SemEval-2013 Task 11: Word Sense Induction & Disambiguation within an End-User Application

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

Abstract

In this paper we describe our Semeval-2013 task on Word Sense Induction and Disambiguation within an end-user application, namely Web search result clustering and diversification. Given a target query, induction and disambiguation systems are requested to cluster and diversify the search results returned by a search engine for that query. The task enables the end-to-end evaluation and comparison of systems.

1 Introduction

Word ambiguity is a pervasive issue in Natural Language Processing. Two main techniques in computational lexical semantics, i.e., Word Sense Disambiguation (WSD) and Word Sense Induction (WSI) address this issue from different perspectives: the former is aimed at assigning word senses from a predefined sense inventory to words in context, whereas the latter automatically identifies the meanings of a word of interest by clustering the contexts in which it occurs (see (Navigli, 2009; Navigli, 2012) for a survey).

Unfortunately, the paradigms of both WSD and WSI suffer from significant issues which hamper their success in real-world applications. In fact, the performance of WSD systems depends heavily on which sense inventory is chosen. For instance, the most popular computational lexicon of English, i.e., WordNet (Fellbaum, 1998), provides fine-grained distinctions which make the disambiguation task quite difficult even for humans (Edmonds and Kilgarriff, 2002; Snyder and Palmer, 2004), although disagreements can be solved to some extent with graph-based methods (Navigli, 2008). On the other hand, although WSI overcomes this issue by allowing unrestrained sets of senses, its evaluation is particularly arduous because there is no easy way of comparing and ranking different representations of senses. In fact, all the proposed measures in the literature tend to favour specific cluster shapes (e.g., singletons or all-in-one clusters) of the senses produced as output. Indeed, WSI evaluation is actually an instance of the more general and difficult problem of evaluating clustering algorithms.

Nonetheless, many everyday tasks carried out by online users would benefit from intelligent systems able to address the lexical ambiguity issue effectively. A case in point is Web information retrieval, a task which is becoming increasingly difficult given the continuously growing pool of Web text of the most wildly disparate kinds. Recent work has addressed this issue by proposing a general evaluation framework for injecting WSI into Web search result clustering and diversification (Navigli and Crisafulli, 2010; Di Marco and Navigli, 2013). In this task the search results returned by a search engine for an input query are grouped into clusters, and diversified by providing a reranking which maximizes the meaning heterogeneity of the top ranking results.

The Semeval-2013 task described in this paper[1] adopts the evaluation framework of Di Marco and Navigli (2013), and extends it to both WSD and WSI systems. The task is aimed at overcoming the well-known limitations of *in vitro* evaluations, such as those of previous SemEval tasks on the topic (Agirre

[1] http://www.cs.york.ac.uk/semeval-2013/task11/

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

and Soroa, 2007; Manandhar et al., 2010), and enabling a fair comparison between the two disambiguation paradigms. Key to our framework is the assumption that search results grouped into a given cluster are semantically related to each other and that each cluster is expected to represent a specific meaning of the input query (even though it is possible for more than one cluster to represent the same meaning). For instance, consider the target query *apple* and the following 3 search result snippets:

1. *Apple* Inc., formerly Apple Computer, Inc., is...
2. The science of *apple* growing is called pomology...
3. *Apple* designs and creates iPod and iTunes...

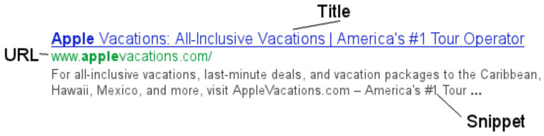
Participating systems were requested to produce a clustering that groups snippets conveying the same meaning of the input query *apple*, i.e., ideally $\{1, 3\}$ and $\{2\}$ in the above example.

2 Task setup

For each ambiguous query the task required participating systems to cluster the top ranking snippets returned by a search engine (we used the Google Search API). WSI systems were required to identify the meanings of the input query and cluster the snippets into semantically-related groups according to their meanings. Instead, WSD systems were requested to sense-tag the given snippets with the appropriate senses of the input query, thereby implicitly determining a clustering of snippets (i.e., one cluster per sense).

