carried out once only before starting processing any input text. We now turn to the candidate identification and disambiguation steps.

6 Candidate Identification

Given a text as input, we apply part-of-speech tagging and identify the set F of all the textual fragments, i.e., all the sequences of words of maximum length five, which contain at least one noun and that are substrings of lexicalizations in BabelNet, i.e., those fragments that can potentially be linked to an entry in BabelNet. For each textual fragment $f \in F$, i.e., a single- or multi-word expression of the input text, we look up the semantic network for candidate meanings, i.e., vertices that contain f or, only for named entities, a superstring of f as their lexicalization. For instance, for sentence (1) in the introduction, we identify the following textual fragments: *Thomas*, *Mario*, *strikers*, *Munich*. This output is obtained thanks to our loose candidate identification routine, i.e., based on superstring matching instead of exact matching, which, for instance, enables us to recognize the right candidate *Mario Gomez* for the mention *Mario* even if this named entity does not have *Mario* as one of its lexicalizations (for an analysis of the impact of this routine against the exact matching approach see the discussion in Section 9).

Moreover, as we stated in Section 3, we allow overlapping fragments, e.g., for *major league* we recognize *league* and *major league*. We denote with $cand(f)$ the set of all the candidate meanings of fragment f. For instance, for the noun *league* we have that $cand(league)$ contains among others the *sport* word sense and the *TV series* named entity.

7 Candidate Disambiguation

Semantic interpretation graph. After the identification of fragments (F) and their candidate meanings ($cand(\cdot)$), we create a directed graph $G_I = (V_I, E_I)$ of the semantic interpretations of the input text. We show the pseudocode in Algorithm 2. V_I contains all the candidate meanings of all fragments, that is, $V_I := \{(v, f) : v \in cand(f), f \in F\}$, where f is a fragment of the input text and v is a candidate Babel synset that has a lexicalization which is equal to or is a superstring of f (see lines 4–8). The set of edges E_I connects related meanings and is populated as follows: we add an edge from (v, f) to (v', f') if and only if $f \neq f'$ and $v' \in semSign_v$ (see lines 9–11). In other words, we connect two candidate meanings of different fragments if one is in the semantic signature of the other. For instance, we add an edge between (*Mario Gomez, Mario*) and (*Thomas Müller, Thomas*), while we do not add one between (*Mario Gomez*, *Mario*) and (*Mario Basler*, *Mario*) since these are two candidate meanings of the same fragment, i.e., *Mario*. In Figure 1, we show an excerpt of our graph for sentence (1).

At this point we have a graph-based representation of all the possible interpretations of the input text. In order to drastically reduce the degree of ambiguity while keeping the interpretation coherence as high as possible, we apply a novel densest subgraph heuristic (see line 12), whose description we defer to the next paragraph. The result is a subgraph which contains those semantic interpretations that are most coherent to each other. However, this subgraph might still contain multiple interpretations for the same fragment, and even unambiguous fragments which are not correct. Therefore, the final

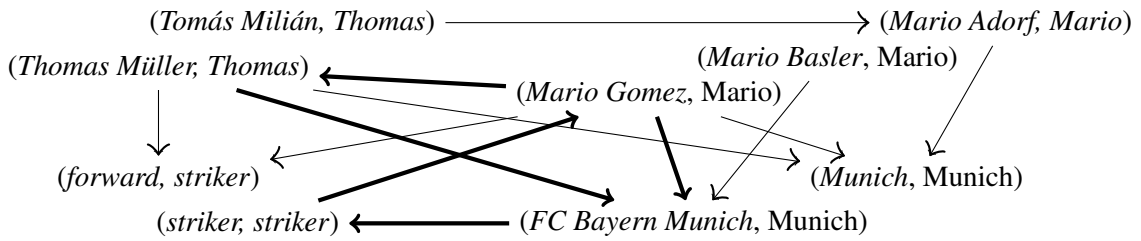

Figure 1: An excerpt of the semantic interpretation graph automatically built for the sentence *Thomas and Mario are strikers playing in Munich* (the edges connecting the correct meanings are in bold).

step is the selection of the most suitable candidate meaning for each fragment f given a threshold θ to discard semantically unrelated candidate meanings. We score each meaning $v \in cand(f)$ with its normalized weighted degree[5] in the densest subgraph:

$$score((v,f)) = \frac{w_{(v,f)} \cdot deg((v,f))}{\sum_{v' \in cand(f)} w_{(v',f)} \cdot deg((v',f))} \quad (2)$$

where $w_{(v,f)}$ is the fraction of fragments the candidate meaning v connects to:

$$w_{(v,f)} := \frac{|\{f' \in F : \exists v' \; s.t. \; ((v,f),(v',f')) \text{ or } ((v',f'),(v,f)) \in E_I\}|}{|F| - 1}$$

The rationale behind this scoring function is to take into account both the semantic coherence, using a graph centrality measure among the candidate meanings, and the lexical coherence, in terms of the number of fragments a candidate relates to.

