Annotating the MASC Corpus with BabelNet

Andrea Moro[†], **Roberto Navigli**[†], **Francesco Maria Tucci**[†], **Rebecca J. Passonneau**[‡]
[†] Dipartimento di Informatica, Sapienza University of Rome, Roma, Italy
{moro,navigli}@di.uniroma1.it, francesco.tucci@yahoo.it
[‡] Center for Computational Learning Systems, Columbia University, New York, NY USA
becky@columbia.edu

Abstract

In this paper we tackle the problem of automatically annotating, with both word senses and named entities, the MASC 3.0 corpus, a large English corpus covering a wide range of genres of written and spoken text. We use BabelNet 2.0, a multilingual semantic network which integrates both lexicographic and encyclopedic knowledge, as our sense/entity inventory together with its semantic structure, to perform the aforementioned annotation task. Word sense annotated corpora have been around for more than twenty years, helping the development of Word Sense Disambiguation algorithms by providing both training and testing grounds. More recently Entity Linking has followed the same path, with the creation of huge resources containing annotated named entities. However, to date, there has been no resource that contains both kinds of annotation. In this paper we present an automatic approach for performing this annotation, together with its output on the MASC corpus. We use this corpus because its goal of integrating different types of annotations goes exactly in our same direction. Our overall aim is to stimulate research on the joint exploitation and disambiguation of word senses and named entities. Finally, we estimate the quality of our annotations using both manually-tagged named entities and word senses, obtaining an accuracy of roughly 70% for both named entities and word sense annotations.

Keywords: Semantic Annotation, Named Entities, Word Senses, Lexical Ambiguity, Semantic Network, Disambiguation

1. Introduction

One of the key tasks of Artificial Intelligence is the automatic understanding of the meaning of text, i.e., Machine Reading (Etzioni et al., 2006). To tackle this problem an important aspect to be pursued is obtaining a correct and, as far as possible, complete representation of both general-purpose and domain-specific knowledge. However, encoding knowledge is a very onerous task, which cannot be performed manually with high accuracy on a large scale. The recent upsurge of interest in the use of semi-structured resources to create novel repositories of knowledge (Hovy et al., 2013) has opened up new opportunities for wide-coverage, general-purpose Natural Language Understanding techniques. Moreover, some of these knowledge repositories integrate both lexicographic and encyclopedic knowledge (Navigli and Ponzetto, 2012a). Given structured resources such as these the logical next step from the point of view of Machine Reading is to link natural language text to them. To this end, the research community has exploited already existing datasets for tasks such as Information and Relation Extraction, e.g., the MUC and ACE datasets, and has developed semi-automatic and manual methods for building semantically annotated datasets (Névéol et al., 2011; Basile et al., 2012). However, no dataset to date has addressed both kinds of lexicographic (i.e., word senses) and encyclopedic knowledge (i.e., Named Entities) a task which on a large scale is hampered by the so-called knowledge acquisition bottleneck (Pilehvar and Navigli, 2014).

In this paper we fill this gap by performing two Natural Language Processing tasks which link raw text to a structured repository of knowledge: Word Sense Disambiguation (WSD) (Navigli, 2009), i.e., the task of determining the sense of a word in a given context, and Entity Linking (EL) (Rao et al., 2013), i.e., the task of discovering which named entities are mentioned in a text. The two main differences between WSD and EL consist in the kind of inventory used, i.e., dictionary vs. encyclopedia, and the assumption that the mention is complete or potentially partial, respectively. Notwithstanding these differences, the tasks are pretty similar in nature, in that they both involve the disambiguation of textual mentions according to a reference inventory. However, the research community has tackled the two tasks separately, often duplicating efforts and solutions. Recently, Moro et al. (2013) have shown that word senses can be used to improve entity relation extraction performance both in terms of precision and accuracy, showing that these two kinds of annotation can be effectively exploited together.

We present a first attempt at the automatic semantic annotation of textual resources with both named entities and word senses. We use BabelNet 2.0 as our sense/entity inventory, together with its semantic structure, for effectively performing the automatic annotation task as described by Moro et al. (2014). We decided to annotate the MASC 3.0 corpus (Ide et al., 2008), because the rationale underlying the organization of this resource goes in exactly the same direction as our idea of integrating different kinds of semantic annotation. Combining these annotations therefore offers the potential for increasing the amount of multilevel information available to a level approaching that which we humans exploit in order to interpret text. Thanks to the application of our integrated WSD and EL approach, we add a complete word sense and named entity annotation on top of the annotations already available in MASC. Evaluations of the annotation quality are carried out using the existing manual word sense annotations available in the MASC corpus and a random sample of automatically annotated Named Entities.