2.1 Dataset

We created a dataset of 100 ambiguous queries. The queries were randomly sampled from the AOL search logs so as to ensure that they had been used in real search sessions. Following previous work on the topic (Bernardini et al., 2009; Di Marco and Navigli, 2013) we selected those queries for which a sense inventory exists as a disambiguation page in the English Wikipedia[2]. This guaranteed that the selected queries consisted of either a single word or a multiword expression for which we had a collaboratively-edited list of meanings, including lexicographic and encyclopedic ones. We discarded all queries made up of > 4 words, since the length of the great majority of queries lay in the range $[1, 4]$. In Table 1 we compare the percentage distribution of 1- to 4-word queries in the AOL query logs against our dataset of queries. Note that we increased the percentage of 3- and 4-word queries in order to have a significant coverage of those lengths. Anyhow, in both cases most queries contained from 1 to 2 words. Note that the reported percentage distributions of query length is different from recent statistics for two reasons: first, over the years users have increased the average number of words per query in order to refine their searches; second, we selected only queries which were either single words (e.g., *apple*) or multi-word expressions (e.g., *mortal kombat*), thereby discarding several long queries composed of different words (such as *angelina jolie actress*).



Apple Vacations: All-Inclusive Vacations | America's #1 Tour Operator
www.applevacations.com/
For all-inclusive vacations, last-minute deals, and vacation packages to the Caribbean, Hawaii, Mexico, and more, visit AppleVacations.com – America's #1 Tour ...

Figure 1: An example of search result for the *apple* query, including: page title, URL and snippet.

query length	1	2	3	4
AOL logs	45.89	40.98	10.98	2.32
our dataset	40.00	40.00	15.00	5.00

Table 1: Percentage distribution of AOL query lengths (first row) vs. the queries sampled for our task (second row).

Finally, we submitted each query to Google search and retrieved the 64 top-ranking results returned for each query. Therefore, overall the dataset consists of 100 queries and 6,400 results. Each search result includes the following information: page title, URL of the page and snippet of the page text. We show an example of search result for the *apple* query in Figure 1.

2.2 Dataset Annotation

For each query q we used Amazon Mechanical Turk[3] to annotate each query result with the

[2] http://en.wikipedia.org/wiki/Disambiguation_page

[3] https://www.mturk.com

most suitable sense. The sense inventory for q was obtained by listing the senses available in the Wikipedia disambiguation page of q augmented with additional options from the classes obtained from the section headings of the disambiguation page plus the OTHER catch-all meaning. For instance, consider the *apple* query. We show its disambiguation page in Figure 2. The sense inventory for *apple* was made up of the senses listed in that page (e.g., MALUS, APPLE INC., APPLE BANK, etc.) plus the set of generic classes OTHER PLANTS AND PLANT PARTS, OTHER COMPANIES, OTHER FILMS, plus OTHER.

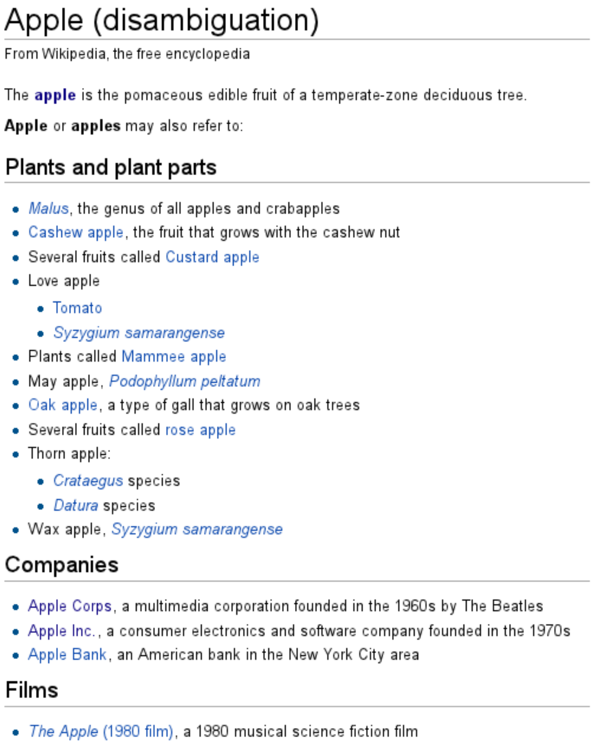
For each query we ensured that three annotators tagged each of the 64 results for that query with the most suitable sense among those in the sense inventory (selecting OTHER if no sense was appropriate). Specifically, each Turker was provided with the following instructions: "The goal is annotating the search result snippets returned by Google for a given query with the appropriate meaning among those available (obtained from the Wikipedia disambiguation page for the query). You have to select the meaning that you consider most appropriate". No constraint on the age, gender and citizenship of the annotators was imposed. However, in order to avoid random tagging of search results, we provided 3 gold-standard result annotations per query, which could be shown to the Turker more than once during the annotation process. In the case (s)he failed to annotate the gold items, the annotator was automatically excluded.