Finally, we link each f to the highest ranking candidate meaning $v^\star$ if $score((v^\star, f)) \geq \theta$, where θ is a fixed threshold (see lines 14–18 of Algorithm 2). For instance, in sentence (1) and for the fragment *Mario* we select *Mario Gomez* as our final candidate meaning and link it to the fragment.

Linking by densest subgraph. We now illustrate our novel densest subgraph heuristic, used in line 12 of Algorithm 2, for reducing the level of ambiguity of the initial semantic interpretation graph G_I. The main idea here is that the most suitable meanings of each text fragment will belong to the densest area of the graph. For instance, in Figure 1 the (candidate, fragment) pairs *(Thomas Müller, Thomas), (Mario Gomez, Mario), (striker, striker)* and *(FC Bayern Munich, Munich)* form a dense subgraph supporting their relevance for sentence (1).

[5]We denote with $deg(v)$ the overall number of incoming and outgoing edges, i.e., $deg(v) := deg^+(v) + deg^-(v)$.

Algorithm 2 Candidate Disambiguation.

1: **input:** F, the fragments in the input text;
   $semSign$, the semantic signatures;
   μ, ambiguity level to be reached;
   $cand$, fragments to candidate meanings.
2: **output:** $selected$, disambiguated fragments.
3: **function** DISAMB($F, semSign, \mu, cand$)
4: $V_I := \emptyset$; $E_I := \emptyset$
5: $G_I := (V_I, E_I)$
6: **for each** fragment $f \in F$ **do**
7: **for each** candidate $v \in cand(f)$ **do**
8: $V_I := V_I \cup \{(v,f)\}$
9: **for each** $((v,f),(v',f')) \in V_I \times V_I$ **do**
10: **if** $f \neq f'$ **and** $v' \in semSign_v$ **then**
11: $E_I := E_I \cup \{((v,f),(v',f'))\}$
12: $G_I^\star := \text{DENSSUB}(F, cand, G_I, \mu)$
13: $selected :=$ **new** $Map < String, Synset >$
14: **for each** $f \in F \; s.t. \; \exists (v,f) \in V_I^\star$ **do**
15: $cand^\star(f) := \{v : (v,f) \in V_I^\star\}$
16: $v^\star := \arg\max_{v \in cand^\star(f)} score((v,f))$
17: **if** $score((v^\star, f)) \geq \theta$ **then**
18: $selected(f) := v^\star$
19: **return** $selected$

The problem of identifying the densest subgraph of size at least k is NP-hard (Feige et al., 1999). Therefore, we define a heuristic for k-partite graphs inspired by a 2-approximation greedy algorithm for arbitrary graphs (Charikar, 2000; Khuller and Saha, 2009). Our adapted strategy for selecting a dense subgraph of G_I is based on the iterative removal of low-coherence vertices, i.e., fragment interpretations. We show the pseudocode in Algorithm 3.

We start with the initial graph $G_I^{(0)}$ at step $t = 0$ (see line 5). For each step t (lines 7–16), first, we identify the most ambiguous fragment f_{max}, i.e., the one with the maximum number of candidate mean-

Algorithm 3 Densest Subgraph.

1: **input:** F, the set of all fragments in the input text; $cand$, from fragments to candidate meanings; $G_I^{(0)}$, the full semantic interpretation graph; μ, ambiguity level to be reached.
2: **output:** $G_I^\star$, a dense subgraph.
3: **function** DENSSUB($F, cand, G_I^{(0)}, \mu$)
4: $t := 0$
5: $G_I^\star := G_I^{(0)}$
6: **while** true **do**
7: $f_{max} := \arg\max_{f \in F} |\{v : \exists(v, f) \in V_I^{(t)}\}|$
8: **if** $|\{v : \exists(v, f_{max}) \in V_I^{(t)}\}| \leq \mu$ **then**
9: **break**;
10: $v_{min} := \underset{v \in cand(f_{max})}{\operatorname{argmin}} score((v, f_{max}))$
11: $V_I^{(t+1)} := V_I^{(t)} \setminus \{(v_{min}, f_{max})\}$
12: $E_I^{(t+1)} := E_I^{(t)} \cap V_I^{(t+1)} \times V_I^{(t+1)}$
13: $G_I^{(t+1)} := (V_I^{(t+1)}, E_I^{(t+1)})$
14: **if** $avgdeg(G_I^{(t+1)}) > avgdeg(G_I^\star)$ **then**
15: $G_I^\star := G_I^{(t+1)}$
16: $t := t + 1$
17: **return** $G_I^\star$

ings in the graph (see line 7). Next, we discard the weakest interpretation of the current fragment f_{max}. To do so, we determine the lexical and semantic coherence of each candidate meaning (v, f_{max}) using Formula 2 (see line 10). We then remove from our graph $G_I^{(t)}$ the lowest-coherence vertex (v_{min}, f_{max}), i.e., the one whose score is minimum (see lines 11–13). For instance, in Figure 1, f_{max} is the fragment *Mario* and we have: $score$((*Mario Gomez, Mario*)) $\propto \frac{3}{3} \cdot 5 = 5$, $score$((*Mario Basler, Mario*)) $\propto \frac{1}{3} \cdot 1 = 0.\overline{3}$ and $score$((*Mario Adorf, Mario*)) $\propto \frac{2}{3} \cdot 2 = 1.\overline{3}$, so we remove (*Mario Basler, Mario*) from the graph since its score is minimum.