The paper is organized as follows: in Section 2 we cover datasets and methods for annotating text with senses and entities, we then introduce the MASC 3.0 corpus in Section

3 and BabelNet in Section 4, we describe our joint approach to WSD and EL in Section 5 and provide statistics and evaluations of our MASC annotations in Sections 6 and 7.

2. Related Work

The WSD task, i.e., the task of determining the sense of a word in a given context, has been investigated for more than fifty years (Navigli, 2009; Navigli, 2012) and many different datasets are available. One long-lasting example is the Senseval/SemEval competition series, thanks to which, during the last fifteen years, many datasets for WSD were released. However, being manually created, these datasets contain few annotations (typically around one-two thousand), hence limiting the scale of the evaluations. Another well-known resource for WSD is the SemCor dataset created at Princeton University by Miller et al. (1993). This contains roughly 700k words, of which 240k are manually annotated with WordNet senses. However, because it links to WordNet senses, this resource contains only lexicographic annotations without considering named entities.

On the other hand, the EL task, i.e., the task of discovering which named entities are mentioned in a text, has become a leading task more recently, see Rao et al. (2013) for a recent survey. During the last two years Google has released two datasets containing semantically annotated web pages with named entities (Singh et al., 2012; Gabrilovich et al., 2013). The first of these is Wikilinks, which is the result of a web crawl on roughly eleven million webpages looking for hyperlinks to Wikipedia pages, i.e., the manual annotations are the hyperlinks found in webpages. The second is the Freebase Annotations of the ClueWeb Corpora v1 (FACC1), which consists of an automatic annotation of roughly 400 million web documents totaling more than six billion annotated mentions of entities from Freebase (Bollacker et al., 2008) with an estimated precision around 80-85% and an estimated recall around 70-85%. However, these resources contain only named entities without taking into account word senses.

State-of-the-art approaches for knowledge-based Word Sense Disambiguation (Ponzetto and Navigli, 2010; Navigli and Ponzetto, 2012b; Agirre et al., 2014) exploit the semantic relations between word senses, as found for example in WordNet or enriched versions of it, to run graph-based algorithms and obtain a ranking over the candidate senses. However, semantic networks containing named entities are an order of magnitude larger than lexical repositories such as WordNet, causing unavoidable slowdowns. For the task of Entity Linking, instead, the most common approach is to consider the subgraph induced by the named entity candidates in the considered knowledge base and then select the candidates that are more connected to each other by means of semantic relations (Ploch, 2011; Hoffart et al., 2011; Rao et al., 2013; Cornolti et al., 2013). However, EL falls short when the input text lacks sufficient encyclopedic context. In recent work, Moro et al. (2014) have proposed a unified approach to WSD and EL which overcomes the above limits by taking advantage of the joint structural information provided by a large semantic network at both the lexicographic and encyclopedic level. In this paper we apply this algorithm to create a large-scale corpus semantically annotated with both word senses and named entities.

3. MASC 3.0

We use the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008) which consists of parts of the American National Corpus covering a wide range of genres of written and spoken textual data amounting to over 500k words. This project aims at organizing and addressing the problems arising against the creation of a resource with multiple annotations. The corpus is available in different formats such as GrAF, in-line XML, token/part of speech sequences, RDF encoding and CoNLL format. The key feature of this corpus is the availability within a single resource of many different linguistic annotations; to date, it contains 17 different types of linguistic annotation, such as sentence boundary, part of speech and syntactic dependency among others. These annotations are the result of a semi-automatic effort in which automatic systems have been coupled with an iterative process of manual evaluations and annotations for retraining the automatic approaches and finetuning annotator guidelines to improve inter-annotator agreement. Moreover, the fact that it is freely available[1] makes it an invaluable resource for both industry and academic communities in order to produce and improve cutting-edge language technologies. Another reason for our use of this corpus is that it already contains around 3k word sense instances manually disambiguated for 53 distinct words (Passonneau et al., 2010; Passonneau et al., 2012). We exploit these manual annotations to show the quality of our automatic annotation.