2.3 Inter-Annotator Agreement and Adjudication

In order to determine the reliability of the Turkers' annotations, we calculated the individual values of Fleiss' kappa κ (Fleiss, 1971) for each query q and then averaged them:

$$\overline{\kappa} = \frac{\sum_{q \in Q} \kappa_q}{|Q|}, \quad (1)$$

where κ_q is the Fleiss' kappa agreement of the three annotators who tagged the 64 snippets returned by the Google search engine for the query $q \in Q$, and Q is our set of 100 queries. We obtained an average value of $\overline{\kappa} = 0.66$, which according to Landis and Koch (1977) can be seen as substantial agreement, with a standard deviation $\sigma = 0.185$.



Apple (disambiguation)

From Wikipedia, the free encyclopedia

The **apple** is the pomaceous edible fruit of a temperate-zone deciduous tree.

Apple or **apples** may also refer to:

Plants and plant parts

- *Malus*, the genus of all apples and crabapples
- Cashew apple, the fruit that grows with the cashew nut
- Several fruits called Custard apple
- Love apple
  - Tomato
  - *Syzygium samarangense*
- Plants called Mammee apple
- May apple, *Podophyllum peltatum*
- Oak apple, a type of gall that grows on oak trees
- Several fruits called rose apple
- Thorn apple:
  - *Crataegus* species
  - *Datura* species
- Wax apple, *Syzygium samarangense*

Companies

- Apple Corps, a multimedia corporation founded in the 1960s by The Beatles
- Apple Inc., a consumer electronics and software company founded in the 1970s
- Apple Bank, an American bank in the New York City area

Films

- *The Apple* (1980 film), a 1980 musical science fiction film

Figure 2: The Wikipedia disambiguation page of Apple.

In Table 2 we show the agreement distribution of our 6400 snippets, distinguishing between full agreement (3 out of 3), majority agreement (2 out of 3), and no agreement. Most of the items were annotated with full or majority agreement, indicating that the manual annotation task was generally doable for the layman. We manually checked all the cases of majority agreement, correcting only 7.92% of the majority adjudications, and manually adjudicated all the snippets for which there was no agreement. We observed during adjudication that in many cases the disagreement was due to the existence of subtle sense distinctions, like between MORTAL KOMBAT (VIDEO GAME) and MORTAL KOMBAT (2011 VIDEO GAME), or between THE DA VINCI CODE and INACCURACIES IN THE DA VINCI CODE.

The average number of senses associated with the search results of each query was 7.69 (higher than in previous datasets, such as AMBIENT[4]+MORESQUE[5], which associates 5.07 senses

[4] http://credo.fub.it/ambient

[5] http://lcl.uniroma1.it/moresque

	Full agr.	Majority	Disagr.
% snippets	66.70	25.85	7.45

Table 2: Percentage of snippets with full agreement, majority agreement and full disagreement.

per query on average).

3 Scoring

Following Di Marco and Navigli (2013), we evaluated the systems' outputs in terms of the snippet clustering quality (Section 3.1) and the snippet diversification quality (Section 3.2). Given a query $q \in Q$ and the corresponding set of 64 snippet results, let $\mathcal{C}$ be the clustering output by a given system and let $\mathcal{G}$ be the gold-standard clustering for those results. Each measure $M(\mathcal{C}, \mathcal{G})$ presented below is calculated for the query q using these two clusterings. The overall results on the entire set of queries Q in the dataset is calculated by averaging the values of $M(\mathcal{C}, \mathcal{G})$ obtained for each single test query $q \in Q$.

3.1 Clustering Quality

The first evaluation concerned the quality of the clusters produced by the participating systems. Since clustering evaluation is a difficult issue, we calculated four distinct measures available in the literature, namely:

- Rand Index (Rand, 1971);
- Adjusted Rand Index (Hubert and Arabie, 1985);
- Jaccard Index (Jaccard, 1901);
- F1 measure (van Rijsbergen, 1979).