We then move to the next step, i.e., we set $t := t + 1$ (see line 16) and repeat the low-coherence removal step. We stop when the number of remaining candidates for each fragment is below a threshold μ, i.e., $|\{v : \exists(v, f) \in V_I^{(t)}\}| \leq \mu \ \forall f \in F$ (see lines 8–9). During each iteration step t we compute the average degree of the current graph $G_I^{(t)}$, i.e., $avgdeg(G_I^{(t)}) = \frac{2|E_I^{(t)}|}{|V_I^{(t)}|}$. Finally, we select as the densest subgraph of the initial semantic interpretation graph G_I the graph $G_I^\star$ that maximizes the average degree (see lines 14–15).

8 Experimental Setup

Datasets. We carried out our experiments on six datasets, four for WSD and two for EL:

- The SemEval-2013 task 12 dataset for multilingual WSD (Navigli et al., 2013), which consists of 13 documents in different domains, available in 5 languages. For each language, all noun occurrences were annotated using BabelNet, thereby providing Wikipedia and WordNet annotations wherever applicable. The number of mentions to be disambiguated roughly ranges from 1K to 2K per language in the different setups.
- The SemEval-2007 task 7 dataset for coarse-grained English all-words WSD (Navigli et al., 2007). We take into account only nominal mentions obtaining a dataset containing 1107 nouns to be disambiguated using WordNet.
- The SemEval-2007 task 17 dataset for fine-grained English all-words WSD (Pradhan et al., 2007). We considered only nominal mentions resulting in 158 nouns annotated with WordNet synsets.
- The Senseval-3 dataset for English all-words WSD (Snyder and Palmer, 2004), which contains 899 nouns to be disambiguated using WordNet.
- KORE50 (Hoffart et al., 2012), which consists of 50 short English sentences (mean length of 14 words) with a total number of 144 mentions manually annotated using YAGO2, for which a Wikipedia mapping is available. This dataset was built with the idea of testing against a high level of ambiguity for the EL task.
- AIDA-CoNLL[6] (Hoffart et al., 2011), which consists of 1392 English articles, for a total of roughly 35K named entity mentions annotated with YAGO concepts separated in development, training and test sets.

We exploited the POS tags already available in the SemEval and Senseval datasets, while we used the Stanford POS tagger (Toutanova et al., 2003) for the English sentences in the last two datasets.

[6] We used AIDA-CoNLL as it is the most recent and largest available dataset for EL (Hachey et al., 2013). The TAC KBP datasets are available only to participants.

Parameters. We fixed the parameters of RWR (Section 5) to the values $\alpha = .85$, $\eta = 100$ and $n = 1\text{M}$ which maximize F1 on a manually created tuning set made up of 10 gold-standard semantic signatures. We tuned our two disambiguation parameters $\mu = 10$ and $\theta = 0.8$ by optimizing $F1$ on the trial dataset of the SemEval-2013 task on multilingual WSD (Navigli et al., 2013). We used the same parameters on all the other WSD datasets. As for EL, we used the training part of AIDA-CoNLL (Hoffart et al., 2011) to set $\mu = 5$ and $\theta = 0.0$.

8.1 Systems

Multilingual WSD. We evaluated our system on the SemEval-2013 task 12 by comparing it with the participating systems:

- UMCC-DLSI (Gutiérrez et al., 2013) a state-of-the-art Personalized PageRank-based approach that exploits the integration of different sources of knowledge, such as WordNet Domains/Affect (Strapparava and Valitutti, 2004), SUMO (Zouaq et al., 2009) and the eXtended WordNet (Mihalcea and Moldovan, 2001);
- DAEBAK! (Manion and Sainudiin, 2013) which performs WSD on the basis of peripheral diversity within subgraphs of BabelNet;
- GETALP (Schwab et al., 2013) which uses an Ant Colony Optimization technique together with the classical measure of Lesk (1986).

We also compared with UKB w2w (Agirre and Soroa, 2009), a state-of-the-art approach for knowledge-based WSD, based on Personalized PageRank (Haveliwala, 2002). We used the same mapping from words to senses that we used in our approach, default parameters[7] and BabelNet as the input graph. Moreover, we compared our system with IMS (Zhong and Ng, 2010), a state-of-the-art supervised English WSD system which uses an SVM trained on sense-annotated corpora, such as SemCor (Miller et al., 1993) and DSO (Ng and Lee, 1996), among others. We used the IMS model out-of-the-box with Most Frequent Sense (MFS) as backoff routine since the model obtained using the task trial data performed worse.

