Paving the Way to a Large-scale Pseudosense-annotated Dataset

Mohammad Taher Pilehvar and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
{pilehvar,navigli}@di.uniroma1.it

Abstract

In this paper we propose a new approach to the generation of pseudowords, i.e., artificial words which model real polysemous words. Our approach simultaneously addresses the two important issues that hamper the generation of large pseudosense-annotated datasets: semantic awareness and coverage. We evaluate these pseudowords from three different perspectives showing that they can be used as reliable substitutes for their real counterparts.

1 Introduction

A fundamental problem in computational linguistics is the paucity of manually annotated data, such as part-of-speech tagged sentences, treebanks, and logical forms, which exist only for few languages (Ide et al., 2010). A case in point is the lack of abundant sense annotated data, which hampers the performance and coverage of lexical semantic tasks such as Word Sense Disambiguation (Navigli, 2009; Navigli, 2012, WSD) and semantic role labeling (Gildea and Jurafsky, 2002). A possible way to break this bottleneck is to use pseudowords, i.e., artificial words constructed by conflating a set of unambiguous words, with the aim of modeling polysemy in real ambiguous words. The idea of pseudowords was originally proposed by Gale et al. (1992) and Schütze (1992) for WSD evaluation, but later found application in other tasks such as selectional preferences (Erk, 2007; Bergsma et al., 2008; Chambers and Jurafsky, 2010), Word Sense Induction (Bordag, 2006; Di Marco and Navigli, 2013) or studies concerning the effects of the amount of data on machine learning for natural language disambiguation (Banko and Brill, 2001). Being made up of monosemous words, pseudowords can potentially be used to create large amounts of pseudosense-annotated data at virtually no cost, hence enabling large-scale studies in lexical semantics. Unfortunately, though, the extent of their usability for such a purpose is hampered by two main issues: semantic awareness and wide coverage.

Semantic awareness corresponds to the constraint that pseudowords, in order to be realistic, are expected to have senses which are in a semantic relationship (thus modeling systematic polysemy). Recent work has focused on this issue and, by exploiting either specific lexical hierarchies (Nakov and Hearst, 2003; Lu et al., 2006), or the WordNet structure (Otrusina and Smrz, 2010), have succeeded in generating pseudowords which are comparable to real words in terms of disambiguation difficulty. The second challenge is coverage, which corresponds to the number of distinct pseudowords an algorithm can generate. When coupled with the semantic awareness issue, wide coverage is hampered by the difficulty in generating thousands of pseudowords which mimic existing polysemous words.

Unfortunately, none of the existing approaches to the generation of pseudowords can meet both these challenges simultaneously, and this has hindered the generation of a large pseudosense-annotated dataset. For instance, approaches which exploit the monosemous neighbors of a target sense in WordNet (Otrusina and Smrz, 2010) can be used to generate pseudowords with good semantic awareness, but

Proceedings of NAACL-HLT 2013, pages 1100–1109,
Atlanta, Georgia, 9–14 June 2013.

they have low coverage of ambiguous nouns when many pseudosense-tagged sentences are needed (cf. Section 2.1.1).

In this paper we propose a new approach, based on Personalized PageRank, which simultaneously addresses the two above-mentioned issues concerning the generation of pseudowords (i.e., semantic awareness and coverage), and hence enables the generation of large-scale pseudosense-annotated datasets. We perform three different experiments to show that our pseudowords are good at modeling existing ambiguous words in terms of disambiguation difficulty, representativeness of real senses and distinguishability of the artificial senses. As a byproduct of this work, we generate a large dataset that provides 1000 tagged sentences for each of the 15,935 pseudowords modeled after real ambiguous nouns in WordNet 3.0.

2 Pseudowords

A pseudoword $p = w_1{*}w_2{*}\ldots{*}w_n$ is an artificially-generated ambiguous word of polysemy degree n which is usually created by conflating n unique unambiguous words w_i called pseudosenses. For instance, *airplane*river* is a pseudoword with two meanings explicitly identified by its pseudosenses: *airplane* and *river*. Pseudowords are particularly interesting as they can be used to introduce controlled artificial ambiguity into a corpus. Given a pseudoword p and an untagged corpus C, this artificial tagging is achieved by substituting all occurrences of w_i in C with p for each pseudosense $i \in \{1, \ldots, n\}$. As a result, each occurrence of the pseudoword p is tagged with the underlying sense w_i. As an example, consider the following two sentences:

a1. The Wright brothers invented the *airplane*.

a2. The Nile is the longest *river* in the world.

