# SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking

**Andrea Moro** and **Roberto Navigli**
Dipartimento di Informatica,
Sapienza Università di Roma,
Viale Regina Elena 295, 00161 Roma, Italy
{moro,navigli}@di.uniroma1.it

## Abstract

In this paper we present the Multilingual All-Words Sense Disambiguation and Entity Linking task. Word Sense Disambiguation (WSD) and Entity Linking (EL) are well-known problems in the Natural Language Processing field and both address the lexical ambiguity of language. Their main difference lies in the kind of meaning inventories that are used: EL uses encyclopedic knowledge, while WSD uses lexicographic information. Our aim with this task is to analyze whether, and if so, how, using a resource that integrates both kinds of inventories (i.e., BabelNet 2.5.1) might enable WSD and EL to be solved by means of similar (even, the same) methods. Moreover, we investigate this task in a multilingual setting and for some specific domains.

## 1 Introduction

The Senseval and SemEval evaluation series represent key moments in the community of computational linguistics and related areas. Their focus has been to provide objective evaluations of methods within the wide spectrum of semantic techniques for tasks mainly related to automatic text understanding.

Through SemEval-2015 task 13 we both continue and renew the longstanding tradition of disambiguation tasks, by addressing multilingual WSD and EL in a joint manner. WSD (Navigli, 2009; Navigli, 2012) is a historical task aimed at explicitly assigning meanings to single-word and multi-word occurrences within text, a task which today is more alive than ever in the research community. EL (Erbs et al., 2011; Cornolti et al., 2013; Rao et al., 2013) is a more recent task which aims at discovering mentions of entities within a text and linking them to the most suitable entry in a knowledge base. Both these tasks aim at handling the inherent ambiguity of natural language, however WSD tackles it from a lexicographic perspective, while EL tackles it from an encyclopedic one. Specifically, the main difference between the two tasks lies in the kind of inventory they use. For instance, WordNet (Miller et al., 1990), a manually curated semantic network for the English language, has become the main reference inventory for English WSD systems thanks to its wide coverage of verbs, adverbs, adjectives and common nouns. More recently, Wikipedia has been shown to be an optimal resource for recovering named entities, and has consequently become - together with all its semi-automatic derivations such as DBpedia (Auer et al., 2007) and Freebase (Bollacker et al., 2008) - the main reference inventory for EL systems.

Over the years, the research community has typically focused on each of these tasks separately. Recently, however, joint approaches have been proposed (Moro et al., 2014b). One of the reasons for pursuing the unification of these tasks derives from the current trend in knowledge acquisition which consists of the seamless integration of encyclopedic and lexicographic knowledge within structured language resources (Hovy et al., 2013). A case in point here is BabelNet[1], a multilingual semantic network and encyclopedic dictionary (Navigli and Ponzetto, 2012). Resources like BabelNet provide a common ground for the tasks of WSD and EL.

[1] http://babelnet.org

*Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

In this task our goal is to promote research in the direction of joint word sense and named entity disambiguation, so as to concentrate research efforts on the aspects that differentiate these two tasks without duplicating research on common problems such as identifying the right meaning in context. However, we are also interested in systems that perform only one of the two tasks, and even systems which tackle one particular setting of WSD, such as all-words sense disambiguation vs. any subset of part-of-speech tags. Moreover, given the recent upsurge of interest in multilingual approaches, we developed the task dataset in three different languages (English, Italian and Spanish) on parallel texts which have been independently and manually annotated by different native/fluent speakers. In contrast to the SemEval-2013 task 12 on Multilingual Word Sense Disambiguation (Navigli et al., 2013), our focus in task 13 is to present a dataset containing both kinds of inventories (i.e., named entities and word senses) in different specific domains (biomedical domain, maths and computer domain, and a broader domain about social issues). Our goal is to further investigate the distance between research efforts regarding the dichotomy EL vs. WSD and those regarding the dichotomy open domain vs. closed domain.

## 2 Task Setup

The task setup consists of annotating four tokenized and part-of-speech tagged documents for which parallel versions in three languages (English, Italian and Spanish) have been provided. Differently from previous editions (Navigli et al., 2013; Lefever and Hoste, 2013; Manandhar et al., 2010; Lefever and Hoste, 2010; Pradhan et al., 2007; Navigli et al., 2007; Snyder and Palmer, 2004; Palmer et al., 2001), in this task we do not make explicit to the participating systems which fragments of the input text should be disambiguated, so as to have, on the one hand, a more realistic scenario, and, on the other hand, to follow the recent trend in EL challenges such as TAC KBP (Ji et al., 2014), MicroPost (Basave et al., 2013) and ERD (Carmel et al., 2014).

### 2.1 Corpora

The documents considered in this task are taken from the OPUS project (http://opus.lingfil.uu.se/), more specifically from the EMEA (European Medicines Agency documents), KDEdoc (the KDE manual corpus) and "The EU bookshop corpus", which make available parallel and POS-tagged documents. We took four documents from these repositories. Two documents contain medical information about drugs. One document consists of the manual of a mathematical graph calculator (i.e., KAlgebra). The remaining document contains a formal discussion about social issues, like supporting elderly workers and, more in general, about issues and solutions to unemployment discussed by the members of the European Commission.