The *Rand Index* (RI) of a clustering $\mathcal{C}$ is a measure of clustering agreement which determines the percentage of correctly bucketed snippet pairs across the two clusterings $\mathcal{C}$ and $\mathcal{G}$. RI is calculated as follows:

$$RI(\mathcal{C}, \mathcal{G}) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \qquad (2)$$

where TP is the number of true positives, i.e., snippet pairs which are in the same cluster both in $\mathcal{C}$ and $\mathcal{G}$, TN is the number of true negatives, i.e., pairs which are in different clusters in both clusterings, and FP and FN are, respectively, the number of false positives and false negatives. RI ranges between 0 and 1, where 1 indicates perfect correspondence.

$\mathcal{C}$ $\mathcal{G}$	C_1	C_2	$\cdots$	C_m	**Sums**
G_1	n_{11}	n_{12}	$\cdots$	n_{1m}	a_1
G_2	n_{21}	n_{22}	$\cdots$	n_{2m}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
G_g	n_{g1}	n_{g2}	$\cdots$	n_{gm}	a_g
Sums	b_1	b_2	$\cdots$	b_m	N

Table 3: Contingency table for the clusterings $\mathcal{G}$ and $\mathcal{C}$.

Adjusted Rand Index (ARI) is a development of Rand Index which corrects the RI for chance agreement and makes it vary according to expectaction:

$$ARI(\mathcal{C}, \mathcal{G}) = \frac{RI(\mathcal{C}, \mathcal{G}) - E(RI(\mathcal{C}, \mathcal{G}))}{\max RI(\mathcal{C}, \mathcal{G}) - E(RI(\mathcal{C}, \mathcal{G}))}. \qquad (3)$$

where $E(RI(\mathcal{C}, \mathcal{G}))$ is the expected value of the RI. Using the contingency table reported in Table 3 we can quantify the degree of overlap between $\mathcal{C}$ and $\mathcal{G}$, where n_{ij} denotes the number of snippets in common between G_i and C_j (namely, $n_{ij} = |G_i \cap C_j|$), a_i and b_j represent, respectively, the number of snippets in G_i and C_j, and N is the total number of snippets, i.e., $N = 64$. Now, the above equation can be reformulated as:

$$ARI(\mathcal{C},\mathcal{G}) = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{N}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{N}{2}}. \qquad (4)$$

The ARI ranges between -1 and $+1$ and is 0 when the index equals its expected value.

Jaccard Index (JI) is a measure which takes into account only the snippet pairs which are in the same cluster both in $\mathcal{C}$ and $\mathcal{G}$, i.e., the true positives (TP), while neglecting true negatives (TN), which are the vast majority of cases. JI is calculated as follows:

$$JI(\mathcal{C}, \mathcal{G}) = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \qquad (5)$$

Finally, the *F1 measure* calculates the harmonic mean of precision (P) and recall (R). Precision determines how accurately the clusters of $\mathcal{C}$ represent the query meanings in the gold standard $\mathcal{G}$, whereas recall measures how accurately the different meanings in $\mathcal{G}$ are covered by the clusters in $\mathcal{C}$. We follow Crabtree et al. (2005) and define the precision of a cluster $C_j \in \mathcal{C}$ as follows:

$$P(C_j) = \frac{|C_j^s|}{|C_j|}, \tag{6}$$

where C_j^s is the intersection between $C_j \in \mathcal{C}$ and the gold cluster $G_s \in \mathcal{G}$ which maximizes the cardinality of the intersection. The recall of a query sense s is instead calculated as:

$$R(s) = \frac{|\bigcup_{C_j \in \mathcal{C}^s} C_j^s|}{n_s}, \tag{7}$$

where $\mathcal{C}^s$ is the subset of clusters of $\mathcal{C}$ whose majority sense is s, and n_s is the number of snippets tagged with query sense s in the gold standard. The total precision and recall of the clustering $\mathcal{C}$ are then calculated as:

$$P = \frac{\sum_{C_j \in \mathcal{C}} P(C_j)|C_j|}{\sum_{C_j \in \mathcal{C}} |C_j|}; \quad R = \frac{\sum_{s \in \mathcal{S}} R(s) n_s}{\sum_{s \in \mathcal{S}} n_s} \tag{8}$$

