SemEval-2013 Task 12: Multilingual Word Sense Disambiguation

Roberto Navigli, David Jurgens and **Daniele Vannella**
Dipartimento di Informatica
Sapienza Università di Roma
Viale Regina Elena, 295 – 00161 Roma Italy
{navigli,jurgens,vannella}@di.uniroma1.it

Abstract

This paper presents the SemEval-2013 task on multilingual Word Sense Disambiguation. We describe our experience in producing a multilingual sense-annotated corpus for the task. The corpus is tagged with BabelNet 1.1.1, a freely-available multilingual encyclopedic dictionary and, as a byproduct, WordNet 3.0 and the Wikipedia sense inventory. We present and analyze the results of participating systems, and discuss future directions.

1 Introduction

Word Sense Disambiguation (WSD), the task of automatically assigning predefined meanings to words occurring in context, is a fundamental task in computational lexical semantics (Navigli, 2009; Navigli, 2012). Several Senseval and SemEval tasks have been organized in the past to study the performance and limits of disambiguation systems and, even more importantly, disambiguation settings. While an ad-hoc sense inventory was originally chosen for the first Senseval edition (Kilgarriff, 1998; Kilgarriff and Palmer, 2000), later tasks (Edmonds and Cotton, 2001; Snyder and Palmer, 2004; Mihalcea et al., 2004) focused on WordNet (Miller et al., 1990; Fellbaum, 1998) as a sense inventory. In 2007 the issue of the fine sense granularity of WordNet was addressed in two different SemEval disambiguation tasks, leading to the beneficial creation of coarser-grained sense inventories from WordNet itself (Navigli et al., 2007) and from OntoNotes (Pradhan et al., 2007).

In recent years, with the exponential growth of the Web and, consequently, the increase of non-English speaking surfers, we have witnessed an upsurge of interest in multilinguality. SemEval-2010 tasks on cross-lingual Word Sense Disambiguation (Lefever and Hoste, 2010) and cross-lingual lexical substitution (Mihalcea et al., 2010) were organized. While these tasks addressed the multilingual aspect of sense-level text understanding, they departed from the traditional WSD paradigm, i.e., the automatic assignment of senses from an existing inventory, and instead focused on lexical substitution (McCarthy and Navigli, 2009). The main factor hampering traditional WSD from going multilingual was the lack of a freely-available large-scale multilingual dictionary.

The recent availability of huge collaboratively-built repositories of knowledge such as Wikipedia has enabled the automated creation of large-scale lexical knowledge resources (Hovy et al., 2013). Over the past few years, a wide-coverage multilingual "encyclopedic" dictionary, called BabelNet, has been developed (Navigli and Ponzetto, 2012a). BabelNet[1] brings together WordNet and Wikipedia and provides a multilingual sense inventory that currently covers 6 languages. We therefore decided to put the BabelNet 1.1.1 sense inventory to the test and organize a traditional Word Sense Disambiguation task on a given English test set translated into 4 other languages (namely, French, German, Spanish and Italian). Not only does BabelNet enable multilinguality, but it also provides coverage for both lexicographic (e.g., apple as fruit) and encyclopedic

[1] http://babelnet.org

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, June 14-15, 2013.

meanings (e.g., Apple Inc. as company). In this paper we describe our task and disambiguation dataset and report on the system results.

2 Task Setup

The task required participating systems to annotate nouns in a test corpus with the most appropriate sense from the BabelNet sense inventory or, alternatively, from two main subsets of it, namely the WordNet or Wikipedia sense inventories. In contrast to previous all-words WSD tasks we did not focus on the other three open classes (i.e., verbs, adjectives and adverbs) since BabelNet does not currently provide non-English coverage for them.

2.1 Test Corpus

The test set consisted of 13 articles obtained from the datasets available from the 2010, 2011 and 2012 editions of the workshop on Statistical Machine Translation (WSMT).[2] The articles cover different domains, ranging from sports to financial news.

The same article was available in 4 different languages (English, French, German and Spanish). In order to cover Italian, an Italian native speaker manually translated each article from English into Italian, with the support of an English mother tongue advisor. In Table 1 we show for each language the number of words of running text, together with the number of multiword expressions and named entities annotated, from the 13 articles.

