Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity

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

Abstract

Semantic similarity is an essential component of many Natural Language Processing applications. However, prior methods for computing semantic similarity often operate at different levels, e.g., single words or entire documents, which requires adapting the method for each data type. We present a unified approach to semantic similarity that operates at multiple levels, all the way from comparing word senses to comparing text documents. Our method leverages a common probabilistic representation over word senses in order to compare different types of linguistic data. This unified representation shows state-of-the-art performance on three tasks: semantic textual similarity, word similarity, and word sense coarsening.

1 Introduction

Semantic similarity is a core technique for many topics in Natural Language Processing such as Textual Entailment (Berant et al., 2012), Semantic Role Labeling (Fürstenau and Lapata, 2012), and Question Answering (Surdeanu et al., 2011). For example, textual similarity enables relevant documents to be identified for information retrieval (Hliaoutakis et al., 2006), while identifying similar words enables tasks such as paraphrasing (Glickman and Dagan, 2003), lexical substitution (McCarthy and Navigli, 2009), lexical simplification (Biran et al., 2011), and Web search result clustering (Di Marco and Navigli, 2013).

Approaches to semantic similarity have often operated at separate levels: methods for word similarity are rarely applied to documents or even single sentences (Budanitsky and Hirst, 2006; Radinsky et al., 2011; Halawi et al., 2012), while document-based similarity methods require more

linguistic features, which often makes them inapplicable at the word or microtext level (Salton et al., 1975; Maguitman et al., 2005; Elsayed et al., 2008; Turney and Pantel, 2010). Despite the potential advantages, few approaches to semantic similarity operate at the sense level due to the challenge in sense-tagging text (Navigli, 2009); for example, none of the top four systems in the recent SemEval-2012 task on textual similarity compared semantic representations that incorporated sense information (Agirre et al., 2012).

We propose a unified approach to semantic similarity across multiple representation levels from senses to documents, which offers two significant advantages. First, the method is applicable independently of the input type, which enables meaningful similarity comparisons across different scales of text or lexical levels. Second, by operating at the sense level, a unified approach is able to identify the semantic similarities that exist independently of the text's lexical forms and any semantic ambiguity therein. For example, consider the sentences:

t1. *A manager fired the worker.*
t2. *An employee was terminated from work by his boss.*

A surface-based approach would label the sentences as dissimilar due to the minimal lexical overlap. However, a sense-based representation enables detection of the similarity between the meanings of the words, e.g., *fire* and *terminate*. Indeed, an accurate, sense-based representation is essential for cases where different words are used to convey the same meaning.

The contributions of this paper are threefold. First, we propose a new unified representation of the meaning of an arbitrarily-sized piece of text, referred to as a **lexical item**, using a sense-based probability distribution. Second, we propose a novel alignment-based method for word sense dis-

ambiguation during semantic comparison. Third, we demonstrate that this single representation can achieve state-of-the-art performance on three similarity tasks, each operating at a different lexical level: (1) surpassing the highest scores on the SemEval-2012 task on textual similarity (Agirre et al., 2012) that compares sentences, (2) achieving a near-perfect performance on the TOEFL synonym selection task proposed by Landauer and Dumais (1997), which measures word pair similarity, and also obtaining state-of-the-art performance in terms of the correlation with human judgments on the RG-65 dataset (Rubenstein and Goodenough, 1965), and finally (3) surpassing the performance of Snow et al. (2007) in a sense-coarsening task that measures sense similarity.

2 A Unified Semantic Representation

We propose a representation of any lexical item as a distribution over a set of word senses, referred to as the item's **semantic signature**. We begin with a formal description of the representation at the sense level (Section 2.1). Following this, we describe our alignment-based disambiguation algorithm which enables us to produce sense-based semantic signatures for those lexical items (e.g., words or sentences) which are not sense annotated (Section 2.2). Finally, we propose three methods for comparing these signatures (Section 2.3). As our sense inventory, we use WordNet 3.0 (Fellbaum, 1998).

2.1 Semantic Signatures

The WordNet ontology provides a rich network structure of semantic relatedness, connecting senses directly with their hypernyms, and providing information on semantically similar senses by virtue of their nearby locality in the network. Given a particular node (sense) in the network, repeated random walks beginning at that node will produce a frequency distribution over the nodes in the graph visited during the walk. To extend beyond a single sense, the random walk may be initialized and restarted from a set of senses (seed nodes), rather than just one; this multi-seed walk produces a multinomial distribution over all the senses in WordNet with higher probability assigned to senses that are frequently visited from the seeds. Prior work has demonstrated that multinomials generated from random walks over WordNet can be successfully applied to linguistic tasks such as word similarity (Hughes and Ramage,

2007; Agirre et al., 2009), paraphrase recognition, textual entailment (Ramage et al., 2009), and pseudoword generation (Pilehvar and Navigli, 2013).