4. BabelNet

To perform the automatic annotation of the MASC 3.0 corpus with word senses and named entities, we exploit BabelNet[2] (Navigli and Ponzetto, 2012a), a large-scale multilingual semantic network created from the algorithmic integration of Wikipedia[3] and WordNet. Its core idea is the automatic mapping between Wikipedia pages and WordNet synsets that represent the same concept. For instance, the concept *hot-air balloon* is defined both in Wikipedia and WordNet. The wide coverage of Wikipedia enables BabelNet to cover novel word senses such as *cluster ballooning* and novel named entities such as *Montgolfier brothers* which are not available in WordNet, while at the same time having fine-grained sense differences thanks to WordNet. The nodes of this semantic network are called Babel synsets and can contain Wikipedia page titles, WordNet synsets and OmegaWiki senses both by themselves or, in the case of the same concept or named entity being covered by more than one resource, within a same Babel synset. For instance, for the concept *plane*, we have a Babel synset containing a reference to the Wikipedia page "Fixed-wing aircraft" and one to the WordNet synset *plane#n#1* together with their synonyms, glosses and translations as

[1] MASC 3.0 and all its previous versions are freely available from http://www.anc.org/data/masc/

[2] http://babelnet.org

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

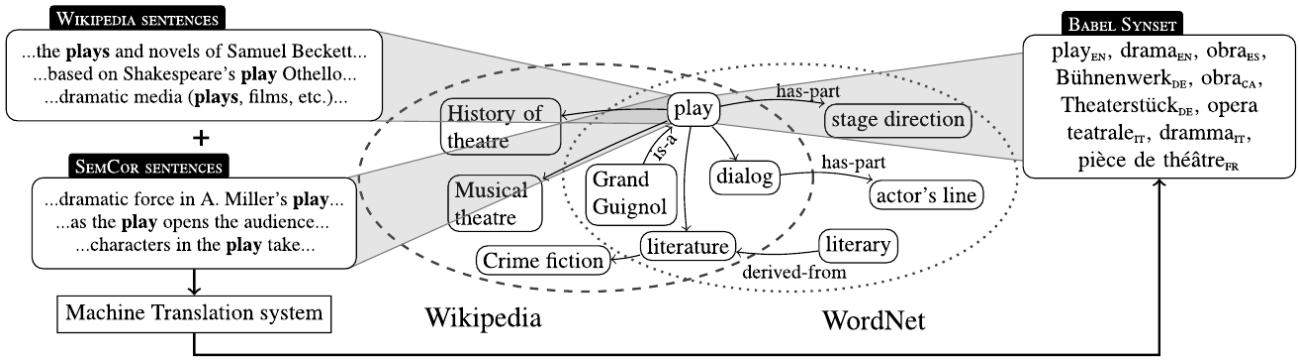

Figure 1: An excerpt of the BabelNet semantic network.

contained in the respective resources. Moreover, BabelNet relates these elements with semantic relations automatically extracted from WordNet and Wikipedia, such as hypernymy, meronymy and semantic relatedness resulting in a rich structured resource containing both encyclopedic and lexicographic knowledge. The latest release of BabelNet, i.e., version 2.0, covers 50 languages and integrates the Open Multilingual WordNet (Bond and Foster, 2013) and OmegaWiki[4] into the network. This multilingual semantic network contains roughly 9M synsets, 50M senses and 250M semantic relation instances. In this latest version, BabelNet has even been released in RDF format so as to enable its connection within the Linked Open Data cloud (Ehrmann et al., 2014).

5. Word Sense Disambiguation and Entity Linking

In this section we describe our approach to Word Sense Disambiguation and Entity Linking. A thorough description of the approach is provided by Moro et al. (2014). Our approach is divided into four steps. The first step deals with the discovery of textual mentions within the input text, i.e., substrings of the text for which at least one candidate named entity or word sense can be found in BabelNet. The second step consists of connecting the candidate meanings by means of pre-computed random walks with restart on BabelNet. The third step regards a reduction of the size of the semantic interpretation graph by means of a dense subgraph extraction routine. The last step regards the selection of the most appropriate candidate for each textual mention and the computation of a confidence score for each of them. We use this approach to perform the automatic annotation of the MASC corpus with word senses and named entities as this is a state-of-the-art system in both WSD and EL (Moro et al., 2014).