2.2 Sense Inventories

2.2.1 BabelNet inventory

To semantically annotate all the single- and multiword expressions, as well as the named entities, occurring in our test corpus we used BabelNet 1.1.1 (Navigli and Ponzetto, 2012a). BabelNet is a multilingual "encyclopedic dictionary" and a semantic network currently covering 6 languages, namely: English, Catalan, French, German, Italian and Spanish. BabelNet is obtained as a result of a novel integration and enrichment methodology. This resource is created by linking the largest multilingual Web encyclopedia – i.e., Wikipedia – to the most popular computational lexicon – i.e., WordNet 3.0. The integration is performed via an automatic mapping and by filling in lexical gaps in resource-poor languages with the aid of Machine Translation (MT).

Its lexicon includes lemmas which denote both lexicographic meanings (e.g., *balloon*) and encyclopedic ones (e.g., *Montgolfier brothers*). The basic meaning unit in BabelNet is the Babel synset, modeled after the WordNet synset (Miller et al., 1990; Fellbaum, 1998). A Babel synset is a set of synonyms which express a concept in different languages. For instance, { Globus aerostàtic$_{\text{CA}}$, Balloon$_{\text{EN}}$, Aérostation$_{\text{FR}}$, Ballon$_{\text{DE}}$, Pallone aerostatico$_{\text{IT}}$, ..., Globo aerostático$_{\text{ES}}$ } is the Babel synset for the balloon aerostat, where the language of each synonym is provided as a subscript label. Thanks to their multilingual nature, we were able to use Babel synsets as interlingual concept tags for nouns occurring within text written in any of the covered languages.

2.2.2 WordNet and Wikipedia inventories

Since BabelNet 1.1.1 is a superset of the WordNet 3.0 and Wikipedia sense inventories,[3] once text is annotated with Babel synsets, it turns out to be annotated also according to either WordNet or Wikipedia, or both. In fact, in order to induce the WordNet annotations, one can restrict to those lexical items annotated with Babel synsets which contain WordNet senses for the target lemma; similarly, for Wikipedia, we restrict to those items tagged with Babel synsets which contain Wikipedia pages for the target lemma.

2.3 BabelNet sense inventory validation

Because BabelNet is an automatic integration of WordNet and Wikipedia, the resulting Babel synsets may contain WordNet and Wikipedia entries about different meanings of the same lemma. The underlying cause is a wrong mapping between the two original resources. For instance, in BabelNet 1.1 the WordNet synset { arsenic, As, atomic number 33 } was mapped to the Wikipedia page AS (ROMAN COIN), and therefore the same Babel synset mixed the two meanings.

In order to avoid an inconsistent semantic tagging of text, we decided to manually check all the mappings in BabelNet 1.1 between Wikipedia pages

[2] http://www.statmt.org/wmt12/

[3] For version 1.1.1 we used the English Wikipedia database dump from October 1, 2012.

Language	Instances	Single-words	Multiword expressions	Named Entities	Mean senses per instance	Mean senses per lemma
			BabelNet			
English	1931	1604	127	200	1.02	1.09
French	1656	1389	89	176	1.05	1.15
German	1467	1267	21	176	1.00	1.05
Italian	1706	1454	211	41	1.22	1.27
Spanish	1481	1103	129	249	1.15	1.19
			Wikipedia			
English	1242	945	102	195	1.15	1.16
French	1039	790	72	175	1.18	1.14
German	1156	957	21	176	1.07	1.08
Italian	1977	869	85	41	1.20	1.18
Spanish	1103	758	107	248	1.11	1.10
			WordNet			
English	1644	1502	85	57	1.01	1.10

Table 1: Statistics for the sense annotations of the test set.

and WordNet senses involving lemmas in our English test set for the task. Overall, we identified 8306 synsets for 978 lemmas to be manually checked. We recruited 8 annotators in our research group and assigned each lemma to two annotators. Each annotator was instructed to check each Babel synset and determine whether any of the following three operations was needed:

- **Delete** a mapping and separate the WordNet sense from the Wikipedia page (like in the *arsenic* vs. AS (ROMAN COIN) example above);
- **Add** a mapping between a WordNet sense and a Wikipedia page (formerly available as two separate Babel synsets);
- **Merge** two Babel synsets which express the same concept.