Formally, we define the semantic signature of a lexical item as the multinomial distribution generated from the random walks over WordNet 3.0 where the set of seed nodes is the set of senses present in the item. This representation encompasses both when the item is itself a single sense and when the item is a sense-tagged sentence.

To construct each semantic signature, we use the iterative method for calculating topic-sensitive PageRank (Haveliwala, 2002). Let M be the adjacency matrix for the WordNet network, where edges connect senses according to the relations defined in WordNet (e.g., hypernymy and meronymy). We further enrich M by connecting a sense with all the other senses that appear in its disambiguated gloss.[1] Let $\vec{v}^{(0)}$ denote the probability distribution for the starting location of the random walker in the network. Given the set of senses S in a lexical item, the probability mass of $\vec{v}^{(0)}$ is uniformly distributed across the senses $s_i \in S$, with the mass for all $s_j \notin S$ set to zero. The PageRank may then be computed using:

$$\vec{v}^{(t)} = (1 - \alpha)\, M\vec{v}^{(t-1)} + \alpha\, \vec{v}^{(0)} \qquad (1)$$

where at each iteration the random walker may jump to any node $s_i \in S$ with probability $\alpha/|S|$. We follow standard convention and set α to 0.15. We repeat the operation in Eq. 1 for 30 iterations, which is sufficient for the distribution to converge. The resulting probability vector $\vec{v}^{(t)}$ is the semantic signature of the lexical item, as it has aggregated its senses' similarities over the entire graph. For our semantic signatures we used the UKB[2] off-the-shelf implementation of topic-sensitive PageRank.

2.2 Alignment-Based Disambiguation

Commonly, semantic comparisons are between word pairs or sentence pairs that do not have their lexical content sense-annotated, despite the potential utility of sense annotation in making semantic comparisons. However, traditional forms of word sense disambiguation are difficult for short texts and single words because little or no contextual information is present to perform the disambiguation task. Therefore, we propose a novel
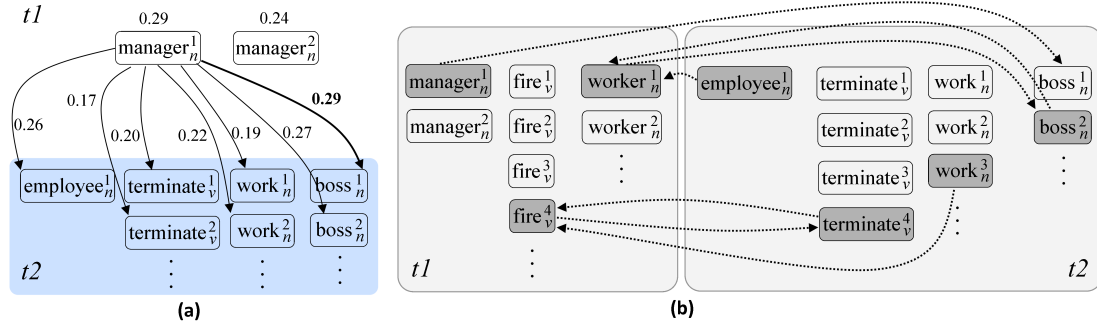
---

[1] http://wordnet.princeton.edu
[2] http://ixa2.si.ehu.es/ukb/

Figure 1: (a) Example alignments of the first sense of term *manager* (in sentence *t1*) to the two first senses of the word types in sentence *t2*, along with the similarity of the two senses' semantic signatures; (b) Alignments which maximize the similarities across words in *t1* and *t2* (the source side of an alignment is taken as the disambiguated sense of its corresponding word).

alignment-based sense disambiguation that leverages the content of the paired item in order to disambiguate each element. Leveraging the paired item enables our approach to disambiguate where traditional sense disambiguation methods can not due to insufficient context.

We view sense disambiguation as an **alignment problem**. Given two arbitrarily ordered texts, we seek the semantic alignment that maximizes the similarity of the senses of the context words in both texts. To find this maximum we use an alignment procedure which, for each word type w_i in item T_1, assigns w_i to the sense that has the maximal similarity to any sense of the word types in the compared text T_2. Algorithm 1 formalizes the alignment process, which produces a sense disambiguated representation as a result. Senses are compared in terms of their semantic signatures, denoted as function $\mathcal{R}$. We consider multiple definitions of $\mathcal{R}$, defined later in Section 2.3.