### 2.2 Sense Inventory

As our sense inventory we use the BabelNet 2.5.1 (http://babelnet.org) multilingual semantic network and encyclopedic dictionary (Navigli and Ponzetto, 2012), which is the result of the automatic integration of multiple language resources: Princeton WordNet, Wikipedia, Wiktionary, OmegaWiki, Wikidata, Open Multi WordNet and automatic translations. The meanings contained within this resource are organized in Babel synsets. Each of these synsets can contain Wikipedia pages, WordNet synsets and items from the other integrated resources. For instance, in BabelNet it is possible to find the concept *"medicine"* (bn:00054128n), which is represented by both the second word sense of *medicine* in WordNet and the Wikipedia page *Pharmaceutical drug*, among others, together with synonyms such as *drug* and *medication* in English and lexicalizations in other languages, such as *farmaco* in Italian and *medicamento* in Spanish.

### 2.3 Dataset Creation

The manual annotation of documents was performed in a language-specific manner, i.e., different taggers worked on the various translated versions of the input documents. More precisely, we had two taggers for each language, who annotated each fragment of text recognized as linkable with all the senses deemed appropriate. During the annotation procedure, for all languages, each tagger was shown an HTML page containing the sentence within which the target fragment was boldfaced. Then a table of checkable meanings identified by their glosses (in English or, if not available, in Spanish or Italian), to-

|  |  |  | Single | Multi | Named | Mean senses | Mean senses | Mean senses per POS |  |  |  | Wikipedia | WordNet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Domain | Language | Instances | words | words | Entities | per instance | per lemma | N | V | R | A | pages | keys |
| Biomedical | EN | 623 | 534 | 41 | 48 | 8.0 | 7.0 | 8.8 | 10.0 | 2.4 | 3.8 | 295 | 549 |
|  | ES | 628 | 552 | 30 | 46 | 6.2 | 6.5 | 5.6 | 9.0 | 3.1 | 5.9 | 251 | - |
|  | IT | 610 | 545 | 29 | 36 | 5.4 | 5.7 | 5.8 | 6.0 | 3.1 | 3.5 | 254 | - |
| Maths and computer | EN | 325 | 292 | 11 | 22 | 9.0 | 9.5 | 10.1 | 10.3 | 2.9 | 5.9 | 135 | 276 |
|  | ES | 308 | 277 | 10 | 21 | 7.5 | 7.6 | 7.9 | 8.0 | 3.8 | 6.0 | 120 | - |
|  | IT | 313 | 275 | 15 | 23 | 6.9 | 6.8 | 7.3 | 7.6 | 3.3 | 4.4 | 136 | - |
| Social issues | EN | 313 | 268 | 29 | 16 | 7.4 | 6.9 | 9.1 | 6.3 | 1.5 | 4.1 | 119 | 294 |
|  | ES | 303 | 259 | 27 | 17 | 7.4 | 7.4 | 8.1 | 7.3 | 3.2 | 5.9 | 102 | - |
|  | IT | 302 | 265 | 22 | 15 | 6.6 | 6.5 | 7.7 | 6.8 | 1.7 | 3.0 | 101 | - |
| All | EN | 1261 | 1094 | 81 | 86 | 8.1 | 7.6 | 9.1 | 9.5 | 2.4 | 4.4 | 549 | 1119 |
|  | ES | 1239 | 1088 | 67 | 84 | 6.8 | 6.8 | 6.8 | 8.4 | 3.2 | 5.9 | 473 | - |
|  | IT | 1225 | 1085 | 66 | 74 | 6.1 | 5.9 | 6.6 | 6.7 | 2.8 | 3.5 | 491 | - |

Table 1: Statistics of the datasets.

gether with the available synonyms and hypernyms (as found in WordNet and the Wikipedia Bitaxonomy (Flati et al., 2014)). The taggers agreed on at least one meaning for 68% of the instances. A third tagger acted as judge by going through all the items and discarding overly general or irrelevant annotations, especially in the case of disagreement between the two taggers. To enforce coherence and spot missing annotations, we projected the English annotations to the other two languages. Finally, the third tagger determined if the projected English annotations that were missing in one of the other two languages were either correctly not included, or if the taggers had actually missed a correct annotation.

As a result of this procedure we obtained a dataset with around 1.2k items, but with only around 80 named entity mentions per language. Please refer to Table 1 for general statistics about the dataset: we show the number of annotated instances per language and domain, together with their classification as single- or multi-word expressions and named entities. We then show the degree of ambiguity both per POS and per instance and lemma (i.e., multiple instances with the same lemma count as a single instance) and, finally, we show how many of the instances have Wikipedia pages or WordNet keys as annotations[2].

### 2.4 Evaluation Measures

To evaluate the performance of the participating systems we used the classical precision, recall and F1 measures:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3)$$

To handle systems that output multiple answers for a single instance we followed the standard scorer of previous Senseval and SemEval challenges in uniformly weighting the multiple answers when computing the TP counts. Moreover, we decided not to take into account fragments annotated by the systems which were not contained in the gold standard, similarly to the D2KB setting of the GERBIL evaluation framework for EL (Usbeck et al., 2015).