After disagreement adjudication carried out by the first author, the number of delete, add and merge operations was 493, 203 and 43, respectively, for a total of 739 operations (i.e., 8.8% of synsets corrected). As a result of our validation of BabelNet 1.1, we obtained version 1.1.1, which is currently available online.

2.4 Sense Annotation

To ensure high quality annotations, the annotation process was completed in three phases. Because BabelNet is a superset of both the WordNet and Wikipedia sense inventories, all annotators used the BabelNet 1.1.1 sense inventory for their respective language. These BabelNet annotations were then projected into WordNet and Wikipedia senses. Annotation was performed by one native speaker each for English, French, German and Spanish and, for Italian, by two native speakers who annotated different subsets of the corpus.

In the first phase, each annotator was instructed to inspect each instance to check that (1) the lemma was tagged with the correct part of speech, (2) lemmas were correctly annotated as named entity or multiword expressions, and (3) the meaning of the instance's lemma had an associated sense in BabelNet. Based on these criteria, annotators removed dozens of instances from the original data.

In the second phase, each instance in the English dataset was annotated using BabelNet senses. To reduce the time required for annotation in the other languages, the sense annotations for the English dataset were then projected onto the other four

Language	Projected instances	Valid projections	Invalid projections
French	1016	791	225
German	592	373	219
Italian	1029	774	255
Spanish	911	669	242

Table 2: Statistics when using the English sense annotations to project the correct sense of a lemma in another language of the sentence-aligned test data.

languages using the sense translation API of BabelNet (Navigli and Ponzetto, 2012d). The projection operated as follows, using the aligned sentences in the English and non-English texts. For an instance in the non-English text, all of the senses for that instance's lemma were compared with the sense annotations in the English sentence. If any of that lemma's senses was used in the English sentence, then that sense was selected for the non-English instance. The matching procedure operates at the sentence-aligned level because the instances themselves are not aligned; i.e., different languages have different numbers of instances per sentence, which are potentially ordered differently due to language-specific construction. Ultimately, this projection labeled approximately 50-70% of the instances in the other four languages. Given the projected senses, the annotators for the other four languages were then asked to (1) correct the projected sense labels and (2) annotate those still without senses.[4] These annotations were recorded in text in a stand-off file; no further annotation tools were used.

The resulting sense projection proved highly useful for selecting the correct sense. Table 2 shows the number of corrections made by the annotators to the projected senses, who changed only 22-37% of the labels. While simple, the projection method offers significant potential for generating good quality sense-annotated data from sentence-aligned multilingual text.

In the third phase, an independent annotator reviewed the labels for the high-frequency lemmas for all languages to check for systematic errors and discuss possible changes to the labeling. This review resulted in only a small number of changes to less than 5% of the total instances, except for German which had a slightly higher percentage of changes.

Table 1 summarizes the sense annotation statistics for the test set. Annotators were allowed to use multiple senses in the case of ambiguity, but encouraged to use a single sense whenever possible. In rare cases, a lemma was annotated with senses from a different lemma. For example, WordNet does not contain a sense for "card" that corresponds to the penalty card meaning (as used in sports such as football). In contrast, BabelNet has a sense for "penalty card" from Wikipedia which, however, is not mapped to the lemma "card". In such cases, we add both the closest meaning from the original lemma (e.g., the rectangual piece of paper sense in WordNet) and the most suitable sense that may have a different lemma form (e.g., PENALTY CARD).

Previous annotation studies have shown that, when a fine-grained sense inventory is used, annotators will often label ambiguous instances with multiple senses if allowed (Erk and McCarthy, 2009; Jurgens and Klapaftis, 2013). Since BabelNet is a combination of a fine-grained inventory (WordNet) and contains additional senses from Wikipedia, we analyzed the average number of BabelNet sense annotations per instance, shown in column six of Table 1. Surprisingly, Table 1 suggests that the rate of multiple sense annotation varies significantly between languages.