As a part of the disambiguation procedure, we leverage the one sense per discourse heuristic of Yarowsky (1995); given all the word types in two compared lexical items, each type is assigned a single sense, even if it is used multiple times. Additionally, if the same word type appears in both sentences, both will always be mapped to the same sense. Although such a sense assignment is potentially incorrect, assigning both types to the same sense results in a representation that does no worse than a surface-level comparison.

We illustrate the alignment-based disambiguation procedure using the two example sentences *t1* and *t2* given in Section 1. Figure 1(a) illustrates example alignments of the first sense of *manager* to the first two senses of the word types in sentence *t2* along with the similarity of the two senses'

Algorithm 1 Alignment-based Sense Disambiguation

Input: T_1 and T_2, the sets of word types being compared
Output: P, the set of disambiguated senses for T_1
1: $P \leftarrow \emptyset$
2: **for each** token $t_i \in T_1$
3:     $max\_sim \leftarrow 0$
4:     $best\_s_i \leftarrow null$
5:     **for each** token $t_j \in T_2$
6:         **for each** $s_i \in \text{Senses}(t_i), s_j \in \text{Senses}(t_j)$
7:             $sim \leftarrow \mathcal{R}(s_i, s_j)$
8:             **if** $sim > max\_sim$ **then**
9:                 $max\_sim = sim$
10:                 $best\_s_i = s_i$
11:     $P \leftarrow P \cup \{best\_s_i\}$
12: **return** P

---

semantic signatures. For the senses of manager, sense $manager_n^1$ obtains the maximal similarity value to $boss_n^1$ among all the possible pairings of the senses for the word types in sentence *t2*, and as a result is selected as the sense labeling for *manager* in sentence *t1*.[3] Figure 1(b) shows the final, maximally-similar sense alignment of the word types in *t1* and *t2*. The resulting alignment produces the following sets of senses:

$$P_{t1} = \{manager_n^1, fire_v^4, worker_n^1\}$$
$$P_{t2} = \{employee_n^1, terminate_v^4, work_n^3, boss_n^2\}$$

where P_x denotes the corresponding set of senses of sentence x.

2.3 Semantic Signature Similarity

Cosine Similarity. In order to compare semantic signatures, we adopt the *Cosine* similarity measure as a baseline method. The measure is computed by treating each multinomial as a vector and then calculating the normalized dot product of the two signatures' vectors.

---

[3] We follow Navigli (2009) and denote with w_p^i the i-th sense of w in WordNet with part of speech p.

1343

However, a semantic signature is, in essence, a weighted ranking of the importance of Word-Net senses for each lexical item. Given that the WordNet graph has a non-uniform structure, and also given that different lexical items may be of different sizes, the magnitudes of the probabilities obtained may differ significantly between the two multinomial distributions. Therefore, for computing the similarity of two signatures, we also consider two nonparametric methods that use the ranking of the senses, rather than their probability values, in the multinomial.

Weighted Overlap. Our first measure provides a nonparametric similarity by comparing the similarity of the rankings for intersection of the senses in both semantic signatures. However, we additionally weight the similarity such that differences in the highest ranks are penalized more than differences in lower ranks. We refer to this measure as the *Weighted Overlap*. Let S denote the intersection of all senses with non-zero probability in both signatures and r_i^j denote the rank of sense $s_i \in S$ in signature j, where rank 1 denotes the highest rank. The sum of the two ranks r_i^1 and r_i^2 for a sense is then inverted, which (1) weights higher ranks more and (2) when summed, provides the maximal value when a sense has the same rank in both signatures. The unnormalized weighted overlap is then calculated as $\sum_{i=1}^{|S|} (r_i^1 + r_i^2)^{-1}$. Then, to bound the similarity value in $[0, 1]$, we normalize the sum by its maximum value, $\sum_{i=1}^{|S|} (2i)^{-1}$, which occurs when each sense has the same rank in both signatures.