### 2.5 Baseline

As baseline we considered the performance of a simple heuristic (called BabelNet first sense or BFS) that exploits the default comparator integrated within the BabelNet 2.5.1 API (i.e., the BabelSynsetComparator Java class). Babel synsets in BabelNet can be viewed as nodes of a semantic network and each of them can contain Wikipedia pages, WordNet synsets and items from the other integrated resources. The comparator takes as input the lemma of the word for which we are ranking the Babel synsets. There are three main cases managed by the comparator. The first case is when both Babel synsets contain a WordNet synset for the considered word. If this is the case, then the WordNet sense numbers are used to rank the synsets. The second case is when only one of the Babel synsets contains a WordNet synset: in this case the Babel synset that

[2] Please note that the sum of Wikipedia pages and WordNet keys does not amount to the number of instances, as BabelNet can have integrated synsets that contain both WordNet keys and Wikipedia pages.

contains the WordNet synset gets ranked first. The last case is when no WordNet synsets are contained within the two Babel synsets. In this case a lexicographic ordering of the Wikipedia pages contained within the Babel synsets is taken into account. As is well known, the first sense heuristic based on WordNet has always proved a really hard to beat baseline, outperforming all the developed systems for the English language over almost all settings and system combinations. In contrast, the BFS heuristic in the other languages shows itself to be weaker, achieving lower performances in almost all settings and system combinations.

## 3 Participating Systems

**DFKI (Supervised).** This system exploits BabelNet as reference inventory and a CRF-based named entity recognizer. The disambiguation system is divided in two parts: one for nouns and another for verbs. For nouns the approach is based on the idea of maximizing multiple objectives at the same time. Similarly to (Hoffart et al., 2011), the disambiguation objectives consist of a global (coherence, unsupervised) part and a local (supervised) part. The global objective makes sure that disambiguation maximizes coherence of the selected synsets and it is based on the semantic signature graph (Moro et al., 2014b). The local objective ensures that the WordNet synset type fits the local context of the noun to be disambiguated. One important aspect of this approach is that, unlike previous work (Hoffart et al., 2011; Moro et al., 2014b), it does not apply discrete optimization, but continuous optimization on the normalized sum of all objectives. The disambiguation procedure aims to optimize the objective function by iteratively updating the candidate probabilities for each fragment. As far as verbs are concerned, a feed-forward neural network is trained using local features such as arguments of the semantic roles of a verb in a sentence, context words, and the verb and its lemma.

**EBL-Hope (Unsupervised + Sense relevance).** This approach uses a modified version of the Lesk algorithm and the Jiang & Conrath similarity measure (Jiang and Conrath, 1997). It validates the output from both techniques for enhanced accuracy and exploits semantic relations and corpus (SemCor) information available in BabelNet and WordNet in an unsupervised manner.

**el92 (Systems mix).** This system is a general-domain system for entity detection and linking. It does not perform WSD. The system combines, via a weighted voting, Entity Linking outputs from four publicly available services: Tagme (Ferragina and Scaiella, 2010), DBpedia Spotlight (Mendes et al., 2011), Wikipedia Miner (Milne and Witten, 2008) and Babelfy (Moro et al., 2014b; Moro et al., 2014a). The different runs correspond to different settings in the weighting formula (De La Clergerie et al., 2008; Fiscus, 1997).

**LIMSI (Unsupervised + Sense relevance).** The system performs WSD by taking advantage of the parallelism of the test data, a feature that was not exploited by the systems that participated in the SemEval-2013 Multilingual Word Sense Disambiguation task 12 (Navigli et al., 2013). The system needs no training and is applied directly to the test dataset, nor does it use distributional (context) information. The texts are sentence- and word-aligned pairwise, and content words are tagged by their translations in another language. The alignments serve to retrieve the BabelNet synsets that are relevant for each instance of a word in the texts (i.e., synsets that contain both the disambiguation target and its aligned translation). If a Babel synset is retained, this is used to annotate the instance of the word in the test set. If more than one synset is retained, these are ranked using the BabelSynsetComparator Java class available in the BabelNet API (please refer to Section 2.5 for a detailed explanation). The highest ranked synset among the ones that contain the aligned translation is used to annotate the instance. The system falls back to the BabelNet first sense (BFS) provided by the BabelSynsetComparator for instances with no aligned translation, or in cases where the translation was not found in any of the synsets available for the word in BabelNet.

**SUDOKU (Unsupervised).** This deterministic constraint-based approach relies on a reasonable degree of "document monosemy" (percentage of unique monosemous lemmas in a document) and exploits Personalised PageRank (Agirre et al., 2014) to select the best candidate. The PPR is started with a surfing vector biased towards monosemous words (i.e., their respective sense). Each submission differs by its imposed constraints: Run1 is the plain approach (Manion and Sainudiin, 2014) applied at the document level; Run2 is the iterative version of the previous approach applied at the document level and with words disambiguated in order of increasing polysemy; Run3 is like Run2, but it is first applied to nouns and then to verbs, adjectives, and adverbs.

**TeamUFAL (Unsupervised).** This system exploits Apache Lucene search engine to index Wikipedia documents, Wiktionary entries and WordNet senses. Then, to perform disambiguation, the Lucene ranking method is used to query the index with multiple queries (consisting of the text fragment and context words). Finally, all query results are merged and the disambiguated meaning is selected thanks to a simple threshold heuristic.

**UNIBA (Unsupervised + Sense relevance).** This system[3] extends two well-known variations of the Lesk WSD method. The main contribution of the approach relies on the use of a word similarity function defined on a distributional semantic space (Word2vec tool (Mikolov et al., 2013)) to compute the gloss-context overlap. Entities are identified by exploiting a list of possible surface forms extracted from BabelNet synsets. Moreover, each synset has a prior probability computed over an annotated corpus. For WordNet synsets, SemCor is exploited, while for Wikipedia entities the number of citations in Wikipedia internal links is counted.