BabelNet may combine multiple Wikipedia pages into a single BabelNet synset. As a result, when Wikipedia is used as a sense inventory, instances are annotated with all of the Wikipedia pages associated with each BabelNet synset. Indeed, Table 1 shows a markedly increased multi-sense annotation rate for three languages when using Wikipedia.

As a second analysis, we considered the observed level of polysemy for each of the unique lemmas. The last column of Table 1 shows the average number of different senses seen for each lemma across the test sets. In all languages, often only a single sense of a lemma was used. Because the test set is constructed based on topical documents, infrequent lemmas mostly occurred within a single document where they were used with a consistent interpreta-

[4]During the second phase, annotators were also allowed to add and remove instances that were missed during the first phase, which resulted in small number of changes.

tion. However, we note that in the case of lemmas that were only seen with a single sense, this sense does not always correspond to the most frequent sense as seen in SemCor.

3 Evaluation

Task 12 uses the standard definitions of precision and recall for WSD evaluation (see, e.g., (Navigli, 2009)). Precision measures the percentage of the sense assignments provided by the system that are identical to the gold standard; Recall measures the percentage of instances that are correctly labeled by the system. When a system provides sense labels for all instances, precision and recall are equivalent. Systems using BabelNet and WordNet senses are compared against the Most Frequent Sense (MFS) baseline obtained by using the WordNet most frequent sense. For the Wikipedia sense inventory, we constructed a pseudo-MFS baseline by selecting (1) the Wikipedia page associated with the highest ranking WordNet sense, as ranked by SemCor frequency, or (2) when no synset for a lemma was associated with a WordNet sense, the first Wikipedia page sorted using BabelNet's ordering criteria, i.e., lexicographic sorting. We note that, in the second case, this procedure frequently selected the page with the same name as the lemma itself. For instance, the first sense of *Dragon Ball* is the cartoon with title DRAGON BALL, followed by two films (DRAGON BALL (1990 FILM) and DRAGON BALL EVOLUTION).

Systems were scored separately for each sense inventory. We note that because the instances in each test set are filtered to include only those that can be labeled with the respective inventory, both the Wikipedia and WordNet test sets are subsets of the instances in the BabelNet test set.

4 Participating Systems

Three teams submitted a total of seven systems for the task, with at least one participant attempting all of the sense inventory and language combinations. Six systems participated in the WSD task with BabelNet senses; two teams submitted four systems using WordNet senses; and one team submitted three systems for Wikipedia-based senses. Notably, all systems used graph-based approaches for sense disambiguation, either using WordNet or BabelNet's synset graphs. We summarize the teams' systems as follows.

DAEBAK! DAEBAK! submitted one system called PD (Peripheral Diversity) based on BabelNet path indices from the BabelNet synset graph. Using a ± 5 sentence window around the target word, a graph is constructed for all senses of co-occurring lemmas following the procedure proposed by Navigli and Lapata (2010). The final sense is selected based on measuring connectivity to the synsets of neighboring lemmas. The MFS is used as a backoff strategy when no appropriate sense can be picked out.

GETALP GETALP submitted three systems, two for BabelNet and one for WordNet, all based on the ant-colony algorithm of (Schwab et al., 2012), which uses the sense inventory network structure to identify paths connecting synsets of the target lemma to the synsets of other lemmas in context. The algorithm requires setting several parameters for the weighting of the structure of the context-based graph, which vary across the three systems. The BN1 system optimizes its parameters from the trial data, while the BN2 and WN1 systems are completely unsupervised and optimize their parameters directly from the structure of the BabelNet and WordNet graphs.

UMCC-DLSI UMCC-DLSI submitted three systems based on the ISR-WN resource (Gutiérrez et al., 2011), which enriches the WordNet semantic network using edges from multiple lexical resources, such as WordNet Domains and the eXtended WordNet. WSD was then performed using the ISR-WN network in combination with the algorithm of Gutiérrez (2012), which is an extension of the Personalized PageRank algorithm for WSD (Agirre and Soroa, 2009) which includes senses frequency. The algorithm requires initializing the PageRank algorithm with a set of seed synsets (vertices) in the network; this initialization represents the key variation among UMCC's three approaches. The RUN-1 system performs WSD using all noun instances from the sentence context. In contrast, the RUN-2 works at the discourse level and initializes the PageRank using the synsets of all