where $\mathcal{S}$ is the set of senses in the gold standard $\mathcal{G}$ for the given query (i.e., $|\mathcal{S}| = |\mathcal{G}|$). The two values of P and R are then combined into their harmonic mean, namely the F1 measure:

$$F1(\mathcal{C}, \mathcal{G}) = \frac{2PR}{P+R}. \tag{9}$$

3.2 Clustering Diversity

Our second evaluation is aimed at determining the impact of the output clustering on the diversification of the top results shown to a Web user. To this end, we applied an automatic procedure for flattening the clusterings produced by the participating systems to a list of search results. Given a clustering $\mathcal{C} = (C_1, C_2, \ldots, C_m)$, we add to the initially empty list the first element of each cluster C_j ($j = 1, \ldots, m$); then we iterate the process by selecting the second element of each cluster C_j such that $|C_j| \geq 2$, and so on. The remaining elements returned by the search engine, but not included in any cluster of $\mathcal{C}$, are appended to the bottom of the list in their original order. Note that systems were asked to sort snippets within clusters, as well as clusters themselves, by relevance.

Since our goal is to determine how many different meanings are covered by the top-ranking search results according to the output clustering, we used the measures of S-recall@K (Subtopic recall at rank K) and S-precision@r (Subtopic precision at recall r) (Zhai et al., 2003).

S-recall@K determines the ratio of different meanings for a given query q in the top-K results returned:

$$\text{S-recall@}K = \frac{|\{sense(r_i) : i \in \{1, \ldots, K\}\}|}{g}, \tag{10}$$

where $sense(r_i)$ is the gold-standard sense associated with the i-th snippet returned by the system, and g is the total number of distinct senses for the query q in our gold standard.

S-precision@r instead determines the ratio of different senses retrieved for query q in the first K_r snippets, where K_r is the minimum number of top results for which the system achieves recall r. The measure is defined as follows:

$$\text{S-precision@}r = \frac{|\cup_{i=1}^{K_r} sense(r_i)|}{K_r}. \tag{11}$$

3.3 Baselines

We compared the participating systems with two simple baselines:

- SINGLETONS: each snippet is clustered as a separate singleton cluster (i.e., $|\mathcal{C}| = 64$).
- ALL-IN-ONE: all snippets are clustered into a single cluster (i.e., $|\mathcal{C}| = 1$).

These baselines are important in that they make explicit the preference of certain quality measures towards clusterings made up with a small or large number of clusters.

4 Systems

5 teams submitted 10 systems, out of which 9 were WSI systems, while 1 was a WSD system, i.e., using the Wikipedia sense inventory for performing the disambiguation task. All systems could exploit the information provided for each search result, i.e., URL, page title and result snippet. WSI systems were requested to use unannotated corpora only.

	System	URLs	Snippets	Wikipedia	YAGO Hierarchy	Distr. Thesaurus	Other
WSI	HDP-CLUSTERS-LEMMA		✓	✓			
	HDP-CLUSTERS-NOLEMMA		✓	✓			
	DULUTH.SYS1.PK2		✓				
	DULUTH.SYS7.PK2		✓				
	DULUTH.SYS9.PK2						Gigaword
	UKP-WSI-WP-LLR2	✓		✓		✓	WaCky
	UKP-WSI-WP-PMI	✓		✓		✓	WaCky
	UKP-WSI-WACKY-LLR	✓		✓		✓	WaCky
	SATTY-APPROACH1		✓				
WSD	RAKESH				✓		DBPedia

Table 4: Resources used for WSI/WSD.

We asked each team to provide information about their systems. In Table 4 we report the resources used by each system. The HDP and UKP systems use Wikipedia as raw text for sampling word counts; DULUTH-SYS9-PK2 uses the first 10,000 paragraphs of the Associated Press wire service data from the English Gigaword Corpus (Graff, 2003, 1st edition), whereas DULUTH-SYS1-PK2 and DULUTH-SYS7-PK2 both use the snippets for inducing the query senses. Finally, the UKP systems were the only ones to retrieve the Web pages from the corresponding URLs and exploit them for WSI purposes. They also use WaCky (Baroni et al., 2009) and a distributional thesaurus obtained from the Leipzig Corpora Collection[6] (Biemann et al., 2007). SATTY-APPROACH1 just uses snippets.