If we replace the individual occurrences of *airplane* and *river* with the pseudoword *airplane*river* while noting the replaced term as the corresponding sense, we obtain the following pseudosense-tagged sentences:

b1. The Wright brothers invented the ***airplane****river*.

b2. The Nile is the longest *airplane**__river__ in the world.

As a result of this procedure, we obtain a corpus of sentences containing the occurrences of an artificially ambiguous word p, for each of which we know its correct sense annotation w_i. Virtually any number of pseudowords can be created, resulting in a large pseudosense-annotated corpus. An obvious restriction on the choice of pseudosenses is that they need to be unambiguous, so as to avoid the introduction of uncontrolled ambiguity. Another constraint is that the constituent w_i must satisfy a minimum occurrence frequency in the corpus C. This minimum frequency corresponds to the number of annotated sentences that are requested for the task of interest which will exploit the resulting annotated corpus.

An immediate way of generating a pseudoword would be to randomly select its constituents from the set of all monosemous words given by a lexicon (e.g., WordNet). However, constructing a pseudoword by merely combining a random set of unambiguous words selected on the basis of their falling in the same range of occurrence frequency (Schütze, 1992), or leveraging homophones and OCR ambiguities (Yarowsky, 1993), does not provide a suitable model of a real polysemous word (Gaustad, 2001; Nakov and Hearst, 2003). This is because in the real world different senses, unless they are homonymous, share some semantic or pragmatic relation. Therefore, random pseudowords will typically model only homonymous distinctions (such as the *centimeter* vs. *curium* senses of *cm*), while they will fall short of modeling systematic polysemy (such as the *lack* vs. *insufficiency* senses of *deficiency*).

2.1 Semantically-aware Pseudowords

In order to cope with the above-mentioned limits of random pseudowords, an artificial word has to model an existing word by providing a one-to-one correspondence between each pseudosense and a corresponding sense of the modeled word. For instance, the pseudoword *lack*shortfall* is a good model of the real word *deficiency* in that its pseudosenses preserve the meanings of their corresponding real word's senses. We call this kind of artificial words semantically-aware pseudowords.

In the next two subsections, we will describe two techniques (the second of which is presented for the first time in this paper) for the generation of

Minimum Frequency	Polysemy												Overall
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**>12**	
0	87	82	74	71	67	70	60	64	45	46	44	28	83
500	41	31	24	15	12	13	10	7	7	0	0	0	35
1000	31	20	16	7	4	6	4	3	0	0	0	0	25

Table 1: Ambiguous noun coverage percentage of vicinity-based pseudowords by degree of polysemy for different values of minimum pseudosense occurrence frequency in Gigaword.

semantically-aware pseudowords. In what follows we focus on nominal pseudowords, and leave the extension to other parts of speech to future work.

2.1.1 Vicinity-based Pseudowords

A computational lexicon such as WordNet (Fellbaum, 1998) can be used as the basis for the automatic generation of semantically-aware pseudowords, an idea which was first proposed by Otrusina and Smrz (2010). WordNet can be viewed as a graph in which synsets act as nodes and the lexical and semantic relationships among them as edges. Given a sense, the approach looks into its surrounding synsets in the WordNet graph in order to find a related monosemous term that can represent that sense. As search space, the approach considers: the other literals in the same synset, the genus phrase from its textual definition, direct siblings, and direct hyponyms. If no monosemous candidate can be found, this space is further extended to hypernyms and meronyms. Hereafter, we term this approach as vicinity-based.

For example, consider the generation process of the vicinity-based pseudoword corresponding to the term *coke*, which has three senses in WordNet 3.0. There exist multiple monosemous candidates for each sense: dozens of candidates (such as *biomass* and *butane*) in the direct siblings' vicinity of the first sense, *coca_cola*, *pepsi*, and *pepsi_cola* for the second sense, and *nose_candy* and *coca_cola* for the third sense. Among these candidates Otrusina and Smrz (2010) select those whose occurrence frequency ratio in a given text corpus is most similar to that of the senses of the corresponding real word as given by a sense-annotated corpus. Clearly, a sufficiently large sense-tagged corpus is required for calculating the occurrence frequency of the individual senses of a word. This is a limitation of the vicinity-based approach.