Top-k Jaccard. Our second measure uses the ranking to identify the top-k senses in a signature, which are treated as the best representatives of the conceptual associates. We hypothesize that a specific rank ordering may be attributed to small differences in the multinomial probabilities, which can lower rank-based similarities when one of the compared orderings is perturbed due to slightly different probability values. Therefore, we consider the top-k senses as an unordered set, with equal importance in the signature. To compare two signatures, we compute the Jaccard Index of the two signatures' sets:

$$\mathcal{R}_{Jac}(U_k, V_k) = \frac{|U_k \cap V_k|}{|U_k \cup V_k|} \qquad (2)$$

where U_k denotes the set of k senses with the highest probability in the semantic signature U.

Dataset	MSRvid	MSRpar	SMTeuroparl	OnWN	SMTnews
Training	750	750	734	-	-
Test	750	750	459	750	399

Table 1: Statistics of the provided datasets for the SemEval-2012 Semantic Textual Similarity task.

3 Experiment 1: Textual Similarity

Measuring semantic similarity of textual items has applications in a wide variety of NLP tasks. As our benchmark, we selected the recent SemEval-2012 task on Semantic Textual Similarity (STS), which was concerned with measuring the semantic similarity of sentence pairs. The task received considerable interest by facilitating a meaningful comparison between approaches.

3.1 Experimental Setup

Data. We follow the experimental setup used in the STS task (Agirre et al., 2012), which provided five test sets, two of which had accompanying training data sets for tuning system performance. Each sentence pair in the datasets was given a score from 0 to 5 (low to high similarity) by human judges, with a high inter-annotator agreement of around 0.90 when measured using the Pearson correlation coefficient. Table 1 lists the number of sentence pairs in training and test portions of each dataset.

Comparison Systems. The top-ranking participating systems in the SemEval-2012 task were generally supervised systems utilizing a variety of lexical resources and similarity measurement techniques. We compare our results against the top three systems of the 88 submissions: TLsim and TLsyn, the two systems of Šarić et al. (2012), and the UKP2 system (Bär et al., 2012). UKP2 utilizes extensive resources among which are a Distributional Thesaurus computed on 10M dependency-parsed English sentences. In addition, the system utilizes techniques such as Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007) and makes use of resources such as Wiktionary and Wikipedia, a lexical substitution system based on supervised word sense disambiguation (Biemann, 2013), and a statistical machine translation system. The TLsim system uses the New York Times Annotated Corpus, Wikipedia, and Google Book Ngrams. The TLsyn system also uses Google Book Ngrams, as well as dependency parsing and named entity recognition.

Ranking			System	Overall			Dataset-specific				
ALL	ALLnrm	Mean		ALL	ALLnrm	Mean	Mpar	Mvid	SMTe	OnWN	SMTn
1	1	1	ADW	**0.866**	**0.871**	**0.711**	0.694	**0.887**	**0.555**	**0.706**	**0.604**
2	3	2	UKP2	0.824	0.858	0.677	0.683	0.873	0.528	0.664	0.493
3	4	6	TLsyn	0.814	0.857	0.660	0.698	0.862	0.361	0.704	0.468
4	2	3	TLsim	0.813	0.864	0.675	**0.734**	0.880	0.477	0.679	0.398

Table 2: Performance of our system (ADW) and the 3 top-ranking participating systems (out of 88) in the SemEval-2012 Semantic Textual Similarity task. Rightmost columns report the corresponding Pearson correlation r for individual datasets, i.e., MSRpar (Mpar), MSRvid (Mvid), SMTeuroparl (SMTe), OnWN (OnWN) and SMTnews (SMTn). We also provide scores according to the three official evaluation metrics (i.e., ALL, ALLnrm, and Mean). Rankings are also presented based on the three metrics.

System Configuration. Here we describe the configuration of our approach, which we have called Align, Disambiguate and Walk (ADW). The STS task uses human similarity judgments on an ordinal scale from 0 to 5. Therefore, in ADW we adopted a similar approach to generating similarity values to that adopted by other participating systems, whereby a supervised system is trained to combine multiple similarity judgments to produce a final rating consistent with the human annotators. We utilized the WEKA toolkit (Hall et al., 2009) to train a Gaussian Processes regression model for each of the three training sets (cf. Table 1). The features discussed hereafter were considered in our regression model.

Main features. We used the scores calculated using all three of our semantic signature comparison methods as individual features. Although the *Jaccard* comparison is parameterized, we avoided tuning and instead used four features for distinct values of k: 250, 500, 1000, and 2500.