We followed the original task formulation and evaluated the synsets in three different settings, i.e., when using BabelNet senses, Wikipedia senses and WordNet senses, thanks to BabelNet being a superset of the other two inventories. We ran our system on a document-by-document basis, i.e., disambiguating each document at once, so as to test its effectiveness on long coherent texts. Performance was calculated in terms of F1 score. We also compared the systems with the MFS baseline computed for the three inventories (Navigli et al., 2013).

Coarse-grained WSD. For the SemEval-2007 task 7 we compared our system with the two top-ranked approaches, i.e., NUS-PT (Chan et al., 2007) and UoR-SSI (Navigli, 2008), which respectively exploited parallel texts and enriched semantic paths in a semantic network, the previously described UKB w2w system,[8] a knowledge-based WSD approach (Ponzetto and Navigli, 2010) which exploits an automatic extension of WordNet, and, as baseline, the MFS.

Fine-grained WSD. For the remaining fine-grained WSD datasets, i.e., Senseval-3 and SemEval-2007 task 17, we compared our approach with the previously described state-of-the-art systems UKB and IMS, and, as baseline, the MFS.

KORE50 and AIDA-CoNLL. For the KORE50 and AIDA-CoNLL datasets we compared our system with six approaches, including state-of-the-art ones (Hoffart et al., 2012; Cornolti et al., 2013):

- MW, i.e., the Normalized Google Distance as defined by Milne and Witten (2008);
- KPCS (Hoffart et al., 2012), which calculates a Mutual Information weighted vector of keyphrases for each candidate and then uses the cosine similarity to obtain candidates' scores;
- KORE and its variants KORE_{LSH-G} and KORE_{LSH-F} (Hoffart et al., 2012), based on similarity measures that exploit the overlap between phrases associated with the considered entities (KORE) and a hashing technique to reduce the space needed by the keyphrases associated with the entities (LSH-G, LSH-F);
- Tagme 2.0[9] (Ferragina and Scaiella, 2012) which uses the relatedness measure defined

[7] ./ukb_wsd -D dict.txt -K kb.bin --ppr_w2w ctx.txt

[8] We report the results as given by Agirre et al. (2014).

[9] We used the out-of-the-box RESTful API available at http://tagme.di.unipi.it

	Sens3	Sem07	SemEval-2013 English			French		German		Italian		Spanish	
System	WN	WN	WN	Wiki	BN	Wiki	BN	Wiki	BN	Wiki	BN	Wiki	BN
Babelfy	68.3	62.7	**65.9**	**87.4**	**69.2**	**71.6**	⋆56.9	81.6	**69.4**	**84.3**	66.6	**83.8**	69.5
IMS	**71.2**	63.3	65.7	–	–	–	–	–	–	–	–	–	–
UKB w2w	⋆65.3	⋆56.0	61.3	–	60.8	–	**60.8**	–	66.2	–	**67.3**	–	70.0
UMCC-DLSI	–	–	64.7	54.8	68.5	⋆60.5	60.5	⋆58.1	62.8	⋆58.3	65.8	⋆61.0	**71.0**
DAEBAK!	–	–	–	–	60.4	–	53.8	–	59.1	–	⋆61.3	–	60.0
GETALP-BN	–	–	51.4	–	58.3	–	48.3	–	52.3	–	52.8	–	57.8
MFS	70.3	**65.8**	⋆63.0	⋆80.3	⋆66.5	69.4	45.3	**83.1**	⋆67.4	82.3	57.5	82.4	⋆64.4
Babelfy unif. weights	67.0	65.2	65.0	87.0	68.5	71.9	57.2	81.2	69.8	83.7	66.8	83.8	70.8
Babelfy w/o dens. sub.	68.3	63.3	65.4	87.3	68.7	71.6	57.0	81.7	69.1	84.4	66.5	83.9	69.5
Babelfy only concepts	68.2	62.7	65.5	83.0	68.7	70.2	56.6	79.3	69.3	83.0	66.3	84.0	69.7
Babelfy on sentences	66.0	65.2	63.5	84.0	67.1	70.7	53.6	82.3	68.1	83.8	64.2	83.5	68.7

Table 1: F1 scores (percentages) of the participating systems of SemEval-2013 task 12 together with MFS, UKB w2w, IMS, our system and its ablated versions on the Senseval-3, SemEval-2007 task 17 and SemEval-2013 datasets. The first system which has a statistically significant difference from the top system is marked with $\star$ (χ^2, $p < 0.05$).

by Milne and Witten (2008) weighted with the commonness of a sense together with the keyphraseness measure defined by Mihalcea and Csomai (2007) to exploit the context around the target word;

- Illinois Wikifier[10] (Cheng and Roth, 2013) which combines local features, such as commonness and TF-IDF between mentions and Wikipedia pages, with global coherence features based on Wikipedia links and relational inference;
- DBpedia Spotlight[11] (Mendes et al., 2011) which uses LingPipe's string matching algorithm implementation together with a weighted cosine similarity measure to recognize and disambiguate mentions.