In addition, as we mentioned earlier, we need pseudowords that can enable the generation of large-scale pseudosense-tagged corpora. For this to be achieved, each pseudosense is required to occur with a relatively high frequency in a given text corpus. The vicinity-based approach can, however, identify at best only a few representatives for each pseudosense, thus undermining its ability to cover many ambiguous nouns. Table 1 shows the percentage of ambiguous nouns in WordNet that can be modeled using the vicinity-based approach when different minimum numbers of annotated sentences are requested, i.e. each pseudosense is required to occur in at least 0 (i.e., no minimum frequency restriction), 500, or 1000 unique sentences in the reference corpus (we use Gigaword (Graff and Cieri, 2003) in our experiments). In the Table, beside the overall coverage percentage, we present the coverage by degree of polysemy and for three different values of minimum pseudosense occurrence frequency. Even though the overall coverage is over 80% when no restriction on minimum frequency is considered (first row in the Table), this high coverage drops rapidly when we request some hundred sentences per sense. For instance, only 25% of the ambiguous nouns in WordNet can be modeled using this approach when a minimum frequency of 1000 noun occurrences is required (last row of Table 1), with most of the covered words having low polysemy (in fact about 93% of them are either 2- or 3-sense nouns). This severe limitation of the vicinity-based approach hinders a wide-coverage modeling of ambiguous nouns in WordNet, thus preventing it from being an option for the generation of a large-scale pseudosense-annotated dataset.

With a view to addressing the above-mentioned issues and to enable wide coverage, in the next subsection we propose a flexible approach for the generation of semantically-aware pseudowords.

2.1.2 Similarity-based Pseudowords

The vicinity-based pseudoword generation approach works on local subgraphs of WordNet, considering mostly all those candidates which are in a direct relationship with a real sense s_i, and treating them as potentially good representatives of s_i. We propose an extension to this approach which exploits the WordNet semantic network in its entirety, hence enabling us to determine a graded degree of similarity between s_i and all the senses of all other words in WordNet.

We chose a graph-based similarity measure for two reasons: firstly, it comes as a natural extension of the vicinity-based method, and, secondly, alternative context-based methods such as Lin's measure (Lin, 1998) have been shown to require a wide-coverage sense-tagged dataset in order to calculate similarities on a sense-by-sense basis for all words in the lexicon (Otrusina and Smrz, 2010). As our similarity measure we selected the Personalized PageRank (Haveliwala, 2002, PPR) algorithm. PPR basically computes the probability according to which a random walker at a specific node in a graph would visit an arbitrary node in the same graph. The algorithm estimates, for a specific node in a graph, a probability distribution (called PPR vector) which determines the importance of any given node in the graph for that specific node. When applied to a semantic graph, this importance can be interpreted as semantic similarity. PPR has previously been used as a core component for semantic similarity[1] (Hughes and Ramage, 2007; Agirre et al., 2009) and Word Sense Disambiguation (Agirre and Soroa, 2009).

Algorithm 1 shows the procedure for the generation of our similarity-based pseudowords. The algorithm takes an ambiguous word w as input, and outputs its corresponding similarity-based pseudoword P_w whose i^{th} pseudosense models the i^{th} sense of w, together with a confidence score which we detail below.

Given w, the algorithm iterates over the synsets corresponding to its individual senses (lines 4-13) and finds the most suitable pseudosenses for P_w. For each synset s of w, we start the PPR algorithm from s (line 5) and collect the probability distribution vector output by PPR (*similarSynsets* in the algorithm), which determines the probability of reaching each synset in WordNet starting from s. We then sort this vector (line 6) and check if each of its nominal synsets (s') contains a monosemous word (line 10). This search continues until a suitable candidate is found that satisfies a certain minimum occurrence frequency *minFreq*. When this occurs, the selected monosemous candidate w' is saved as the corresponding pseudosense for the i^{th} sense of P_w (line 11). We iterate these steps for all synsets of w.

Algorithm 1 Generate a similarity-based pseudoword

Input: an ambiguous word w in WordNet
Output: a "similarity-based" pseudoword P_w
a confidence score *averageRank*

1: $P_w \leftarrow \emptyset$
2: *totalRank* $\leftarrow 0$
3: $i \leftarrow 1$
4: **for each** $s \in$ *Synsets*(w)
5: *similarSynsets* $\leftarrow$ *PersonalizedPageRank*(s)
6: sort *similarSynsets* in descending order
7: **for each** $s' \in$ *similarSynsets*
8: *totalRank* $\leftarrow$ *totalRank* $+ 1$
9: **for each** $w' \in$ *SynsetLiterals*(s')
10: **if** $|$*Synsets*$(w')|$=1 **&** *Freq*(w')>*minFreq* **then**
11: $P_w \leftarrow P_w \cup \{(i, w')\}$
12: **break**
13: $i \leftarrow i + 1$
14: *averageRank* $\leftarrow$ *totalRank*$/|$*Synsets*$(w)|$
15: **return** $(P_w,$ *averageRank*$)$