String-based features. Additionally, because the texts often contain named entities which are not present in WordNet, we incorporated the similarity values produced by four string-based measures, which were used by other teams in the STS task: (1) *longest common substring* which takes into account the length of the longest overlapping contiguous sequence of characters (substring) across two strings (Gusfield, 1997), (2) *longest common subsequence* which, instead, finds the longest overlapping subsequence of two strings (Allison and Dix, 1986), (3) *Greedy String Tiling* which allows reordering in strings (Wise, 1993), and (4) the character/word n-gram similarity proposed by Barrón-Cedeño et al. (2010).

We followed Šarić et al. (2012) and used the models trained on the SMTeuroparl and MSRpar datasets for testing on the SMTnews and OnWN test sets, respectively.

3.2 STS Results

Three evaluation metrics are provided by the organizers of the SemEval-2012 STS task, all of which are based on Pearson correlation r of human judgments with system outputs: (1) the correlation value for the concatenation of all five datasets (ALL), (2) a correlation value obtained on a concatenation of the outputs, separately normalized by least square (ALLnrm), and (3) the weighted average of Pearson correlations across datasets (Mean). Table 2 shows the scores obtained by ADW for the three evaluation metrics, as well as the Pearson correlation values obtained on each of the five test sets (rightmost columns). We also show the results obtained by the three top-ranking participating systems (i.e., UKP2, TLsim, and TLsyn). The leftmost three columns show the system rankings according to the three metrics.

As can be seen from Table 2, our system (ADW) outperforms all the 88 participating systems according to all the evaluation metrics. Our system shows a statistically significant improvement on the SMTnews dataset, with an increase in the Pearson correlation of over 0.10. MSRpar (MPar) is the only dataset in which TLsim (Šarić et al., 2012) achieves a higher correlation with human judgments. Named entity features used by the TLsim system could be the reason for its better performance on the MSRpar dataset, which contains a large number of named entities.

3.3 Similarity Measure Analysis

To gain more insight into the impact of our alignment-based disambiguation approach, we carried out a 10-fold cross-validation on the three training datasets (cf. Table 1) using the systems described hereafter.

ADW-MF. To build this system, we utilized our main features only; i.e., we did not make use of additional string-based features.

DW. Similarly to ADW-MF, this system utilized the main features only. In DW, however, we replaced our alignment-based disambiguation phase with a random walk-based WSD system that disambiguated the sentences separately, without performing any alignment. As our WSD system, we used UKB, a state-of-the-art knowledge-based WSD system that is based on the same topic-sensitive PageRank algorithm used by our approach. UKB initializes the algorithm from all senses of the words in the context of a word to be disambiguated. It then picks the most relevant sense of the word according to the resulting probability vector. As the lexical knowledge base of UKB, we used the same semantic network as that utilized by our approach for calculating semantic signatures.

Table 3 lists the performance values of the two above-mentioned systems on the three training sets in terms of Pearson correlation. In addition, we present in the table correlation scores for four other similarity measures reported by Bär et al. (2012):

- Pairwise Word Similarity that comprises of a set of WordNet-based similarity measures proposed by Resnik (1995), Jiang and Conrath (1997), and Lin (1998b). The aggregation strategy proposed by Corley and Mihalcea (2005) has been utilized for extending these word-to-word similarity measures for calculating text-to-text similarities.

- Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007) where the high-dimensional vectors are obtained on WordNet, Wikipedia and Wiktionary.

- Distributional Thesaurus where a similarity score is computed similarly to that of Lin (1998a) using a distributional thesaurus obtained from a 10M dependency-parsed sentences of English newswire.

- Character n-grams which were also used as one of our additional features.

As can be seen from Table 3, our alignment-based disambiguation approach (ADW-MF) is better suited to the task than a conventional WSD approach (DW). Another interesting point is the high scores achieved by the Character n-grams

Similarity measure	Dataset		
	Mpar	Mvid	SMTe
DW	0.448	0.820	0.660
ADW-MF	0.485	**0.842**	**0.721**
Explicit Semantic Analysis	0.427	0.781	0.619
Pairwise Word Similarity	0.564	0.835	0.527
Distributional Thesaurus	0.494	0.481	0.365
Character n-grams	**0.658**	0.771	0.554

Table 3: Performance of our main-feature system with conventional WSD (DW) and with the alignment-based disambiguation approach (ADW-MF) vs. four other similarity measures, using 10-fold cross validation on the training datasets MSR-par (Mpar), MSRvid (Mvid), and SMTeuroparl (SMTe).

measure. This confirms that string-based methods are strong baselines for semantic textual similarity. Except for the MSRpar (Mpar) dataset, our system (ADW-MF) outperforms all other similarity measures. The scores obtained by Explicit Semantic Analysis and Distributional Thesaurus are not competitive on any dataset. On the other hand, Pairwise Word Similarity achieves a high performance on MSRpar and MSRvid datasets, but performs surprisingly low on the SMTeuroparl dataset.