The only participating WSD system, RAKESH, uses the YAGO hierarchy (Suchanek et al., 2008) together with DBPedia abstracts (Bizer et al., 2009).

5 Results

We show the results of RI and ARI in Table 5. The best performing systems are those from the HDP team, with considerably higher RI and ARI. The next best systems are SATTY-APPROACH1, which uses only the words in the snippets, and the only WSD system, i.e., RAKESH. SINGLETONS perform well with RI, but badly when chance agreement is taken into account.

As for F1 and JI, whose values are shown in Table 6, the two HDP systems again perform best in terms of F1, and are on par with UKP-WSI-WACKY-LLR in terms of JI. The third best approach in terms of F1 is again SATTY-APPROACH1, which however performs badly in terms of JI. The SINGLETONS baseline clearly obtains the best F1 performance, but the worst JI results. The ALL-IN-ONE baseline outperforms all other systems with the JI measure, because TN are not considered, which favours large clusters.

	System	RI	ARI
WSI	HDP-CLUSTERS-LEMMA	**65.22**	21.31
	HDP-CLUSTERS-NOLEMMA	64.86	**21.49**
	SATTY-APPROACH1	59.55	7.19
	DULUTH.SYS9.PK2	54.63	2.59
	DULUTH.SYS1.PK2	52.18	5.74
	DULUTH.SYS7.PK2	52.04	6.78
	UKP-WSI-WP-LLR2	51.09	3.77
	UKP-WSI-WP-PMI	50.50	3.64
	UKP-WSI-WACKY-LLR	50.02	2.53
WSD	RAKESH	58.76	8.11
BL	SINGLETONS	**60.09**	0.00
	ALL-IN-ONE	39.90	0.00

Table 5: Results for Rand Index (RI) and Adjusted Rand Index (ARI), sorted by RI.

To get more insights into the performance of the various systems, we calculated the average number of clusters per clustering produced by each system and compared it with the gold standard average. We also computed the average cluster size, i.e., the average number of snippets per cluster. The statistics are shown in Table 7. Interestingly, the best performing systems are those with the cluster number and average number of clusters closest to the gold standard ones. This finding is also confirmed by Figure 3, where we draw each system according to its average values regarding cluster number and size: again the distance from the gold standard is meaningful.

We now move to the diversification perfor-

[6] http://corpora.uni-leipzig.de/

	System	JI	F1
WSI	UKP-WSI-WACKY-LLR	**33.94**	58.26
	HDP-CLUSTERS-NOLEMMA	33.75	68.03
	HDP-CLUSTERS-LEMMA	33.02	**68.30**
	DULUTH.SYS1.PK2	31.79	56.83
	UKP-WSI-WP-LLR2	31.77	58.64
	DULUTH.SYS7.PK2	31.03	58.78
	UKP-WSI-WP-PMI	29.32	60.48
	DULUTH.SYS9.PK2	22.24	57.02
	SATTY-APPROACH1	15.05	67.09
WSD	RAKESH	30.52	39.49
BL	SINGLETONS	0.00	**100.00**
	ALL-IN-ONE	**39.90**	54.42

Table 6: Results for Jaccard Index (JI) and F1 measure.

	System	# cl.	ACS
	GOLD STANDARD	7.69	11.56
WSI	HDP-CLUSTERS-LEMMA	**6.63**	**11.07**
	HDP-CLUSTERS-NOLEMMA	**6.54**	**11.68**
	SATTY-APPROACH1	9.90	6.46
	UKP-WSI-WP-PMI	5.86	30.30
	DULUTH.SYS7.PK2	3.01	25.15
	UKP-WSI-WP-LLR2	4.17	21.87
	UKP-WSI-WACKY-LLR	3.64	32.34
	DULUTH.SYS9.PK2	3.32	19.84
	DULUTH.SYS1.PK2	2.53	26.45
WSD	RAKESH	9.07	2.94

Table 7: Average number of clusters (# cl.) and average cluster size (ACS).