In line 14 we calculate the *averageRank*, a value given by the average of synset positions in the *similarSynsets* lists from which the pseudosenses of P_w are picked out. We later use this value as a confidence score while evaluating our pseudowords. Finally, the algorithm returns the corresponding pseudoword P_w along with its *averageRank* score (line 15). We show in Table 2 some examples of ambiguous words together with their similarity-based pseudowords.

Thanks to the large search space of our similarity-based approach, we are always able to select a monosemous candidate for each pseudosense, thus resolving the coverage issue regarding vicinity-based pseudowords. A question that arises here is that of how often our algorithm needs to resort to lower-ranking items in the *similarSynsets* list. To

[1] Top-ranking synsets will contain words which are most likely similar to the target sense, whereas we move to a graded notion of relatedness as far as lower-ranking ones are concerned (Agirre et al., 2009).

Word	Similarity-based Pseudoword
bernoulli	physicist*mathematician*__astronomer__
coach	football_coach*tutor*passenger_car*clarence*public_transport
green	__greenery__*common*labor_leader*green_party*river*golf_course*greens*__max__
horoscope	forecast*diagram
sunray	sunbeam*__vine__*sunlight
lifter	athlete*thief

Table 2: Similarity-based pseudowords generated for six different nouns in WordNet 3.0 (with minimum frequency of 1000 occurrences in Gigaword). Pseudosenses which could not be modeled using the vicinity-based approach are shown in bold.

verify this, we analyzed the *averageRank* values output by Algorithm 1. Table 3 shows for each polysemy degree and for three different values of *minFreq*, the mean and mode statistics of the *averageRank* scores of the generated similarity-based pseudowords for all the 15,935 polysemous nouns in WordNet 3.0. As expected, the higher the number of required sentences per pseudosense (*minFreq*), the further the algorithm descends through the list *similarSynsets* to select a pseudosense. However, as can be seen from the mode statistics in the Table, even when *minFreq* is set to a large value, most of the pseudosenses are picked from the highest-ranking positions in the *similarSynsets* list.

3 Evaluation

Our novel similarity-based algorithm for the generation of pseudowords inherently tackles the coverage issue. To test whether our generated pseudowords also cope with the issue of semantic awareness we carried out three separate evaluations so as to assess their strength in modeling semantic properties of their corresponding real senses from different perspectives. These will be described in the next three subsections. Since our aim was to leverage pseudowords for the creation of a large-scale pseudosense-annotated dataset, we performed evaluations on pseudowords generated with *minFreq* per pseudosense set to 1000 (i.e., we can generate at least 1000 annotated sentences for each pseudosense).

minFreq	0		500		1000	
poly.	mean	mode	mean	mode	mean	mode
2	2.0	1.0	14.8	2.0	25.4	4.0
3	2.3	1.7	13.4	2.7	21.0	5.5
4	2.3	1.8	12.3	5.8	19.8	6.8
5	2.3	1.8	12.9	5.6	20.0	10.0
6	2.4	2.0	13.7	4.5	18.7	8.8
7	2.3	2.1	11.5	6.3	16.0	6.1
8	2.2	1.8	11.3	9.6	17.2	10.8
9	2.4	2.0	10.7	10.9	15.6	15.1
10	2.2	2.0	10.1	7.0	14.3	12.1
11	2.4	2.1	10.2	7.1	14.2	17.3
12	2.5	2.4	11.0	4.4	14.4	14.4
>12	2.6	1.0	9.3	2.0	13.7	4.0
overall	2.1	1.0	14.1	2.0	23.4	4.0

Table 3: Statistics of *averageRank* scores for the full set of 15,935 similarity-based pseudowords modeled after ambiguous nouns in WordNet 3.0: we show mean and mode statistics for three different values of minimum occurrence frequency (0, 500, and 1000). We show the average value in the case of multiple modes.

3.1 Disambiguation Difficulty of Pseudowords

Our first experiment is an extrinsic evaluation of pseudowords. Ideally, pseudowords are expected to show a similar degree of difficulty to real ambiguous words in a disambiguation task (Otrusina and Smrz, 2010; Lu et al., 2006). We thus experimentally tested this assumption on similarity-based and random pseudowords. Given its low coverage, we excluded the vicinity-based approach from this experiment.