**vua-background (Partially supervised).** This approach exploits the Named Entities contained in the test data to generate a background corpus. This is done by finding similar DBpedia entities for the entities in the input documents. Using this background corpus, the system tries to find the predominant sense of the words in the test data (McCarthy et al., 2004). If a predominant sense is recognized for a specific lemma, then it is used, otherwise the system falls back to the "It Makes Sense" WSD system (Zhong and Ng, 2010).

**WSD-games (Unsupervised).** This approach is formulated in terms of Evolutionary Game Theory, where each word to be disambiguated is represented as a node in a graph and each sense as a class. The proposed algorithm performs a consistent class assignment of senses according to the similarity information of each word with the others, so that similar words are constrained to similar classes. The propagation of the information over the graph is formulated in terms of a non-cooperative multi-player game, where the players are the data points, in order to decide their class memberships, and equilibria correspond to consistent labeling of the data.

## 4 Results and Discussion

The results obtained by the participating systems are shown in Tables 2-6. In Table 2 we show the precision, recall and F1 scores of the participating systems that annotated all classes of items (named entities, nouns, verbs, adverbs, adjectives) over the whole dataset. Six out of the nine participating teams annotated the full set of items. We also show the F1 performance on each considered domain independently and for different kinds of subsets of the item classes (i.e., we show the F1 score over all items, then only on named entities, all open-class word senses and individually).

### 4.1 Overall Performance

From Table 2 we can see that the best system for English (i.e., LIMSI) is able to obtain a performance more than five percentage points higher than the second ranked system. This is due to the good-quality indirect supervision provided by the alignments combined with the use of the BabelSynsetComparator. However, on the other two languages this system obtains lower performance than the other competing systems. The performance of the SUDOKU system is of a particular interest, as it obtains the second best scores on the English part of the dataset and the top scores overall on the other two languages. It exploits monosemous words within the input documents to run Personalized PageRank. The three runs differ mainly in respect of the order in which the words get disambiguated.

In Table 3 we show the F1 scores of all the systems over the whole dataset for each class of the

[3] During the evaluation period the system did not return any annotation for adjectives due to a misinterpretation of the POS tag set. For full evaluations see the system paper.

|  | EN |  |  | ES |  |  | IT |  |  |
|---|---|---|---|---|---|---|---|---|---|
| System | P | R | F1 | P | R | F1 | P | R | F1 |
| LIMSI | **68.7** | **63.1** | **65.8** | 47.9 | 42.4 | 45.0 | 51.3 | 45.7 | 48.4 |
| SUDOKU-Run2 | 62.9 | 60.4 | 61.6 | 59.9 | **54.6** | **57.1** | 59.7 | 54.3 | 56.9 |
| SUDOKU-Run3 | 61.9 | 59.4 | 60.7 | 59.5 | 54.2 | 56.8 | 59.7 | 54.3 | 56.9 |
| vua-background | 67.5 | 51.5 | 58.4 | - | - | - | - | - | - |
| SUDOKU-Run1 | 60.1 | 52.1 | 55.8 | **60.2** | 52.3 | 56.0 | **64.4** | **55.9** | **59.9** |
| WSD-games-Run2 | 58.8 | 50.0 | 54.1 | - | - | - | - | - | - |
| WSD-games-Run1 | 57.4 | 48.9 | 52.8 | - | - | - | - | - | - |
| WSD-games-Run3 | 53.5 | 45.4 | 49.1 | - | - | - | - | - | - |
| EBL-Hope | 48.4 | 44.4 | 46.3 | - | - | - | - | - | - |
| TeamUFAL | 40.4 | 36.5 | 38.3 | - | - | - | - | - | - |
| BFS | 67.9 | 67.2 | 67.5 | 38.9 | 36.2 | 37.5 | 41.7 | 38.8 | 40.2 |
| # items | 1261 |  |  | 1239 |  |  | 1225 |  |  |

Table 2: Precision, Recall and F1 on all domains.

manually annotated items and for each language. In the English part of the datasets the DFKI system performs best for verb, noun and named entity disambiguation, thanks to precomputed random walks called semantic signatures, along the lines of Babelfy (Moro et al., 2014b), and supervised techniques. The UNIBA system on the English dataset obtains the best result on adverbs. Finally, in the Spanish dataset the EBL-Hope system based on a combination of a Lesk-based measure together with the Jiang & Conrath similarity measure shows the best performance for named entities.

### 4.2 Domain-based Evaluation

In Tables 4-6 we show the detailed performances of all the systems over different classes of items, and on different domains. One of the main goals of this task is to investigate the performance of disambiguation methods over different domains. Our documents derive from the biomedical domain, the maths and computer domain, and a broader domain (a document discussing social issues, especially for elderly workers and possible solutions).

**Biomedical domain.** In Table 4 we show the performance of the systems on the biomedical documents. The first thing to notice is the much higher best score of the first ranked system (i.e., LIMSI), which attains an F1 score of 71.3%. This is due to the lower ambiguity of nouns and named entities (see Table 1) resulting from the greater numbers of domain-specific concepts used within this kind of documents. This can also be seen from the higher scores obtained by the BFS. Overall, all systems obtained a better performance than in the other domains, with a gain of more than four percentage points each. The second ranked system (i.e., SUDOKU) shows its ability to exploit monosemous words obtaining a 0.1 difference from the first ranked system and a 0.9 point distance from the BFS baseline. This is of particular interest as the system does not explicitly exploit any sense relevance information. Moreover, the DFKI system obtains the best scores for nouns and verbs, and is the only system able to obtain a 100% F1 score on NE disambiguation. However, several other systems performed above 90%, showing that in this particular set of documents named entities are easy to disambiguate.