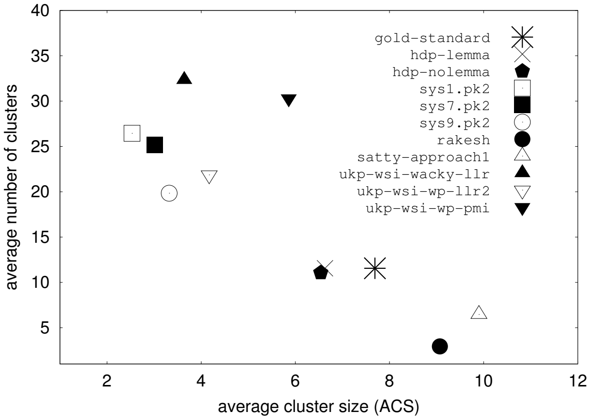


Figure 3: Average cluster size (ACS) vs. average number of clusters.

	System	K			
		5	10	20	40
WSI	HDP-CL.-NOLEMMA	**50.80**	63.21	**79.26**	**92.48**
	HDP-CL.-LEMMA	48.13	**65.51**	78.86	91.68
	UKP-WACKY-LLR	41.19	55.41	68.61	83.90
	UKP-WP-LLR2	41.07	53.76	68.87	85.87
	UKP-WP-PMI	40.45	56.25	68.70	84.92
	SATTY-APPROACH1	38.97	48.90	62.72	82.14
	DULUTH.SYS7.PK2	38.88	53.79	70.38	86.23
	DULUTH.SYS9.PK2	37.15	49.90	68.91	83.65
	DULUTH.SYS1.PK2	37.11	53.29	71.24	88.48
WSD	RAKESH	46.48	62.36	78.66	90.72

Table 8: S-recall@K.

	System	r			
		50	60	70	80
WSI	HDP-CL.-LEMMA	**48.85**	42.93	**35.19**	27.62
	HDP-CL.-NOLEMMA	48.18	**43.88**	34.85	**29.30**
	UKP-WP-PMI	42.83	33.40	26.63	22.92
	UKP-WACKY-LLR	42.47	31.73	25.39	22.71
	UKP-WP-LLR2	42.06	32.04	26.57	22.41
	DULUTH.SYS1.PK2	40.08	31.31	26.73	24.51
	DULUTH.SYS7.PK2	39.11	30.42	26.54	23.43
	DULUTH.SYS9.PK2	35.90	29.72	25.26	21.26
	SATTY-APPROACH1	34.94	26.88	23.55	20.40
WSD	RAKESH	48.00	39.04	32.72	27.92

Table 9: S-precision@r.

mance, calculated in terms of S-recall@K and S-precision@r, whose results are shown in Tables 8 and 9, respectively. Here we find that, again, the HDP team obtains the best performance, followed by RAKESH. We note however that not all systems optimized the order of clusters and cluster snippets by relevance.

We also graph the diversification performance trend of S-recall@K and S-precision@r in Figures 4 and 5 for $K = 1, \ldots, 25$ and $r \in \{40, 50, \ldots, 100\}$.

6 Conclusions and Future Directions

One of the aims of the SemEval-2013 task on Word Sense Induction & Disambiguation within an End User Application was to enable an objective comparison of WSI and WSD systems when integrated into Web search result clustering and diversification. The task is a hard one, in that it involves clustering, but provides clear-cut evidence that our end-to-end application framework overcomes the limits of previous in-vitro evaluations. Indeed, the systems which create good clusters and better diversify search results, i.e., those from the HDP team, achieve good performance across all the proposed measures, with no contradictory evidence.

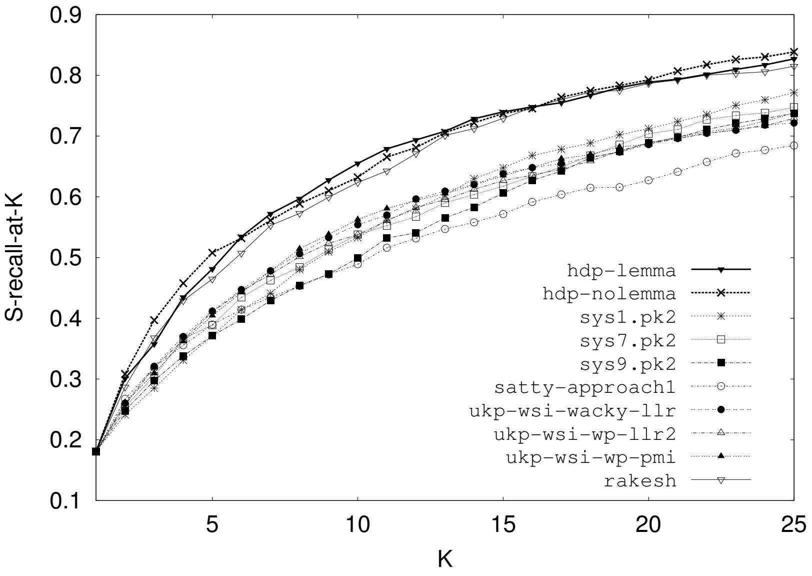


Figure 4: S-recall@K.