Starting from a sense-tagged lexical sample dataset for a set of ambiguous nouns, for each such noun and for each kind of pseudoword, we automatically generated a pseudosense-annotated dataset by enforcing the same sense distribution as the corresponding real ambiguous noun. This constraint was particularly important for random pseudowords since they do not model the corresponding real ambiguous words (see Section 2). An analysis was then performed to compare the disambiguation performance of a supervised WSD system on a given ambiguous word against its corresponding pseudoword.

Specifically, for our manually sense-tagged corpus we used the Senseval-3 English lexical sample dataset (Mihalcea et al., 2004), which contains 3593 and 1807 sense-tagged sentences for 20 ambiguous nouns (with an average polysemy degree of 5.8) in its training and test sets, respectively. We generated,

with *minFreq* = 1000, the similarity-based pseudowords corresponding to these 20 nouns, as well as a set of 20 random pseudowords with the same polysemy degrees. We note that, in this setting, the vicinity-based approach could only generate pseudowords corresponding to 5 of the 20 nouns.

In order to create the datasets for our experiments, for each of our similarity-based and random pseudowords, we sampled unique sentences from the English Gigaword corpus (Graff and Cieri, 2003) according to the same sense distributions given by the Senseval-3 training and test datasets for the corresponding real word. Next, we performed WSD on our three datasets, namely: the Senseval-3 dataset of real words, and the two artificially sense-tagged datasets for the similarity-based and random pseudowords. As our WSD system for this experiment, we used It Makes Sense (IMS), a state-of-the-art supervised WSD system (Zhong and Ng, 2010).

WSD recall[2] performance values on the above-mentioned datasets are shown in Table 4. For the random setting, in order to ensure stability, the results are averaged on a set of 25 different pseudowords modeling a given ambiguous noun. We can see from the Table that the overall system performance with the similarity-based pseudowords (75.14%) is much closer to the real setting (73.26%) than it is with random pseudowords (78.80%). For random pseudowords, the overall recall over 25 runs ranges from 75.40% to 80.80%.

Moreover, the similarity-based approach exhibits a closer WSD recall performance to that of real data (|RE−SB| column in the table) for 15 of the 20 nouns (shown in bold in the Table). Accordingly, the overall sum of the differences (distance) between the recall values is 129.3 for similarity-based pseudowords, which is considerably lower than the 196.4 for random pseudowords (averaged over 25 runs whose distances range from 158.3 to 262.0).

To further corroborate our findings, we calculated the Pearson's r correlation between recall values on real words with those obtained on the corresponding pseudowords. Similarity-based pseudowords obtain the high correlation of 0.74, whereas this value drops to 0.54 for random pseudowords. Even worse, we observed a high variation of correlation (in the range of [0.18, 0.67]) over the 25 sets of random pseudowords (0.54 being the average).

[2]Since in our experiments the WSD system always provides an answer for each item in the test set, the values of precision, recall and $F1$ will be equal.

Word	RE	SB	RND	|RE−SB|	|RE−RND|
argument	50.44	68.79	77.15	**18.35**	26.71
arm	92.30	85.69	88.11	6.61	**4.19**
atmosphere	70.52	69.15	80.44	**1.37**	10.32
audience	81.28	73.74	83.76	7.54	**4.22**
bank	85.76	83.07	82.46	**2.69**	3.99
degree	78.42	81.58	80.59	**3.16**	4.35
difference	62.46	61.43	75.17	**1.03**	12.90
difficulty	52.72	51.82	67.23	**0.90**	14.97
disc	78.62	76.48	78.07	**2.14**	6.18
image	71.78	75.76	81.50	**3.98**	10.02
interest	77.34	73.19	71.70	**4.15**	6.85
judgment	55.64	66.87	59.64	11.23	**9.01**
organization	80.36	72.86	78.65	7.50	**3.65**
paper	60.84	66.29	73.14	**5.45**	12.59
party	82.94	80.00	81.04	**2.94**	3.74
performance	58.56	64.76	73.86	**6.20**	15.52
plan	88.42	85.41	87.39	**3.01**	3.12
shelter	58.48	74.75	80.21	**16.27**	21.73
sort	67.64	88.15	77.37	20.51	**9.73**
source	63.46	67.74	66.26	**4.28**	7.03
overall	73.26	75.14	78.80	**129.31**	196.35

Table 4: Recall percentage of IMS on the 20 nouns of the Senseval-3 lexical-sample test set (RE) compared to the corresponding similarity-based (SB) and random (RND) pseudowords. The last 2 columns show absolute differences between the real and the two pseudoword settings.