5.1. Mentions and Candidates

To perform the first step we first obtain the part-of-speech (POS) tags of the input document[5]. Next, we discover mentions and their respective candidates using the psedocode in Algorithm 1. We consider each substring of maximum length M in the POS-tagged document and check if it is present as an English lexicalization of some Babel synset in BabelNet (see lines 4–6). As a result of this procedure we obtain, for each substring, a list of candidate Babel synsets. We enforce POS-coherence by requiring that at least one POS tag of the words of the considered substring matches the tags of the candidate synsets (see line 8). We remark that we allow the recognition of overlapping mentions. For instance, for the multi-word expression *bus driver*, we obtain three nominal mentions, i.e., *bus, bus driver,* and *driver* and the corresponding candidate synsets.

Algorithm 1 Discover mentions and candidates.

```
1:  input: d, the input document;
           cand, a map from the lexicalizations in
           BabelNet to the Babel synsets.
2:  output: cand, a map from the mentions to the set
            of candidates.
3:  function MENTCAND(d, cand)
4:      List<WordLemmaTag> t := POStag(d);
5:      for each substring s of maximum length M in t do
6:          List<BabelSynset> lb := cand.get(s);
7:          for each Babel synset b in lb do
8:              if b.pos() is equal to s.pos() then
9:                  cand.get(s).add(b)
10:     return cand
```

5.2. Semantic Relations between Candidates

As a result of the previous step we obtain a list of mentions together with their respective candidate meanings. In this step we focus on the relations between the candidates of different mentions so as to give a better semantic characterization of their relevance within the considered context. To obtain a good semantic characterization of the candidate meanings we exploit the semantic relations within BabelNet by performing random walk with restart (RWR) and obtain sets of semantically relevant concepts and named entities for each node of the network.

More formally, for each node c of the considered semantic network we simulate a RWR, as shown in Algorithm 2, and obtain a weighted distribution over the other nodes of the semantic network, i.e. $synDistr_c$. Starting from a node c the RWR will either choose one of its neighbors uniformly at random and then continue, or it will restart at the starting

[4] http://www.omegawiki.org

[5] We exploit the POS tag annotations of the MASC 3.0 corpus although in general any POS tagger should work.

Algorithm 2 Random walk with restart.

input: c, the starting vertex;
α, the restart probability;
n, the number of steps to be executed;
P, the transition probabilities.
output: $synDistr_c$, distribution over the nodes of the graph for c.
function RWR(c, α, n, P)
  $v := c$
  $synDistr_c :=$ new Map<BabelSynset,Integer> ()
  while $n > 0$ **do**
    if *random()* $> \alpha$ **then**
      // continue the walk
      choose a neighbor v' of the current node v uniformly at random
      $synDistr_c.get(v')$++
      $v := v'$
    else
      // restart the walk
      $v := c$
    $n := n - 1$
  return $synDistr_c$

node c. While the RWR is simulated we keep track of the number of times it visits each synset. By taking the synsets that are visited a number of times above a fixed threshold θ we obtain a new set of semantic relations going from the starting node to highly related nodes. In this manner we are able to add edges between our candidate synsets obtaining a semantic graph $G_I := (V_I, E_I)$ containing all the pairs substring/candidate, (s, c), as nodes and all the high-quality semantic relations as edges ((s, c) and (s', c') are connected iff the candidate c is semantically related to c', i.e., iff $synDistr_c.get(c') > \theta$).

The rationale behind this step is that of building a single graph containing all the possible semantic interpretations of the input text for both word senses and named entities, so as to obtain a better and more detailed semantic context within which to perform the disambiguation step. Finally, we want to emphasise that this is a preliminary step which needs to be performed only once, independently of the input text.

5.3. Linking by Densest Subgraph

In this section we illustrate how we reduce the initial size of our semantic interpretation graph. We perform this step to further improve the quality of the considered semantic context. To do this we developed a novel densest subgraph algorithm. The main idea here is that the most suitable meanings of each mention will belong to the densest area of the graph.

The problem of identifying the densest subgraph of size at least k is NP-hard. Therefore, we define a heuristic for k-partite graphs inspired by a 2-approximation greedy algorithm for general graphs (Charikar, 2000; Khuller and Saha, 2009). Our adapted strategy for selecting the densest subgraph of G_I is based on the iterative removal of low-coherence vertices, i.e., mention 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 mention f_{max}, i.e., the one with the maximum number of candidate meanings in the graph (see line 7). Next, we discard the weakest interpretation of the current mention f_{max}. To do so, we determine the lexical and semantic coherence of each candidate meaning (v, f_{max}) using Formula 1 (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).