On the other two languages the performances are a little bit lower, but the SUDOKU system confirms its ability to exploit monosemous words at a quality comparable to the one obtained in the English dataset. The LIMSI system, instead, obtains a reduction of around 20% due to its exploitation of the BabelSynsetComparator, which performs badly in these languages (see the BFS scores).

**Maths and computer domain.** In Table 5 we show the results for the maths and computer domain. As can be seen in Table 1, this is the most ambiguous domain and the best systems obtain much lower performances than in the other domains. Interestingly, the DFKI system is not able to achieve the best performance on any of the considered item classes, while UNIBA and SUDOKU show the best results for nouns and verbs. As regards named en-

| | EN | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| LIMSI | **65.8** | 82.9 | **64.7** | 64.8 | 56.0 | 76.5 | **79.5** |
| SUDOKU-Run2 | 61.6 | 87.0 | 59.9 | 62.5 | 49.6 | 70.4 | 71.7 |
| SUDOKU-Run3 | 60.7 | 87.0 | 58.9 | 62.7 | 46.0 | 71.7 | 68.1 |
| vua-background | 58.4 | 14.9 | 60.3 | 53.8 | 55.2 | 77.2 | 72.5 |
| SUDOKU-Run1 | 55.8 | 16.8 | 57.5 | 53.4 | 52.2 | 48.9 | 74.4 |
| WSD-games-Run2 | 54.1 | 12.6 | 55.8 | 51.4 | 43.7 | 75.3 | 69.9 |
| WSD-games-Run1 | 52.8 | 12.6 | 54.5 | 49.6 | 42.5 | 75.3 | 69.9 |
| WSD-games-Run3 | 49.1 | 12.6 | 50.7 | 47.4 | 35.8 | 74.1 | 64.0 |
| EBL-Hope | 46.3 | 84.2 | 43.8 | 45.7 | 30.6 | 76.5 | 57.8 |
| TeamUFAL | 38.3 | 79.8 | 35.5 | 46.4 | 18.8 | 45.8 | 28.8 |
| DFKI | - | **88.9** | - | **70.3** | **57.7** | - | - |
| el92-Run1 | - | 86.1 | - | - | - | - | - |
| UNIBA-Run1 | - | 84.4 | - | 63.3 | 57.1 | **79.0** | - |
| UNIBA-Run2 | - | 82.9 | - | 63.2 | 57.1 | **79.0** | - |
| UNIBA-Run3 | - | 82.9 | - | 63.2 | 57.1 | **79.0** | - |
| el92-Run3 | - | 79.7 | - | - | - | - | - |
| el92-Run2 | - | 79.2 | - | - | - | - | - |
| BFS | 67.5 | 85.7 | 66.3 | 66.7 | 55.1 | 82.1 | 82.5 |

| | ES | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| SUDOKU-Run2 | **57.1** | 36.9 | **58.0** | **56.3** | 55.6 | 61.9 | 61.1 |
| SUDOKU-Run3 | 56.8 | 36.9 | 57.7 | 54.9 | **57.9** | 60.3 | 61.5 |
| SUDOKU-Run1 | 56.0 | 17.4 | 57.6 | 54.0 | 56.4 | 61.4 | **62.0** |
| LIMSI | 45.0 | 30.8 | 45.6 | 48.3 | 28.6 | **64.6** | 49.7 |
| EBL-Hope | - | **70.8** | - | 48.2 | - | - | - |
| BFS | 37.5 | 37.0 | 37.6 | 40.6 | 19.8 | 55.1 | 46.2 |

| | IT | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| SUDOKU-Run1 | **59.9** | 21.7 | **61.3** | **56.6** | **62.7** | **62.5** | 68.3 |
| SUDOKU-Run3 | 56.9 | **54.9** | 57.0 | 56.3 | 51.5 | 57.1 | 65.8 |
| SUDOKU-Run2 | 56.9 | **54.9** | 57.0 | 54.1 | 60.9 | 61.2 | 62.0 |
| LIMSI | 48.4 | 46.5 | 48.4 | 43.9 | 44.2 | 56.0 | **69.6** |
| UNIBA-Run3 | - | 50.0 | - | 53.7 | 61.1 | 60.0 | - |
| UNIBA-Run2 | - | 48.5 | - | 53.8 | 61.1 | 60.0 | - |
| UNIBA-Run1 | - | 48.5 | - | 53.7 | 61.1 | 60.0 | - |
| EBL-Hope | - | 48.5 | - | 38.8 | - | - | - |
| BFS | 40.2 | 50.0 | 39.8 | 35.4 | 38.3 | 48.0 | 61.0 |

Table 3: F1 performance by item class and language on all domains.