We also compared with UKB w2w, introduced above. Note that we could not use supervised systems, as the training data of AIDA-CoNLL covers less than half of the mentions used in the testing part and less than 10% of the entities considered in KORE50. To enable a fair comparison, we ran our system by restricting the BabelNet sense inventory of the target mentions to the English Wikipedia. As is customary in the literature, we calculated the systems' accuracy for both Entity Linking datasets.

[10]We used the out-of-the-box Java API available from http://cogcomp.cs.illinois.edu/page/download_view/Wikifier

[11]We used the 2011 version of DBpedia Spotlight as it obtains better scores on the considered datasets in comparison to the new version (Daiber et al., 2013). We used the out-of-the-box RESTful API available at http://spotlight.dbpedia.org

9 Results

Multilingual WSD. In Table 1 we show the F1 performance on the SemEval-2013 task 12 for the three setups: WordNet, Wikipedia and BabelNet. Using BabelNet we surpass all systems on English and German and obtain performance comparable with the best systems on two other languages (UKB on Italian and UMCC-DLSI on Spanish). Using the WordNet sense inventory, our results are on a par with the best system, i.e., IMS. On Wikipedia our results range between 71.6% (French) and 87.4% F1 (English), i.e., more than 10 points higher than the current state of the art (UMCC-DLSI) in all 5 languages. As for the MFS baseline, which is known to be very competitive in WSD (Navigli, 2009), we beat it in all setups except for German on Wikipedia. Interestingly, we surpass the WordNet MFS by 2.9 points, a significant result for a knowledge-based system (see also (Pilehvar and Navigli, 2014)).

Coarse- and fine-grained WSD. In Table 2, we show the results of the systems on the SemEval-2007 coarse-grained WSD dataset. As can be seen, we obtain the second best result after Ponzetto and Navigli (2010). In Table 1 (first two columns), we show the results of IMS and UKB on the Senseval-3 and SemEval-2007 task 17 datasets. We rank second on both datasets after IMS. However, the differences are not statistically significant. Moreover, Agirre et al. (2014, Table 5) note that using WordNet 3.0, instead of 1.7 or 2.1, to annotate these datasets can cause a more than one percent drop in performance.

System	F1
(Ponzetto and Navigli, 2010)	**85.5**
Babelfy	84.6
UoR-SSI	84.1
UKB w2w	83.6
NUS-PT	⋆82.3
MFS	77.4
Babelfy unif. weights	85.7
Babelfy w/o dens. sub.	84.9
Babelfy only concepts	85.3
Babelfy on sentences	82.3

Table 2: F1 score (percentages) on the SemEval-2007 task 7. The first system which has a statistically significant difference from the top system is marked with ⋆ (χ^2, $p < 0.05$).

Entity Linking. In Table 3 we show the results on the two Entity Linking datasets, i.e., KORE50 and AIDA-CoNLL. Our system outperforms all other approaches, with KORE-LSH-G getting closest, and Tagme and Wikifier lagging behind on the KORE50 dataset. For the AIDA-CoNLL dataset we obtain the third best performance after MW and KPCS, however the difference is not statistically significant.

We note the low performance of DBpedia Spotlight which, even if it achieves almost 100% precision on the identified mentions on both datasets, suffers from low recall due to its candidate identification step, confirming previous evaluations (Derczynski et al., 2013; Hakimov et al., 2012; Ludwig and Sack, 2011). This problem becomes even more accentuated in the latest version of this system (Daiber et al., 2013). Finally, UKB using BabelNet obtains low performance on EL, i.e., 19.4-10.5 points below the state of the art. This result is discussed below.

Discussion. The results obtained by UKB show that the high performance of our unified approach to EL and WSD is not just a mere artifact of the use of a rich multilingual semantic network, that is, BabelNet. In other words, it is not true that any graph-based algorithm could be applied to perform both EL and WSD at the same time equally well. This also shows that BabelNet by itself is not sufficient for achieving high performances for both tasks and that, instead, an appropriate processing of the structural and lexical information of the semantic network is needed. A manual analysis revealed that the main cause of error for UKB in the EL setup stems from its inability to enforce high coherence, e.g., by jointly disambiguating all the words, which is instead needed when considering the high level of ambiguity that we have in our semantic interpretation graph (Cucerzan, 2007). For instance, for sentence (1) in the introduction, UKB disambiguates *Thomas* as a cricket player and *Mario* as the popular video game rather than the two well-known soccer players, and *Munich* as the German city, rather than the soccer team in which they play. Our approach, instead, by enforcing highly coherent semantic interpretations, correctly identifies all the soccer-related entities.