We then move to the next step, i.e., we set $t := t+1$ and repeat the low-coherence removal step (see line 16). We stop when the number of remaining candidates for each mention is below a threshold, 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 $G_I^{(t)}$ that maximizes the average degree (see lines 14–15).

Algorithm 3 Densest subgraph.

1: **input:** F, the set of all mentions in the input text;
$cand$, the map from mentions 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** $G_I^\star$

5.4. Candidate Disambiguation

At this point we have a dense graph $G_I^\star$ representing the most coherent semantic interpretations of the given input text in terms of (mention, candidate) pairs, i.e., a node for each pair of substring/synset candidate, and semantic relations between them, i.e., the edges obtained with the RWR. We now select the most appropriate candidate for each mention and compute a confidence score. For each node (s, c) in $G_I^\star$ we compute the number of different substrings that it relates to and then we multiply it by its degree

# Documents		392
# Content Words	305,960	
# Non-Content Words	286,512	
# Words		592,472
# Adjective Word Senses	30,015	
# Adverb Word Senses	23,685	
# Noun Word Senses	131,688	
# Verb Word Senses	82,489	
# Word Senses		267,877
# Named Entities		18,539
Total number of annotations		286,416

Table 1: Statistics of the MASC 3.0 corpus and of our automatic annotation.

in $G_I^\star$:

$$ds_{(s,c)} := \frac{|\{s' \in S : \exists c' \; s.t.((s,c),(s',c')) \text{ or } ((s',c'),(s,c)) \in E_I\}|}{|S|-1} \cdot deg((s,c)) \quad (1)$$

where $S = \{s : \exists\,(s,c) \in V_I\}$ is the set of all the substrings extracted from the document and E_I is the set of edges of $G_I^\star$. The rationale behind this scoring function is to take into account both the semantic coherence, using a graph centrality measure among the candidate meanings such as the degree, and the lexical coherence, in terms of the number of substrings that a candidate relates to. Given a substring s and a set of candidates $C = \{c : \exists\,(s,c) \in E_I\}$ we select the most suitable candidate as the one that maximizes the ds score:

$$dis(s) = \arg\max_{c \in C} \; ds_{(s,c)}$$

In this manner we can disambiguate all the considered substrings. To assign a confidence score between 0 and 1 to the disambiguated entry we compute the number of substrings that each disambiguated entry connects to, and we normalize it by the total number of substrings found in the document.

6. Statistics

We now present the statistics of our automatically-annotated dataset. We annotated 592K words of running text in 392 documents of the MASC corpus. Overall, we extracted from the corpus 286K mentions (including 30K unique mentions). Among these 286K mentions, we have 41K unambiguous mentions and 18K multiword expressions.

In Table 1, we show, together with the number of words and documents in MASC 3.0, the number of annotated named entities and word senses disambiguated by our system. We can see that we disambiguate most of the mentions within the corpus. The average polysemy of the disambiguated mentions is roughly 9 for nouns, 5 for verbs, 5 for adjectives and 4 for adverbs.

In Figure 2 we show the distribution of the disambiguated word senses by part of speech (excluding named entities which are names by definition); as expected, there is a greater number of nouns and verbs.

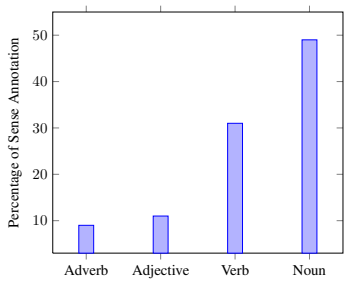

Figure 2: Distribution of Word Senses by Part of Speech

7. Evaluations

To evaluate our approach we used the accuracy measure, which is defined as the number of correct senses/entities over the whole number of manually annotated substrings. We performed two evaluations to estimate the quality of the automatically created semantic annotations. As regards, named entities we manually evaluated a random sample of $1,000$ entities obtaining an estimated accuracy of 72.4% for our annotations. As regards evaluating the quality of word senses, instead, we exploited the MASC word sense sentence annotation corpus[6] (Passonneau et al., 2010), which consists of the manual annotations performed by multiple judges of 53 words in roughly $3,000$ sentences. By using this dataset we calculate a word sense disambiguation accuracy of 54.5%. Note, however, that each of those 53 words is ambiguous, hence the dataset provides a lowerbound estimate for the word sense annotation quality of our approach. To estimate general accuracy on the full set of annotations, we randomly sampled 500 word tokens from the MASC corpus and manually validated the automatic annotations of these items, obtaining 68.8% accuracy.