| | EN | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| LIMSI | **71.3** | 98.9 | **68.9** | 76.5 | 50.6 | 77.5 | 75.0 |
| SUDOKU-Run3 | 71.2 | 98.9 | 68.8 | 75.8 | 50.6 | 75.3 | 77.8 |
| SUDOKU-Run2 | 68.9 | 98.9 | 66.4 | 71.9 | 47.3 | 77.9 | **83.3** |
| vua-background | 63.6 | 4.1 | 66.4 | 62.7 | 53.8 | 76.9 | 77.4 |
| SUDOKU-Run1 | 62.4 | 4.1 | 65.0 | 62.8 | 52.5 | 50.7 | 82.3 |
| WSD-games-Run2 | 58.4 | 4.1 | 60.8 | 55.8 | 45.8 | **80.0** | 79.2 |
| WSD-games-Run1 | 56.3 | 4.1 | 58.6 | 52.2 | 45.8 | **80.0** | 79.2 |
| WSD-games-Run3 | 54.4 | 4.1 | 56.6 | 54.1 | 35.0 | 72.5 | 77.8 |
| EBL-Hope | 52.0 | 98.9 | 48.0 | 54.1 | 28.2 | **80.0** | 65.3 |
| TeamUFAL | 45.6 | 93.5 | 41.6 | 57.2 | 18.6 | 39.7 | 30.9 |
| DFKI | - | **100.0** | - | **79.1** | **58.3** | - | - |
| UNIBA-Run3 | - | 98.9 | - | 72.1 | 52.3 | **80.0** | - |
| UNIBA-Run1 | - | 98.9 | - | 71.9 | 52.3 | **80.0** | - |
| UNIBA-Run2 | - | 98.9 | - | 71.9 | 52.3 | **80.0** | - |
| el92-Run1 | - | 90.9 | - | - | - | - | - |
| el92-Run2 | - | 81.5 | - | - | - | - | - |
| el92-Run3 | - | 81.5 | - | - | - | - | - |
| BFS | 72.1 | 98.9 | 69.9 | 75.3 | 52.5 | 82.9 | 81.9 |

| | ES | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| SUDOKU-Run1 | **62.7** | 8.3 | **65.1** | **65.5** | 54.3 | **65.7** | 62.1 |
| SUDOKU-Run3 | 62.6 | 12.2 | 64.7 | 64.3 | **56.7** | 52.6 | **71.2** |
| SUDOKU-Run2 | 60.8 | 12.2 | 62.9 | 64.5 | 51.2 | 52.6 | 63.2 |
| LIMSI | 51.0 | 12.2 | 52.7 | 59.6 | 28.3 | 59.7 | 40.7 |
| EBL-Hope | - | **77.3** | - | 59.6 | - | - | - |
| BFS | 43.7 | 12.2 | 45.1 | 51.7 | 20.5 | 49.4 | 39.0 |

| | IT | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| SUDOKU-Run1 | **65.1** | 10.5 | **67.0** | **65.9** | **64.2** | 48.0 | 64.3 |
| SUDOKU-Run3 | 61.4 | **28.6** | 62.7 | 62.3 | 52.3 | 48.0 | **70.6** |
| SUDOKU-Run2 | 58.8 | **28.6** | 60.0 | 56.7 | 61.5 | **56.0** | 64.7 |
| LIMSI | 53.1 | 24.4 | 54.1 | 54.2 | 42.2 | 38.5 | 63.5 |
| UNIBA-Run3 | - | **28.6** | - | 62.4 | 63.6 | 46.2 | - |
| UNIBA-Run1 | - | 24.4 | - | 62.2 | 63.6 | 46.2 | - |
| UNIBA-Run2 | - | 24.4 | - | 62.2 | 63.6 | 46.2 | - |
| EBL-Hope | - | 24.4 | - | 50.5 | - | - | - |
| BFS | 44.3 | 28.6 | 44.9 | 43.3 | 38.7 | 38.5 | 56.8 |

Table 4: F1 performance by item class and language on biomedical domain.

tities, the system EBL-Hope obtains the best results in all languages. This system, in addition to exploiting a Lesk-based measure combined with the Jiang & Conrath similarity measure, uses the BabelNet semantic relations, which have already been shown to be useful for attaining state-of-the-art performances in EL (Moro et al., 2014b). Interestingly, in the Italian dataset the system UNIBA (which is based on an extended version of the Lesk measure and a semantic relatedness measure) obtains the same performance for NE as the EBL-Hope system.

**Social issues domain.** In Table 6 we show the performance on our last domain. In this social issues domain DFKI confirms its quality on disambiguating nouns and named entities, while for verbs the best system is vua-background, which is based on the predominant sense algorithm (McCarthy et al., 2004) and, as a fallback routine, on the "It Makes Sense" supervised WSD system (Zhong and Ng, 2010). For the other two languages the SUDOKU system obtains the best scores, with the exception of adverbs in the Italian dataset where the UNIBA system is able to reach an F1 score of 100%.

## 5 Conclusion and Future Directions

In this paper we described the organization and results obtained within the SemEval 2015 task 13: Multilingual Word Sense Disambiguation. Our analysis of the results revealed interesting aspects of the integration of WSD and EL tasks, such as the effectiveness of techniques like semantic signatures, PPR and similarity measures for noun and named entity