Team	System	English	French	German	Italian	Spanish
DAEBAK!	PD	0.604	0.538	0.591	0.613	0.600
GETALP	BN-1	0.263	0.261	0.404	0.324	-
GETALP	BN-2	0.266	0.257	0.400	0.324	0.371
UMCC-DLSI	RUN-1	0.677	**0.605**	0.618	0.657	0.705
UMCC-DLSI	RUN-2	**0.685**	**0.605**	**0.621**	**0.658**	**0.710**
UMCC-DLSI	RUN-3	0.680	-	-	-	-
MFS		0.665	0.453	0.674	0.575	0.645

Table 3: System performance, reported as F1, for all five languages in the test set when using BabelNet senses. Top performing systems are marked in bold.

nouns in the document. Finally, the RUN-3 system initializes using all words in the sentence.

5 Results and Discussion

All teams submitted at least one system using the BabelNet inventory, shown in Table 3. The UMCC-DLSI systems were consistently able to outperform the MFS baseline (a notoriously hard-to-beat heuristic) in all languages except German. Additionally, the DAEBAK! system outperformed the MFS baseline on French and Italian. The UMCC-DLSI RUN-2 system performed the best for all languages. Notably, this system leverages the single-sense per discourse heuristic (Yarowsky, 1995), which uses the same sense label for all occurrences of a lemma in a document.

UMCC-DLSI submitted the only three systems to use Wikipedia-based senses. Table 4 shows their performance. Of the three sense inventories, Wikipedia had the most competitive MFS baseline, scoring at least 0.694 on all languages. Notably, the Wikipedia-based system has the lowest recall of all systems. Despite having superior precision to the MFS baseline, the low recall brought the resulting F1 measure below the MFS.

Two teams submitted four total systems for WordNet, shown in Table 5. The UMCC-DLSI RUN-2 system was again the top-performing system, underscoring the benefit of using discourse information in selecting senses. The other two UMCC-DLSI systems also surpassed the MFS baseline. Though still performing worse than the MFS baseline, when using the WordNet sense graph, the GETALP system sees a noticeable improvement of 0.14 over its performance on English data when using the WordNet sense graph.

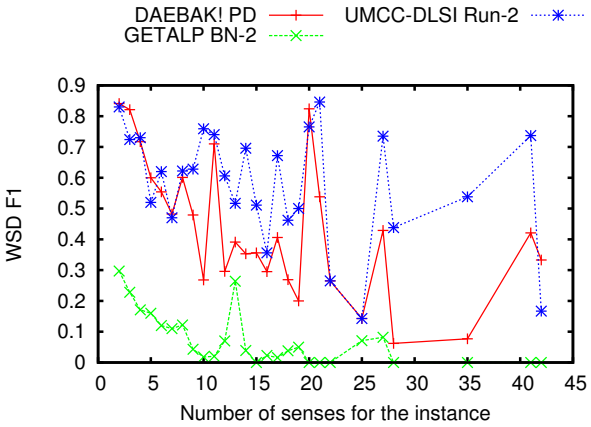

Figure 1: F1 measure according to the degree of instance polysemy, reported when at least ten instances have the specified polysemy.

The disambiguation task encompasses multiple types of entities. Therefore, we partitioned the BabelNet test data according to the type of instance being disambiguated; Table 6 highlights the results per instance type, averaged across all languages.[5] Both multiword expressions and named entities are less polysemous, resulting in a substantially higher MFS baseline that no system was able to outperform on the two classes. However, for instances made of a single term, both of the UMCC-DLSI systems were able to outperform the MFS baseline.

BabelNet adds many Wikipedia senses to the existing WordNet senses, which increases the poly-

[5]We omit the UMCC-DLSI Run-3 system from analysis, as it participated in only a single language.

		English			French			German			Italian			Spanish		
Team	System	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
UMCC-DLSI	RUN-1	0.619	0.484	0.543	0.817	0.480	0.605	0.758	0.460	0.572	0.785	0.458	0.578	0.773	0.493	0.602
UMCC-DLSI	RUN-2	0.620	0.487	0.546	0.815	0.478	0.603	0.769	0.467	0.581	0.787	0.463	0.583	0.778	0.502	0.610
UMCC-DLSI	RUN-3	0.622	0.489	0.548	-	-	-	-	-	-	-	-	-	-	-	-
MFS		0.860	0.753	0.803	0.698	0.691	0.694	0.836	0.827	0.831	0.833	0.813	0.823	0.830	0.819	0.824

Table 4: The F1 measure for each system across all five languages in the test set when using Wikipedia-based senses.