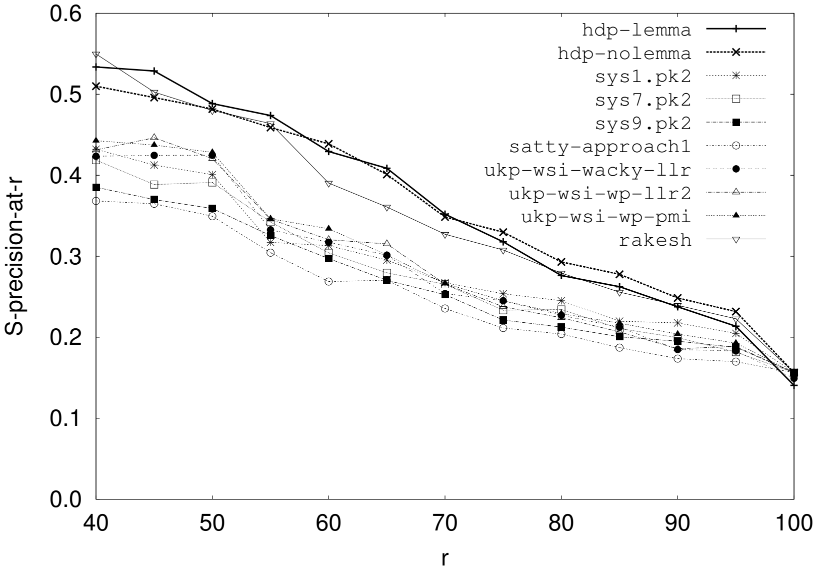


Figure 5: S-precision@r.

Our annotation experience showed that the Wikipedia sense inventory, augmented with our generic classes, is a good choice for semantically tagging search results, in that it covers most of the meanings a Web user might be interested in. In fact, only 20% of the snippets was annotated with the OTHER class.

Future work might consider large-scale multilingual lexical resources, such as BabelNet (Navigli and Ponzetto, 2012), both as sense inventory and for performing the search result clustering and diversification task.

Acknowledgments

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

We thank Antonio Di Marco and David A. Jurgens for their help.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4^{th} International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Andrea Bernardini, Claudio Carpineto, and Massimiliano D'Amico. 2009. Full-subtopic retrieval with keyphrase-based search results clustering. In *Proceedings of Web Intelligence 2009*, volume 1, pages 206–213, Los Alamitos, CA, USA.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig corpora collection - monolingual corpora of standard size. In *Proceedings of Corpus Linguistic 2007*, Birmingham, UK.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165.

Daniel Crabtree, Xiaoying Gao, and Peter Andreae. 2005. Improving web clustering by cluster selection. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 172–178, Washington, DC, USA.

Antonio Di Marco and Roberto Navigli. 2013. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3).

Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8(4):279–291.

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

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. In *Psychological Bulletin*, volume 76, page 378–382.

David Graff. 2003. English Gigaword. In *Technical Report, LDC2003T05, Linguistic Data Consortium*, Philadelphia, PA, USA.

Lawrence Hubert and Phipps Arabie. 1985. Comparing Partitions. *Journal of Classification*, 2(1):193–218.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. In *Bulletin de la Société Vaudoise des Sciences Naturelles*, volume 37, page 547–579.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Suresh Manandhar, Ioannis P. Klapaftis, Dmitriy Dligach, and Sameer S. Pradhan. 2010. SemEval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden.

Roberto Navigli and Giuseppe Crisafulli. 2010. Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 116–126, Boston, USA.

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

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

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

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of the 3^{rd} International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 41–43, Barcelona, Spain.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203–217.

Cornelis Joost van Rijsbergen. 1979. *Information Retrieval*. Butterworths, second edition.

ChengXiang Zhai, William W. Cohen, and John Lafferty. 2003. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17, Toronto, Canada.