| | EN | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| LIMSI | **54.1** | 57.1 | **53.9** | 39.3 | 59.4 | 71.7 | **90.0** |
| SUDOKU-Run2 | 53.2 | 56.3 | 53.1 | **51.4** | 49.1 | 56.6 | 67.5 |
| SUDOKU-Run3 | 49.4 | 56.3 | 49.1 | 48.9 | 42.3 | 64.2 | 57.5 |
| EBL-Hope | 41.7 | **74.3** | 39.8 | 42.5 | 28.6 | 67.9 | 50.0 |
| TeamUFAL | 29.8 | 54.5 | 28.4 | 35.8 | 12.6 | 37.8 | 39.2 |
| el92-Run1 | - | 70.6 | - | - | - | - | - |
| el92-Run3 | - | 66.7 | - | - | - | - | - |
| el92-Run2 | - | 64.7 | - | - | - | - | - |
| DFKI | - | 57.1 | - | 44.9 | 52.3 | - | - |
| UNIBA-Run1 | - | 57.1 | - | 44.1 | **60.6** | **75.5** | - |
| UNIBA-Run2 | - | 57.1 | - | 44.1 | **60.6** | **75.5** | - |
| UNIBA-Run3 | - | 57.1 | - | 44.1 | **60.6** | **75.5** | - |
| WSD-games-Run2 | - | - | 48.5 | 39.6 | 37.7 | 64.2 | 80.0 |
| vua-background | - | - | 47.7 | 30.5 | 49.7 | 70.6 | 73.0 |
| WSD-games-Run1 | - | - | 47.4 | 39.6 | 34.3 | 64.2 | 80.0 |
| SUDOKU-Run1 | - | - | 44.7 | 28.5 | 51.4 | 52.0 | 75.0 |
| WSD-games-Run3 | - | - | 43.4 | 36.2 | 35.4 | 67.9 | 58.2 |
| BFS | 55.3 | 57.1 | 55.2 | 43.6 | 55.7 | 77.8 | 87.5 |

| | ES | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| SUDOKU-Run2 | **49.7** | 50.0 | **49.7** | 42.4 | **60.9** | 66.7 | 44.1 |
| SUDOKU-Run3 | 48.4 | 50.0 | 48.3 | 39.2 | 58.7 | 66.7 | **52.9** |
| SUDOKU-Run1 | 44.2 | - | 45.9 | 32.0 | 58.7 | 56.0 | **52.9** |
| LIMSI | 34.8 | 56.3 | 33.6 | 32.2 | 27.2 | **81.5** | 47.1 |
| EBL-Hope | - | **68.8** | - | **45.4** | - | - | - |
| BFS | 28.7 | 62.5 | 26.8 | 27.1 | 16.3 | 74.1 | 50.0 |

| | IT | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| SUDOKU-Run2 | **52.1** | 68.6 | **51.1** | **46.6** | 59.0 | **66.7** | 58.5 |
| SUDOKU-Run3 | 49.1 | 68.6 | 47.9 | 43.0 | 53.0 | **66.7** | 63.4 |
| SUDOKU-Run1 | 48.4 | - | 50.5 | 35.8 | **60.2** | **66.7** | 70.7 |
| LIMSI | 44.6 | 64.9 | 43.3 | 33.4 | 45.8 | **66.7** | **85.4** |
| UNIBA-Run1 | - | **75.7** | - | 43.4 | 57.8 | 50.0 | - |
| UNIBA-Run2 | - | **75.7** | - | 43.4 | 57.8 | 50.0 | - |
| UNIBA-Run3 | - | **75.7** | - | 42.2 | 57.8 | 50.0 | - |
| EBL-Hope | - | **75.7** | - | 37.1 | - | - | - |
| BFS | 36.7 | 64.9 | 34.8 | 27.4 | 37.3 | 66.7 | 70.7 |

Table 5: F1 performance by item class and language on maths and computer domain.

| | EN | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| LIMSI | **67.2** | 54.5 | **67.7** | 63.7 | 63.6 | 82.8 | **77.8** |
| vua-background | 60.8 | 54.5 | 61.1 | 54.8 | **70.6** | **89.7** | 65.3 |
| SUDOKU-Run1 | 56.4 | 60.9 | 56.2 | 56.4 | 52.9 | 36.4 | 63.6 |
| SUDOKU-Run2 | 55.6 | 81.5 | 54.5 | 52.8 | 56.8 | 75.9 | 59.3 |
| WSD-games-Run1 | 53.5 | 45.5 | 53.8 | 53.0 | 50.0 | 82.8 | 50.0 |
| WSD-games-Run2 | 53.5 | 45.5 | 53.8 | 53.0 | 50.0 | 82.8 | 50.0 |
| SUDOKU-Run3 | 51.1 | 81.5 | 49.7 | 48.2 | 40.9 | 75.9 | 63.0 |
| WSD-games-Run3 | 46.7 | 45.5 | 46.7 | 44.2 | 38.6 | **89.7** | 50.0 |
| EBL-Hope | 39.5 | 36.4 | 39.6 | 31.5 | 40.9 | 82.8 | 53.7 |
| TeamUFAL | 32.5 | 64.2 | 31.0 | 33.6 | 31.8 | 72.4 | 18.4 |
| DFKI | - | **90.3** | - | **73.4** | 66.7 | - | - |
| el92-Run1 | - | 89.7 | - | - | - | - | - |
| el92-Run2 | - | 89.7 | - | - | - | - | - |
| el92-Run3 | - | 89.7 | - | - | - | - | - |
| UNIBA-Run1 | - | 66.7 | - | 63.0 | 63.6 | 82.8 | - |
| UNIBA-Run2 | - | 54.5 | - | 62.3 | 63.6 | 82.8 | - |
| UNIBA-Run3 | - | 54.5 | - | 61.9 | 63.6 | 82.8 | - |
| BFS | 70.8 | 77.4 | 70.5 | 69.2 | 61.4 | 87.5 | 79.6 |