4 Experiment 2: Word Similarity

We now proceed from the sentence level to the word level. Word similarity has been a key problem for lexical semantics, with significant efforts being made by approaches in distributional semantics to accurately identify synonymous words (Turney and Pantel, 2010). Different evaluation methods exist in the literature for evaluating the performance of a word-level semantic similarity measure; we adopted two well-established benchmarks: synonym recognition and correlating word similarity judgments with those from human annotators.

For synonym recognition, we used the TOEFL dataset created by Landauer and Dumais (1997). The dataset consists of 80 multiple-choice synonym questions from the TOEFL test; a word is paired with four options, one of which is a valid synonym. Test takers with English as a second language averaged 64.5% correct. Despite multiple approaches, only recently has the test been answered perfectly (Bullinaria and Levy, 2012), underscoring the challenge of synonym recognition.

Approach	Accuracy
PPMIC (Bullinaria and Levy, 2007)	85.00%
GLSA (Matveeva et al., 2005)	86.25%
LSA (Rapp, 2003)	92.50%
ADW_{Jac}	93.75±2.5%
ADW_{WO}	95.00%
ADW_{Cos}	96.25%
PR (Turney et al., 2003)	97.50%
PCCP (Bullinaria and Levy, 2012)	100.00%

Table 4: Accuracy on the 80-question TOEFL Synonym test. ADW_{Jac}, ADW_{WO}, and ADW_{Cos} correspond to results with the *Jaccard*, *Weighted Overlap* and *Cosine* signature comparison measures, respectively.

Word	Synonym choices (correct in bold)			
fanciful	familiar	apparent⋆	**imaginative**†	logical
verbal	**oral**†	overt	fitting	verbose⋆
resolved	**settled**⋆	forgotten†	publicized	examined
percentage	volume	sample	**proportion**	profit†⋆
figure	list	**solve**⋆	divide†	express
highlight	alter†	imitate	**accentuate**⋆	restore

Table 5: Questions answered incorrectly by our approach. Symbols † and ⋆ correspond to the choices of our approach with the *Weighted Overlap* and *Cosine* signature comparisons respectively. We do not include the mistakes made when the *Jaccard* measure was used as they vary with the k value.

For the similarity judgment evaluation, we used as benchmark the RG-65 dataset created by Rubenstein and Goodenough (1965). The dataset contains 65 word pairs judged by 51 human subjects on a scale of 0 to 4 according to their semantic similarity. Ideally, a measure's similarity judgments are expected to be highly correlated with those of humans. To be consistent with the previous literature (Hughes and Ramage, 2007; Agirre et al., 2009), we used Spearman's rank correlation in our experiment.

4.1 Experimental Setup

Our alignment-based sense disambiguation transforms the task of comparing individual words into that of calculating the similarity of the best-matching sense pair across the two words. As there is no training data we do not optimize the k value for computing signature similarity with the *Jaccard* index; instead, we report, for the synonym recognition and the similarity judgment evaluations, the respective range of accuracies and the average correlation obtained upon using five values of k randomly selected in the range $[50, 2500]$: $678, 1412, 1692, 2358, 2387$.

4.2 Word Similarity Results: TOEFL dataset

Table 4 lists the accuracy performance of the system in comparison to the existing state of the art on the TOEFL test. ADW_{WO}, ADW_{Cos}, and ADW_{Jac} correspond to our approach when *Weighted Overlap*, *Cosine*, and *Jaccard* signature comparison measures are used, respectively. Despite not being tuned for the task, our model achieves near-perfect performance, answering all but three questions correctly with the *Cosine* measure. Among the top-performing approaches, only

that of Rapp (2003) uses word senses, an approach that is outperformed by our method.

The errors produced by our system were largely the result of sense locality in the WordNet network. Table 5 highlights the incorrect responses. The synonym mistakes reveal cases where senses of the two words are close in WordNet, indicating some relatedness. For example, *percentage* may be interpreted colloquially as monetary value (e.g., "give me my percentage") and elicits the synonym of *profit* in the economic domain, which ADW incorrectly selects as a synonym.