System	KORE50	CoNLL
Babelfy	**71.5**	82.1
KORE-LSH-G	64.6	81.8
KORE	63.9	⋆80.7
MW	⋆57.6	**82.3**
Tagme	56.3	70.1
KPCS	55.6	82.2
KORE-LSH-F	53.2	81.2
UKB w2w (on BabelNet)	52.1	71.8
Illinois Wikifier	41.7	72.4
DBpedia Spotlight	35.4	34.0
Babelfy unif. weights	69.4	81.7
Babelfy w/o dens. sub.	62.5	78.1
Babelfy only NE	68.1	78.8

Table 3: Accuracy (percentages) of state-of-the-art EL systems and our system on KORE50 and AIDA-CoNLL. The first system with a statistically significant difference from the top system is marked with ⋆ (χ^2, $p < 0.05$).

In order to determine the need of our loose candidate identification heuristic (see Section 6), we compared the percentage of times a candidate set contains the correct entity against that obtained by an exact string matching between the mention and the sense inventory. On KORE50, our heuristic retrieves the correct entity 98.6% of the time vs. 42.4% when exact matching is used. This demonstrates the inadequacy of exact matching for EL, and the need for a comprehensive sense inventory, as is done in our approach.

We also performed different ablation tests by experimenting with the following variants of our system (reported at the bottom of Tables 1, 2 and 3):

- Babelfy using uniform distribution during the RWR to obtain the concepts' semantic signatures; this test assesses the impact of our weighting and edge creation strategy.

- Babelfy without performing the densest subgraph heuristic, i.e., when line 12 in Algorithm 2 is $G_I^\star = G_I$, so as to verify the impact of identifying the most coherent interpretations.
- Babelfy applied to the BabelNet subgraph induced by the entire set of named entity vertices, for the EL task, and that induced by word senses only, for the WSD task; this test aims to stress the impact of our unified approach.
- Babelfy applied on sentences instead of on whole documents.

The component which has a smaller impact on the performance is our triangle-based weighting scheme. The main exception is on the smallest dataset, i.e., SemEval-2007 task 17, for which this version attains an improvement of 2.5 percentage points.

Babelfy without the densest subgraph algorithm is the version which attains the lowest performances on the EL task, with a 9% performance drop on the KORE50 dataset, showing the need for a specially designed approach to cope with the high level of ambiguity that is encountered on this task. On the other hand, in the WSD datasets this version attains almost the same results as the full version, due to the lower number of candidate word senses.

Babelfy applied on sentences instead of on whole documents shows a lower performance, confirming the significance of higher semantic coherence on whole documents (notwithstanding the two exceptions on the SemEval-2007 task 17 and on the SemEval-2013 German Wikipedia datasets).

Finally, the version in which we restrict our system to named entities only (for EL) and concepts only (for WSD) consistently obtains lower results (notwithstanding the three exceptions on the Spanish SemEval-2013 task 12 using BabelNet and Wikipedia, and on the SemEval 2007 coarse-grained task). This highlights the benefit of our joint use of lexicographic and encyclopedic structured knowledge, on each of the two tasks. The 3.4% performance drop attained on KORE50 is of particular interest, since this dataset aims at testing performance on highly ambiguous mentions within short sentences. This indicates that the semantic analysis of small contexts can be improved by leveraging the coherence between concepts and named entities.

10 Conclusion

In this paper we presented Babelfy, a novel, integrated approach to Entity Linking and Word Sense Disambiguation, available at `http://babelfy.org`. Our joint solution is based on three key steps: i) the automatic creation of semantic signatures, i.e., related concepts and named entities, for each node in the reference semantic network; ii) the unconstrained identification of candidate meanings for all possible textual fragments; iii) linking based on a high-coherence densest subgraph algorithm. We used BabelNet 1.1.1 as our multilingual semantic network.

Our graph-based approach exploits the semantic network structure to its advantage: two key features of BabelNet, that is, its multilinguality and its integration of lexicographic and encyclopedic knowledge, make it possible to run our general, unified approach on the two tasks of Entity Linking and WSD in any of the languages covered by the semantic network. However, we also demonstrated that BabelNet in itself does not lead to state-of-the-art accuracy on both tasks, even when used in conjunction with a high-performance graph-based algorithm like Personalized PageRank. This shows the need for our novel unified approach to EL and WSD.

At the core of our approach lies the effective treatment of the high degree of ambiguity of partial textual mentions by means of a 2-approximation algorithm for the densest subgraph problem, which enables us to output a semantic interpretation of the input text with drastically reduced ambiguity, as was previously done with SSI (Navigli, 2008).