3.2 Representative Power of Pseudosenses

The ideal case for pseudosenses would be that of being in a synonymous relationship with the corresponding real sense, i.e., selected from the same WordNet synset. But given that many of the WordNet synsets do not contain monosemous terms, the similarity-based approach often needs to look further into other related synsets to find a suitable pseudosense. To get a clear idea of the exact statistics, we went through all our similarity-based pseudowords and, for each pseudosense w_i, checked the relationship in WordNet between the synset containing w_i and the corresponding real sense. Table 5 shows for three values of *minFreq* the distribution of pseudosenses across different types of WordNet relationships, also including indirect ones. As can be seen in the Table, when *minFreq* is set to 0, a large portion of pseudosenses (around 75%) are selected from synonyms or generalization/specialization relations

	minFreq	0	500	1000
Relation type	Synonyms	33.0	7.6	5.4
	Hypernyms	33.4	16.1	13.0
	Hyponyms	9.1	6.1	4.9
	Meronyms	0.2	0.2	0.2
	Siblings	8.2	17.2	16.6
	Indirect relations	16.1	52.8	59.9

Table 5: Percentage of similarity-based pseudosenses obtained from different types of WordNet relations.

(hypernym and hyponyms). However, this percentage drops to about 23% when *minFreq* = 1000. This suggests that many of our pseudosenses are modeled from indirect relations when higher values of *minFreq* are used. This can potentially increase the risk of an undesirable modeling in which meanings are not properly preserved. For this reason, we carried out another experiment to assess the representative power of similarity-based pseudosenses. To this end, we randomly sampled 110 pseudowords (from the entire set of 15,935 pseudowords generated with minimum frequency of 1000), 10 for each degree of polysemy, from 2 to 12, totaling 770 pseudosenses. Then we presented each of these pseudowords[3] to two annotators who were asked to judge the degree of representativeness of its pseudosenses based on the following scores: 1: completely unrelated, 2: somewhat related, 3: good substitute, or 4: perfect substitute.

As an example, the scores assigned by the two annotators to different pseudosenses of the pseudoword generated for the noun *representative* are shown in Table 6. The overall representativeness score for each pseudoword is calculated by averaging the scores assigned to its individual pseudosenses. For instance, the overall scores calculated for the pseudoword *representative* are 3.75 and 3.50 (as given by the two annotators). The first row in Table 7 shows the average representativeness scores for each degree of polysemy on the full set of 770 pseudosenses. It can be seen that the score remains around 3.0 for all polysemy degrees from 2 to 12. Despite the fact that only one fifth of pseudosenses are taken from synonyms, hypernyms and hyponyms (when *minFreq* is 1000, cf. Table 5), the overall

[3]For each pseudoword, we provided annotators with the corresponding real word, as well as its synsets and glosses as given by WordNet.

Sense Definition (in short) {**Synset**} > **Corresponding Pseudosense**	**Score 1**	**Score 2**
a person who represents others {representative} > negotiator	3	3
an advocate who represents someone else's policy {spokesperson, interpreter, representative, voice} > spokesperson	4	4
a member of the U.S. House of Representatives {congressman, congresswoman, representative} > congressman	4	4
an item of information that is typical of a group {example, illustration, instance, representative} > case_in_point	4	3
average score	3.75	3.50

Table 6: Examples of representativeness scores assigned by the annotators to pseudosenses of the term *representative*.

representativeness score of 3.12 shows that most of these pseudosenses can be considered as good substitutes for their corresponding real senses. Therefore we conclude that not only does our similarity-based pseudoword generation approach extend the coverage of the vicinity-based method from 25% to 100% (when *minFreq* = 1000), but also that the pseudosenses coming from more distant synsets as ranked by PPR are still good representatives on average.

3.3 Distinguishability of Pseudosenses

In addition to assessing the representativeness of pseudosenses, their degree of distinguishability has to be determined. In other words, we have to determine how easily each pseudosense can be distinguished from the others in a pseudoword. Our reason for having such an experiment is readily illustrated by way of an example: consider the similarity-based pseudoword *philanthropist*benefactor*[4] corresponding to the noun *donor*[5]. Even though both pseudosenses are good representatives for their corresponding senses, the distinguishability of the two

[4]From WordNet: "Philanthropist: someone who makes charitable donations intended to increase human well-being"; "Benefactor: a person who helps people or institutions (especially with financial help)".