8. Conclusion

In this paper we presented a joint word senses disambiguation and entity linking of the MASC corpus. We performed the disambiguation of all the word senses and named entities as found in the MASC 3.0 corpus using BabelNet 2.0, a multilingual semantic network that integrates encyclopedic and lexicographic knowledge automatically extracted from WordNet, Wikipedia and OmegaWiki, as our word sense/named entity inventory. To perform this automatic annotation we exploited a novel joint disambiguation system for word senses and named entities. Moreover, to validate the quality of our annotations we evaluated both manually annotated named entities and word senses, with an estimated accuracy of 72.4% for named entities and 68.8% for word sense annotations. Finally, following the MASC licensing philosophy, our automatically generated semantic annotations are accessible at `http://lcl.uniroma1.it/MASC-NEWS`.

Acknowledgments

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

[6] We used the file masc_wordsense.zip downloadable from http://www.anc.org/MASC/download/

9. References

Agirre, E., de Lacalle, O. L., and Soroa, A. (2014). Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, 40(1).

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). Developing a large semantically annotated corpus. In *Proc. of LREC*, volume 12, pages 3196–3200.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. of SIGMOD*, pages 1247–1250.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proc. of ACL*, pages 1352–1362.

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

Cornolti, M., Ferragina, P., and Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. In *Proc. of WWW*, pages 249–260.

Ehrmann, M., Cecconi, F., Vannella, D., M^{c}Crae, J., Cimiano, P., and Navigli, R. (2014). Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In *Proc. of LREC*.

Etzioni, O., Banko, M., and Cafarella, M. J. (2006). Machine Reading. In *Proc. of AAAI*, pages 1517–1519.

Gabrilovich, E., Ringgaard, M., and Subramanya, A. (2013). FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format vers. 1, Correction level 0).

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust Disambiguation of Named Entities in Text. In *Proc. of EMNLP*, pages 782–792.

Hovy, E. H., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Ide, N., Baker, C., Fellbaum, C., Fillmore, C., and Passonneau, R. (2008). MASC: the Manually Annotated Sub-Corpus of American English. In *Proc. of LREC*.

Khuller, S. and Saha, B. (2009). On Finding Dense Subgraphs. In *Proc. of ICALP*, pages 597–608.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Proc. of HLT*, pages 303–308.

Moro, A., Li, H., Krause, S., Xu, F., Navigli, R., and Uszkoreit, H. (2013). Semantic Rule Filtering for Web-Scale Relation Extraction. In *Proc. of ISWC*.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics*.

Navigli, R. and Ponzetto, S. P. (2012a). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. and Ponzetto, S. P. (2012b). Joining Forces Pays Off: Multilingual Joint Word Sense Disambiguation. In *Proc. of EMNLP*, pages 1399–1410, Jeju, Korea.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):1–69.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pages 115–129.

Névéol, A., Doğan, R. I., and Lu, Z. (2011). Semi-automatic semantic annotation of PubMed queries: A study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics*, 44(2):310–318.

Passonneau, R. J., Salleb-Aoussi, A., Bhardwaj, V., and Ide, N. (2010). Word Sense Annotation of Polysemous Words by Multiple Annotators. In *Proc. of LREC*.

Passonneau, R. J., Baker, C., Fellbaum, C., and Ide, N. (2012). The MASC word sense sentence corpus. In *Proc. of LREC*.

Pilehvar, M. T. and Navigli, R. (2014). A Large-scale Pseudoword-based Evaluation Framework for State-of-the-Art Word Sense Disambiguation. *Computational Linguistics*.

Ploch, D. (2011). Exploring Entity Relations for Named Entity Disambiguation. In *Proc. of ACL*, pages 18–23.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich Word Sense Disambiguation rivaling supervised system. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531, Uppsala, Sweden.

Rao, D., McNamee, P., and Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 93–115. Springer.

Singh, S., Subramanya, A., Pereira, F., and McCallum, A. (2012). Wikilinks: A Large-scale Cross-Document Coreference Corpus Labeled via Links to Wikipedia. Technical Report UM-CS-2012-015.