4.3 Word Similarity Results: RG-65 dataset

Table 6 shows the Spearman's ρ rank correlation coefficients with human judgments on the RG-65 dataset. As can be seen from the Table, our approach with the *Weighted Overlap* signature comparison improves over the similar approach of Hughes and Ramage (2007) which, however, does not involve the disambiguation step and considers a word as a whole unit as represented by the set of its senses.

5 Experiment 3: Sense Similarity

WordNet is known to be a fine-grained sense inventory with many related word senses (Palmer et al., 2007). Accordingly, multiple approaches have attempted to identify highly similar senses in order to produce a coarse-grained sense inventory. We adopt this task as a way of evaluating our similarity measure at the sense level.

5.1 Coarse-graining Background

Earlier work on reducing the polysemy of sense inventories has considered WordNet-based sense relatedness measures (Mihalcea and Moldovan, 2001) and corpus-based vector representations of

Approach	Correlation
ADW_{Cos}	0.825
Agirre et al. (2009)	0.830
Hughes and Ramage (2007)	0.838
Zesch et al. (2008)	0.840
ADW_{Jac}	0.841
ADW_{WO}	0.868

Table 6: Spearman's ρ correlation coefficients with human judgments on the RG-65 dataset. ADW_{Jac}, ADW_{WO}, and ADW_{Cos} correspond to results with the *Jaccard*, *Weighted Overlap* and *Cosine* signature comparison measures respectively.

Method	Onto		SE-2			Onto + SE-2	
	Noun	Verb	Noun	Verb	Adj	Noun	Verb
$\mathcal{R}_{Cos}$	0.406	0.522	0.450	0.465	0.484	0.441	0.485
$\mathcal{R}_{WO}$	**0.421**	**0.544**	**0.483**	**0.482**	**0.531**	**0.470**	**0.503**
$\mathcal{R}_{Jac}$	0.418	0.531	0.478	0.473	0.501	0.465	0.493
SVM	0.370	0.455	NA	NA	0.473	0.423	0.432
ODE	0.218	0.396	NA	NA	0.371	0.331	0.288

Table 7: F-score sense merging evaluation on three hand-labeled datasets: OntoNotes (Onto), Senseval-2 (SE-2), and combined (Onto+SE-2). Results are reported for all three of our signature comparison measures and also for two previous works (last two rows).

word senses (Agirre and Lopez, 2003; McCarthy, 2006). Navigli (2006) proposed an automatic approach for mapping WordNet senses to the coarse-grained sense distinctions of the Oxford Dictionary of English (ODE). The approach leverages semantic similarities in gloss definitions and the hierarchical relations between senses in the ODE to cluster WordNet senses. As current state of the art, Snow et al. (2007) developed a supervised SVM classifier that utilized, as its features, several earlier sense relatedness techniques such as those implemented in the WordNet::Similarity package (Pedersen et al., 2004). The classifier also made use of resources such as topic signatures data (Agirre and de Lacalle, 2004), the WordNet domain dataset (Magnini and Cavaglià, 2000), and the mappings of WordNet senses to ODE senses produced by Navigli (2006).

5.2 Experimental Setup

We benchmark the accuracy of our similarity measure in grouping word senses against those of Navigli (2006) and Snow et al. (2007) on two datasets of manually-labeled sense groupings of WordNet senses: (1) sense groupings provided as a part of the Senseval-2 English Lexical Sample WSD task (Kilgarriff, 2001) which includes nouns, verbs and adjectives; (2) sense groupings included in the OntoNotes project[4] (Hovy et al., 2006) for nouns and verbs. Following the evaluation methodology of Snow et al. (2007), we combine the Senseval-2 and OntoNotes datasets into a third dataset.

Snow et al. (2007) considered sense grouping as a binary classification task whereby for each word every possible pairing of senses has to be classified

---

[4]Sense groupings belong to a pre-version 1.0: `http://cemantix.org/download/sense/ontonotes-sense-groups.tar.gz`

as either *merged* or *not-merged*. We constructed a simple threshold-based classifier to perform the same binary classification. To this end, we calculated the semantic similarity of each sense pair and then used a threshold value t to classify the pair as merged if similarity $\geq t$ and not-merged otherwise. We sampled out 10% of the dataset for tuning the value of t, thus adapting our classifier to the fine granularity of the dataset. We used the same held-out instances to perform a tuning of the k value used for *Jaccard* index, over the same values of k as in Experiment 1 (cf. Section 3).