[5]*donor* has 2 senses according to WordNet 3.0: (1) "person who makes a gift of property"; (2) "(medicine) someone who gives blood or tissue or an organ to be used in another person".

Polysemy	2	3	4	5	6	7	8	9	10	11	12	Overall
Representativeness score	3.3	3.4	3.1	3.1	2.9	3.1	2.9	2.8	3.3	3.1	3.3	3.12
Distinguishability score	0.90	0.83	0.83	0.82	0.81	0.77	0.75	0.73	0.80	0.71	0.70	0.79

Table 7: Average representativeness and distinguishability scores for pseudosenses of different polysemy classes (scores range from 1 to 4 for representativeness and from 0 to 1 for distinguishability evaluation).

real senses is not preserved in the pseudoword. For instance, *benefactor* is a suitable pseudosense for both senses of *donor*, whereas *philanthropist* cannot be used in the blood donation sense.

Therefore we carried out another manual evaluation to test the efficacy of pseudowords in preserving the distinguishability of senses of real words. To this end, for each pseudoword P_w (from the same set of 110 sampled pseudowords used in Section 3.2) we presented its corresponding pseudosenses in random order to two annotators and asked them to associate each pseudosense with the most appropriate WordNet sense of the real word w. Then we calculated a distinguishability score for each polysemy degree by dividing the number of correct mappings by the total number of senses.

For instance, for the similarity-based pseudoword corresponding to the word *representative* (shown in Table 6), we provided the shuffled list of pseudosenses [*spokesperson, case_in_point, negotiator, congressman*] to each annotator and asked them to sort the list according to the WordNet sense inventory of *representative* (i.e., map each pseudosense to its most suitable real sense). Both annotators correctly mapped all pseudosenses of this pseudoword; hence, the distinguishability score given by each annotator for this pseudoword was 4/4 = 1.

The average distinguishability scores for each degree of polysemy, as well as the overall score, is shown in Table 7 (second row). Each value is an average of the scores obtained from the two annotators. It can be seen that the distinguishability score decreases for higher degrees of polysemy. The score, however, remains above 0.70 with highly-polysemous pseudowords. The overall score of 0.79 shows that similarity-based pseudowords effectively preserve the distinguishability of senses of their real counterparts. In other words, they do not tend to have over-generalized pseudosenses which cover more than one sense.

4 Related Work

The idea of pseudowords dates back to 1992, when it was first proposed as a means of generating large amounts of artificially annotated evaluation data for WSD algorithms (Gale et al., 1992; Schütze, 1992). However, as mentioned earlier in Section 2, constructing a pseudoword by combining a random set of unambiguous words, as was done in these early works, can not model systematic polysemy (Gaustad, 2001; Nakov and Hearst, 2003), since different senses of a real ambiguous word, unless it is homonymous, share some semantic or pragmatic relation.

Several researchers addressed the issue of producing semantically-aware pseudowords that can model semantic relationships between senses. Nakov and Hearst (2003) used lexical category membership from a medical term hierarchy (extracted from MeSH[6] (Medical Subject Headings)) to create "more plausibly-motivated" pseudowords. By considering the frequency distributions from lexical category co-occurrence, they produced a set of pseudowords which were closer to real ambiguous words in terms of disambiguation difficulty than random pseudowords. However, this approach requires a specific hierarchical lexicon and falls short of creating many pseudowords with high polysemy (the authors report generating pseudowords with two senses only).

More recent work has focused on the identification of monosemous representatives in the surrounding of a sense, i.e., selected among concepts directly related to the given sense. Lu et al. (2006) modeled senses of a real ambiguous word by picking out the most similar monosemous morpheme from a Chinese hierarchical lexicon. Pseudowords are then constructed by conflating these morphemes accordingly. However, this method leverages a specific Chinese hierarchical lexicon, in which different lev-

[6] http://www.nlm.nih.gov/mesh

els of the hierarchy correspond to different levels of sense granularity. A more flexible technique is proposed by Otrusina and Smrz (2010) who model ambiguous words in WordNet. Their vicinity-based approach searches the surroundings of each particular sense in the WordNet graph in order to find an unambiguous representative for that sense. However, as we described in Section 2.1.1, while the approach addresses the semantic awareness issue, it falls short of providing a high coverage, an issue which we tackle in our novel similarity-based approach.