Team	System	Precision	Recall	F1
GETALP	WN-1	0.406	0.406	0.406
UMCC-DLSI	RUN-1	0.639	0.635	0.637
UMCC-DLSI	RUN-2	**0.649**	**0.645**	**0.647**
UMCC-DLSI	RUN-3	0.642	0.639	0.640
MFS		0.630	0.630	0.630

Table 5: System performance when using WordNet senses. Top performing systems are marked in bold.

Team	System	Single term	Multiword expression	Named Entity
DAEBAK!	PD	0.502	0.801	**0.910**
GETALP	BN-1	0.232	0.724	0.677
GETALP	BN-2	0.235	0.740	0.656
UMCC-DLSI	RUN-1	0.582	0.806	0.865
UMCC-DLSI	RUN-2	**0.584**	**0.809**	0.864
MFS		0.511	0.853	0.920

Table 6: System F1 per instance type, averaged across all submitted languages, with the highest system scores in bold.

		English			French			German			Italian			Spanish		
Team	System	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
DAEBAK	PD	0.769	0.364	0.494	0.747	0.387	0.510	0.762	0.307	0.438	0.778	0.425	0.550	0.778	0.450	0.570
GETALP	BN-2	0.793	0.111	0.195	0.623	0.130	0.215	0.679	0.124	0.210	0.647	0.141	0.231	0.688	0.177	0.282
UMCC-DLSI	RUN-1	0.787	0.421	0.549	0.754	0.441	0.557	0.741	0.330	0.457	0.796	0.461	0.584	0.830	0.525	0.643
UMCC-DLSI	RUN-2	0.791	0.419	0.548	0.760	0.436	0.554	0.746	0.332	0.460	0.799	0.453	0.578	0.837	0.530	0.649

Table 7: System performance when the system's annotations are restricted to only those senses that it also uses in the aligned sentences of at least two other languages.

semy of most instances. As a further analysis, we consider the relationship between the polysemy of an instance's target and system performance. Instances were grouped according to the number of BabelNet senses that their lemma had; following, systems were scored on each grouping. Figure 1 shows the performance of the best system from each team on each polysemy-based instance grouping, with a general trend of performance decay as the number of senses increases. Indeed, all systems' performances are negatively correlated with the degree of polysemy, ranging from -0.401 (UMCC-DLSI RUN-1) to -0.654 (GETALP BN-1) when measured using Pearson's correlation. All systems'

correlations are significant at $p < 0.05$.

Last, we note that all systems operated by sense-annotating each language individually without taking advantage of either the multilingual structure of BabelNet or the sentence alignment of the test data. For example, the sense projection method used to create the initial set of multilingual annotations on our test data (cf. Table 2) suggests that the sense translation API could be used as a reliable source for estimating the correctness of an annotation; specifically, given the sense annotations for each language, the translation API could be used to test whether the sense is also present in the aligned sentence in the other languages.

Therefore, we performed a post-hoc analysis of the benefit of multilingual sense alignment using the results of the four systems that submitted for all languages in BabelNet. For each language, we filter the sense annotations such that an annotation for an instance is retained only if the system assigned the same sense to some word in the aligned sentence from at least two other languages.

Table 7 shows the resulting performance for the four systems. As expected, the systems exhibit significantly lower recall due to omitting all language-specific instances. However, the resulting precision is significantly higher than the original performance, shown in Table 3. Additionally, we analyzed the set of instances reported for each system and confirmed that the improvement is not due to selecting only monosemous lemmas. Despite the GETALP system having the lower performance of the four systems when all instances are considered, the system obtains the highest precision for the English dataset. Furthermore, the UMCC-DLSI systems still obtain moderate recall, while enjoying 0.106-0.155 absolute improvements in precision across all languages. While the resulting F1 is lower due to a loss of recall, we view this result as a solid starting point for other methods to sense-tag the remaining instances. Overall, these results corroborate previous studies suggesting that highly precise sense annotations can be obtained by leveraging multiple languages (Navigli and Ponzetto, 2012b; Navigli and Ponzetto, 2012c).