Our experiments on six gold-standard datasets show the state-of-the-art performance of our approach, as well as its robustness across languages. Our evaluation also demonstrates that our approach fares well both on long texts, such as those of the WSD tasks, and short and highly-ambiguous sentences, such as the ones in KORE50. Finally, ablation tests and further analysis demonstrate that each component of our system is needed to contribute state-of-the-art performances on both EL and WSD.

As future work, we plan to use Babelfy for information extraction, where semantics is taking the lead (Moro and Navigli, 2013), and for the validation of semantic annotations (Vannella et al., 2014).

Acknowledgments

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

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proc. of EACL*, pages 33–41.

Eneko Agirre, David Martínez, Oier López de Lacalle, and Aitor Soroa. 2006. Two graph-based algorithms for state-of-the-art WSD. In *Proc. of EMNLP*, pages 585–593.

Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2010. Graph-based Word Sense Disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896.

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proc. of ISWC/ASWC*, pages 722–735.

Carmen Banea and Rada Mihalcea. 2011. Word Sense Disambiguation with multilingual features. In *Proc. of IWCS*, pages 25–34.

Christoph Böhm, Gerard de Melo, Felix Naumann, and Gerhard Weikum. 2012. LINDA: distributed web-of-data-scale entity matching. In *Proc. of CIKM*, pages 2104–2108.

Samuel Brody and Mirella Lapata. 2009. Bayesian Word Sense Induction. In *Proc. of EACL*, pages 103–111.

Razvan C. Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proc. of EACL*, pages 9–16.

Yee Seng Chan, Hwee Tou Ng, and Zhi Zhong. 2007. NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. In *Proc. of SemEval-2007*, pages 253–256.

Moses Charikar. 2000. Greedy approximation algorithms for finding dense components in a graph. In *Proc. of APPROX*, pages 84–95.

Xiao Cheng and Dan Roth. 2013. Relational Inference for Wikification. In *Proc. of EMNLP*, pages 1787–1796.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proc. of WWW*, pages 249–260.

Silviu Cucerzan. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proc. of EMNLP-CoNLL*, pages 708–716.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proc. of I-Semantics*, pages 121–124.

Leon Derczynski, Diana Maynard, Niraj Aswani, and Kalina Bontcheva. 2013. Microblog-genre noise and impact on semantic annotation accuracy. In *Proc. of Hypertext*, pages 21–30.

Antonio Di Marco and Roberto Navigli. 2013. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. *Computational Linguistics*, 39(3):709–754.

Nicolai Erbs, Torsten Zesch, and Iryna Gurevych. 2011. Link discovery: A comprehensive analysis. In *Proc. of ICSC*, pages 83–86.

Oren Etzioni, Michele Banko, and Michael J Cafarella. 2006. Machine Reading. In *Proc. of AAAI*, pages 1517–1519.

Uriel Feige, Guy Kortsarz, and David Peleg. 1999. The dense k-subgraph problem. *Algorithmica*, 29:2001.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). In *Proc. of CIKM*, pages 1625–1628.

Paolo Ferragina and Ugo Scaiella. 2012. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. *IEEE Software*, 29(1):70–75.

Weiwei Guo and Mona T. Diab. 2010. Combining Orthogonal Monolingual and Multilingual Sources of Evidence for All Words WSD. In *Proc. of ACL*, pages 1542–1551.

Yoan Gutiérrez, Yenier Castañeda, Andy González, Rainel Estrada, Dennys D. Piug, Jose I. Abreu, Roger Pérez, Antonio Fernández Orquín, Andrés Montoyo, Rafael Muñoz, and Franc Camara. 2013. UMCC_DLSI: Reinforcing a Ranking Algorithm with Sense Frequencies and Multidimensional Semantic Resources to solve Multilingual Word Sense Disambiguation. In *Proc. of SemEval-2013*, pages 241–249.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating Entity Linking with Wikipedia. *Artificial Intelligence*, 194:130–150.

Sherzod Hakimov, Salih Atilay Oto, and Erdogan Dogdu. 2012. Named entity recognition and disambiguation using linked data and graph-based centrality scoring. In *Proc. of SWIM*, pages 4:1–4:7.

Taher H. Haveliwala. 2002. Topic-sensitive PageRank. In *Proc. of WWW*, pages 517–526.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proc. of EMNLP*, pages 782–792.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *Proc. of CIKM*, pages 545–554.

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61.

Eduard H. Hovy, Roberto Navigli, and Simone P. Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Samir Khuller and Barna Saha. 2009. On finding dense subgraphs. In *Proc. of ICALP*, pages 597–608.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective Annotation of Wikipedia Entities in Web Text. In *Proc. of KDD*, pages 457–466.

Els Lefever and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual Word Sense Disambiguation. In *Proc. of SemEval-2010*, pages 15–20.

Els Lefever and Véronique Hoste. 2013. SemEval-2013 Task 10: Cross-lingual Word Sense Disambiguation. In *Proc. of SemEval-2013*, pages 158–166.