| | ES | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| SUDOKU-Run2 | **57.0** | **69.2** | **56.5** | **51.6** | 57.5 | **87.0** | **70.0** |
| SUDOKU-Run1 | 54.2 | 52.2 | 54.3 | 49.7 | 57.5 | 52.6 | 68.0 |
| SUDOKU-Run3 | 53.3 | **69.2** | 52.5 | 49.5 | **59.8** | 78.3 | 56.0 |
| LIMSI | 43.1 | 34.8 | 43.5 | 39.3 | 32.2 | 60.9 | 62.0 |
| EBL-Hope | - | 52.2 | - | 26.6 | - | - | - |
| BFS | 34.0 | 51.9 | 33.1 | 30.2 | 25.0 | 52.2 | 52.0 |

| | IT | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Named | Word Senses | | | | |
| System | All | Entities | All | N | V | R | A |
| SUDOKU-Run1 | **61.0** | 63.6 | **60.9** | 56.0 | **63.4** | 90.9 | **72.4** |
| SUDOKU-Run2 | 57.9 | **80.0** | 56.9 | 55.6 | **63.4** | 66.7 | 60.3 |
| SUDOKU-Run3 | 55.8 | **80.0** | 54.7 | **56.1** | 46.3 | 66.7 | 60.3 |
| LIMSI | 42.9 | 57.1 | 42.4 | 33.1 | 46.3 | 83.3 | 67.2 |
| UNIBA-Run3 | - | 47.6 | - | 47.1 | 61.0 | **100.0** | - |
| UNIBA-Run2 | - | 47.6 | - | 46.7 | 61.0 | **100.0** | - |
| UNIBA-Run1 | - | 47.6 | - | 46.3 | 61.0 | **100.0** | - |
| EBL-Hope | - | 47.6 | - | 16.7 | - | - | - |
| BFS | 35.7 | 64.0 | 34.5 | 27.0 | 39.0 | 50.0 | 60.3 |

Table 6: F1 performance by item class and language on social issues domain.

disambiguation, and Lesk-based measures for verb, adjective and adverb disambiguation. Another interesting outcome that emerges from this task is that supervised approaches are difficult to generalize in a multilingual setting. In fact, the supervised systems that participated in this task took into account only the English language. Moreover, the task confirms yet again that the WordNet first sense heuristic is a hard baseline to beat. Unfortunately, no domain-specific disambiguation system participated in the task. However, in the biomedical domain, the participating systems show higher quality performances than in the other considered domains.

As future directions, we would like to continue to investigate the nature of this novel joint task, and to concentrate on the differences between named entity disambiguation and word sense disambiguation with a special focus on non-European languages.

## Acknowledgments

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

The organization of this task could not have been possible without the help of many people. In particular, we would like to thank José Camacho Collados, Claudio Delli Bovi, Tiziano Flati, Darío Garigliotti, Ignacio Iacobacci, Luca Matteis, Mohammad Taher Pilehvar, Alessandro Raganato, Daniele Vannella for the help provided with the annotations, all the participating teams for the useful discussions and Jim McManus for his comments on the manuscript.

## References

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.

Amparo Elizabeth Cano Basave, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. 2013. Making Sense of Microposts (# MSM2013) Concept Extraction Challenge. In *Proc. of # MSM*, pages 1–15.

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. of SIGMOD*, pages 1247–1250.

David Carmel, Ming-Wei Chang, Evgeniy Gabrilovich, Bo-June (Paul) Hsu, and Kuansan Wang. 2014. ERD'14: Entity Recognition and Disambiguation Challenge. *SIGIR Forum*, 48(2):63–77.

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

Éric Villemonte De La Clergerie, Olivier Hamon, Djamel Mostefa, Christelle Ayache, Patrick Paroubek, and Anne Vilnat. 2008. Passage: from french parser evaluation to large sized treebank. *Proc. of LREC*.

Nicolai Erbs, Torsten Zesch, and Iryna Gurevych. 2011. Link Discovery: A Comprehensive Analysis. In *Proc. of ICSC*, pages 83–86.

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.

Jonathan G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Proc. of Automatic Speech Recognition and Understanding*, pages 347–354.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2014. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 945–955.

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.

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.

Heng Ji, Joel Nothman, and Ben Hachey. 2014. Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In *Proc. Text Analysis Conference (TAC2014)*.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the International Conference on Research in Computational Linguistics*, pages 19–33.

Els Lefever and Veronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proc. of SemEval*, pages 15–20.

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

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

Steve L. Manion and Raazesh Sainudiin. 2014. An iterative 'sudoku style' approach to subgraph-based word sense disambiguation. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 40–50, Dublin, Ireland, August.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 279–286.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of Advances in Neural Information Processing Systems*, pages 3111–3119.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-Line Lexical Database. *Int. Journal of Lexicography*, 3(4):235–244.

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

Andrea Moro, Francesco Cecconi, and Roberto Navigli. 2014a. Multilingual word sense disambiguation and entity linking for everybody. In *Proc. of ISWC (P&D)*, pages 25–28.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014b. Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of* *the Association for Computational Linguistics*, 2:231–244.

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

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 *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proc. of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA.

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

Roberto Navigli. 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.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proc. of Senseval-2*, pages 21–24.

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.

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*, pages 93–115.

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

Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. 2015. GERBIL - General Entity Annotator Benchmark. In *Proc. of WWW*.

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