6 Conclusion and Future Directions

Following recent SemEval efforts with word senses in multilingual settings, we have introduced a new task on multilingual WSD that uses the recently released BabelNet 1.1.1 sense inventory. Using a data set of 13 articles in five languages, all nominal instances were annotated with BabelNet senses. Because BabelNet is a superset of WordNet and Wikipedia, the task also facilitates analysis in those sense inventories.

Three teams submitted seven systems, with all systems leveraging the graph-based structure of WordNet and BabelNet. Several systems were able to outperform the competitive MFS baseline, except in the case of Wikipedia, but current performance leaves significant room for future improvement. In addition, we believe that future research could leverage sense parallelism available in sentence-aligned multilingual corpora, together with enriched information available in future versions of BabelNet. All of the resources for this task, including the newest 1.1.1 version of BabelNet, were released on the task website.[6]

Acknowledgments

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

A large group of people assisted with SemEval-2013 Task 12, and without whose help this task would not have been possible. In particular, we would like to thank Philipp Cimiano, Maud Erhmann, Sascha Hinte, Jesús Roque Campaña Gómez, and Andreas Soos for their assistance in sense annotation; our fellow LCL team members: Moreno De Vincenzi, Stefano Faralli, Tiziano Flati, Marc Franco Salvador, Andrea Moro, Silvia Necşulescu, and Taher Pilehvar for their invaluable assistance in creating BabelNet 1.1.1, preparing and validating sense annotations, and sense-tagging the Italian corpus; last, we thank Jim McManus for his help in producing the Italian test data.

[6] http://www.cs.york.ac.uk/semeval-2013/task12/

References

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

Philip Edmonds and Scott Cotton. 2001. Senseval-2: Overview. In *Proceedings of The Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–6, Toulouse, France.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 440–449, Singapore.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Yoan Gutiérrez, Antonio Fernández Orquín, Sonia Vázquez, and Andrés Montoyo. 2011. Enriching the integration of semantic resources based on wordnet. *Procesamiento del Lenguaje Natural*, 47:249–257.

Yoan Gutiérrez. 2012. *Análisis semántico multidimensional aplicado a la desambiguación del lenguaje natural*. Ph.D. thesis, Universidad de Alicante.

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

David Jurgens and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Proceedings of the 7th International Workshop on Semantic Evaluation*.

Adam Kilgarriff and Martha Palmer. 2000. Introduction to the special issue on senseval. *Computers and the Humanities*, 34(1-2):1–13.

Adam Kilgarriff. 1998. Senseval: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 1255–1258, Granada, Spain.

Els Lefever and Veronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04, Barcelona, Spain, 25–26 July 2004*, pages 25–28.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 9–14, Uppsala, Sweden. Association for Computational Linguistics.

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244.

Roberto Navigli and Mirella Lapata. 2010. An experimental study on graph connectivity for unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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. BabelRelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, Toronto, Ontario, Canada.

Roberto Navigli and Simone Paolo Ponzetto. 2012c. Joining forces pays off: Multilingual Joint Word Sense Disambiguation. In *Proceedings of EMNLP-CoNLL*, pages 1399–1410, Jeju Island, Korea.

Roberto Navigli and Simone Paolo Ponzetto. 2012d. Multilingual WSD with just a few lines of code: the BabelNet API. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Korea.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 Task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic*, pages 30–35.

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 *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pages 115–129.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 Task-17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic*, pages 87–92.

Didier Schwab, Jérôme Goulian, Andon Tchechmedjiev, and Hervé Blanchon. 2012. Ant colony algorithm for

the unsupervised word sense disambiguation of texts: Comparison and evaluation. In *Proceedings of the* 24^{th} *International Conference on Computational Linguistics (COLING)*, pages 8–15, Mumbai, India.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proceedings of ACL 2004 SENSEVAL-3 Workshop*, pages 41–43, Barcelona, Spain.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the* 33^{rd} *Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA, USA.