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for Word Sense Disambiguation. In *Proc. of ACL-HLT*, pages 317–322.

Michael E. Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proc. of the International Conference on Systems Documentation*, pages 24–26.

Nadine Ludwig and Harald Sack. 2011. Named entity recognition for user-generated tags. In *Proc. of DEXA*, pages 177–181.

Suresh Manandhar, Ioannis P. Klapaftis, Dmitriy Dligach, and Sameer S. Pradhan. 2010. SemEval-2010 task 14: Word sense induction & disambiguation. In *Proc. of SemEval-2010*, pages 63–68.

Steve L. Manion and Raazesh Sainudiin. 2013. DAEBAK!: Peripheral Diversity for Multilingual Word Sense Disambiguation. In *Proc. of SemEval-2013*, pages 250–254.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proc. of I-Semantics*, pages 1–8.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proc. of CIKM*, pages 233–242.

Rada Mihalcea and Dan I Moldovan. 2001. Extended WordNet: Progress report. In *Proc. of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100.

Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proc. of HLT/EMNLP*, pages 411–418.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proc. of HLT*, pages 303–308.

Tristan Miller, Chris Biemann, Torsten Zesch, and Iryna Gurevych. 2012. Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation. In *Proc. of COLING*, pages 1781–1796.

David Milne and Ian H. Witten. 2008. Learning to link with Wikipedia. In *Proc. of CIKM*, pages 509–518.

Andrea Moro and Roberto Navigli. 2013. Integrating Syntactic and Semantic Analysis into the Open Information Extraction Paradigm. In *Proc. of IJCAI*, pages 2148–2154.

Roberto Navigli and Mirella Lapata. 2010. An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *TPAMI*, 32(4):678–692.

Roberto Navigli and Simone Paolo Ponzetto. 2012a. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli and Simone Paolo Ponzetto. 2012b. Joining forces pays off: Multilingual Joint Word Sense Disambiguation. In *Proc. of EMNLP*, pages 1399–1410.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 Task 07: Coarse-Grained English All-Words Task. In *Proc. of SemEval-2007*, pages 30–35.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proc. of SemEval-2013*, pages 222–231.

Roberto Navigli. 2008. A structural approach to the automatic adjudication of word sense disagreements. *Natural Language Engineering*, 14(4):293–310.

Roberto Navigli. 2009. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Roberto Navigli. 2012. A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In *Proc. of SOFSEM*, pages 115–129.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proc. of ACL*, pages 40–47.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A Large-scale Pseudoword-based Evaluation Framework for State-of-the-Art Word Sense Disambiguation. *Computational Linguistics*.

Simone P. Ponzetto and Roberto Navigli. 2010. Knowledge-rich Word Sense Disambiguation rivaling supervised system. In *Proc. of ACL*, pages 1522–1531.

Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proc. of SemEval-2007*, pages 87–92. Association for Computational Linguistics.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity Linking: Finding Extracted Entities in a Knowledge Base. In *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing, pages 93–115. Springer Berlin Heidelberg.

Lev-Arie Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proc. of ACL*, pages 1375–1384.

Lenhart K. Schubert. 2006. Turing's dream and the knowledge challenge. In *Proc. of NCAI*, pages 1534–1538.

Didier Schwab, Andon Tchechmedjiev, Jérôme Goulian, Mohammad Nasiruddin, Gilles Sérasset, and Hervé Blanchon. 2013. GETALP System: Propagation of a Lesk Measure through an Ant Colony Algorithm. In *Proc. of SemEval-2013*, pages 232–240.

Hui Shen, Razvan Bunescu, and Rada Mihalcea. 2013. Coarse to Fine Grained Sense Disambiguation in Wikipedia. In *Proc. of *SEM*, pages 22–31.

Ravi Sinha and Rada Mihalcea. 2007. Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. In *Proc. of ICSC*, pages 363–369.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proc. of Senseval-3*, pages 41–43.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an Affective Extension of WordNet. In *Proc. of LREC*, pages 1083–1086.

Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. 2006. Fast Random Walk with Restart and Its Applications. In *Proc. of ICDM*, pages 613–622.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of NAACL-HLT*, pages 173–180.

George Tsatsaronis, Michalis Vazirgiannis, and Ion Androutsopoulos. 2007. Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri. In *Proc. of IJCAI*, pages 1725–1730.

Tim Van de Cruys and Marianna Apidianaki. 2011. Latent Semantic Word Sense Induction and Disambiguation. In *Proc. of ACL*, pages 1476–1485.

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. 2014. Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. In *Proc. of ACL*.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–10.

Zhi Zhong and Hwee Tou Ng. 2010. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. In *Proc. of ACL (Demo)*, pages 78–83.

Amal Zouaq, Michel Gagnon, and Benoit Ozell. 2009. A SUMO-based Semantic Analysis for Knowledge Extraction. In *Proc of LTC*.