5 Conclusion and Future Work

In this paper we proposed a new technique for the generation of pseudowords which, in contrast to existing work, can simultaneously tackle the two major issues associated with pseudowords, i.e., semantic awareness and coverage. Our approach can be used to model any given ambiguous noun in WordNet, hence enabling the generation of large-scale pseudosense-annotated datasets for thousands of pseudowords. We performed three experiments to evaluate the reliability of our pseudowords. We showed that the similarity-based pseudowords are highly correlated with their real counterparts in terms of disambiguation difficulty. Further evaluations demonstrated that this approach is able to provide a good semantic modeling of individual senses of real words while preserving their distinguishability.

We are releasing to the research community the entire set of 15,935 pseudowords, i.e., for all WordNet polysemous nouns (`http://lcl.uniroma1.it/pseudowords/`). This set of pseudowords (together with the English Gigaword corpus) can be used to generate a large pseudosense-tagged dataset containing $\geq$1000 annotated sentences for every sense of all the pseudowords modeled after real ambiguous nouns in WordNet. The resulting dataset could be a good complement for MASC (Ide et al., 2010) which, being human-created, can provide 1000 sense-annotated sentences for just a few words.

We hope that the availability of this resource will enable large-scale experiments in tasks such as semantic role labeling, semantic parsing, and Word Sense Disambiguation. Specifically, as future work, we plan to utilize the generated pseudosense-tagged dataset to perform an in-depth study of different WSD paradigms. We also plan to extend our work to other part-of-speech tags.

Acknowledgments

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

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Athens, Greece.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Boulder, Colorado.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 26–33, Toulouse, France.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 59–68, Honolulu, Hawaii.

Stefan Bordag. 2006. Word Sense Induction: Triplet-based clustering and automatic evaluation. In *Proceedings of the* 11^{th} *Conference on European chapter of the Association for Computational Linguistics*, EACL '06, pages 137–144, Trento, Italy.

Nathanael Chambers and Dan Jurafsky. 2010. Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 445–453, Uppsala, Sweden.

Antonio Di Marco and Roberto Navigli. 2013. Clustering and diversifying Web search results with graph-based Word Sense Induction. *Computational Linguistics*, 39(3).

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th*

Annual Meeting of the Association of Computational Linguistics, ACL '07, pages 216–223, Prague, Czech Republic.

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

William Gale, Kenneth Church, and David Yarowsky. 1992. Work on statistical methods for Word Sense Disambiguation. In *Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 54–60, Menlo Park, CA.

Tanja Gaustad. 2001. Statistical corpus-based Word Sense Disambiguation: Pseudowords vs real ambiguous words. In *Companion Volume to the Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics Proceedings ot the Student Research Workshop*, ACL/EACL '01, pages 61–66, Toulouse, France.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

David Graff and Christopher Cieri. 2003. English Gigaword, LDC2003T05. In *Linguistic Data Consortium*, Philadelphia.

Taher H. Haveliwala. 2002. Topic-sensitive PageRank. In *Proceedings of* 11^{th} *International Conference on World Wide Web*, WWW '02, pages 517–526, Honolulu, Hawaii, USA.

Thad Hughes and Daniel Ramage. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 581–589, Prague, Czech Republic.

Nancy Ide, Collin F. Baker, Christiane Fellbaum, and Rebecca J. Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 68–73, Uppsala, Sweden.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, Madison, USA.

Zhimao Lu, Haifeng Wang, Jianmin Yao, Ting Liu, and Sheng Li. 2006. An equivalent pseudoword solution to Chinese Word Sense Disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL '06, pages 457–464, Sydney, Australia.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain.

Preslav I. Nakov and Marti A. Hearst. 2003. Category-based pseudowords. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics – short papers*, HLT-NAACL '03, pages 67–69, Edmonton, Canada.

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 '12, pages 115–129, Spindleruv Mlyn, Czech Republic.

Lubomir Otrusina and Pavel Smrz. 2010. A new approach to pseudoword generation. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC'10, pages 1195–1199, Valletta, Malta.

Hinrich Schütze. 1992. Dimensions of meaning. In *Supercomputing '92: Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, pages 787–796, Minneapolis, Minnesota, USA.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 266–271, Princeton, New Jersey.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage Word Sense Disambiguation system for free text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL'10*, pages 78–83, Uppsala, Sweden.