5.3 Sense Merging Results

For a binary classification task, we can directly calculate precision, recall and F-score by constructing a contingency table. We show in Table 7 the F-score performance of our classifier as obtained by an averaged 10-fold cross-validation. Results are presented for all three of the measures of semantic signature comparison and for the three datasets: OntoNotes, Senseval-2, and the two combined. In addition, we show in Table 7 the F-score results provided by Snow et al. (2007) for their SVM-based system and for the mapping-based approach of Navigli (2006), denoted by ODE.

Table 7 shows that our methodology yields improvements over previous work on both datasets and for all parts of speech, irrespective of the semantic signature comparison method used. Among the three methods, *Weighted Overlap* achieves the best performance, which demonstrates that our transformation of semantic signatures into ordered lists of concepts and calculating similarity by rank comparison has been helpful.

1348

6 Related Work

Due to the wide applicability of semantic similarity, significant efforts have been made at different lexical levels. Early work on document-level similarity was driven by information retrieval. Vector space methods provided initial successes (Salton et al., 1975), but often suffer from data sparsity when using small documents, or when documents use different word types, as in the case of paraphrases. Later efforts such as LSI (Deerwester et al., 1990), PLSA (Hofmann, 2001) and Topic Models (Blei et al., 2003; Steyvers and Griffiths, 2007) overcame these sparsity issues using dimensionality reduction techniques or modeling the document using latent variables. However, such methods were still most suitable for comparing longer texts. Complementary approaches have been developed specifically for comparing shorter texts, such as those used in the SemEval-2012 STS task (Agirre et al., 2012). Most similar to our approach are the methods of Islam and Inkpen (2008) and Corley and Mihalcea (2005), who performed a word-to-word similarity alignment; however, they did not operate at the sense level. Ramage et al. (2009) used a similar semantic representation of short texts from random walks on WordNet, which was applied to paraphrase recognition and textual entailment. However, unlike our approach, their method does not perform sense disambiguation prior to building the representation and therefore potentially suffers from ambiguity.

A significant amount of effort has also been put into measuring similarity at the word level, frequently by approaches that use distributional semantics (Turney and Pantel, 2010). These methods use contextual features to represent semantics at the word level, whereas our approach represents word semantics at the sense level. Most similar to our approach are those of Agirre et al. (2009) and Hughes and Ramage (2007), which represent word meaning as the multinomials produced from random walks on the WordNet graph. However, unlike our approach, neither of these disambiguates the two words being compared, which potentially conflates the meanings and lowers the similarity judgment.

Measures of sense relatedness have frequently leveraged the structural properties of WordNet (e.g., path lengths) to compare senses. Budanitsky and Hirst (2006) provided a survey of such WordNet-based measures. The main drawback with these approaches lies in the WordNet structure itself, where frequently two semantically similar senses are distant in the WordNet hierarchy. Possible solutions include relying on wider-coverage networks such as WikiNet (Nastase and Strube, 2013) or multilingual ones such as BabelNet (Navigli and Ponzetto, 2012b). Fewer works have focused on measuring the similarity – as opposed to relatedness – between senses. The topic signatures method of Agirre and Lopez (2003) represents each sense as a vector over corpus-derived features in order to build comparable sense representations. However, topic signatures often produce lower quality representations due to sparsity in the local structure of WordNet, especially for rare senses. In contrast, the random walk used in our approach provides a denser, and thus more comparable, representation for all WordNet senses.

7 Conclusions

This paper presents a unified approach for computing semantic similarity at multiple lexical levels, from word senses to texts. Our method leverages a common probabilistic representation at the sense level for all types of linguistic data. We demonstrate that our semantic representation achieves state-of-the-art performance in three experiments using semantic similarity at different lexical levels (i.e., sense, word, and text), surpassing the performance of previous similarity measures that are often specifically targeted for each level.

In future work, we plan to explore the impact of the sense inventory-based network used in our semantic signatures. Specifically, we plan to investigate higher coverage inventories such as BabelNet (Navigli and Ponzetto, 2012a), which will handle texts with named entities and rare senses that are not in WordNet, and will also enable cross-lingual semantic similarity. Second, we plan to evaluate our method on larger units of text and formalize comparison methods between different lexical levels.

Acknowledgments

References

Eneko Agirre and Oier Lopez de Lacalle. 2004. Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of LREC*, pages 1123–1126, Lisbon, Portugal.

Eneko Agirre and Oier Lopez. 2003. Clustering WordNet word senses. In *Proceedings of RANLP*, pages 121–130, Borovets, Bulgaria.

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 NAACL*, pages 19–27, Boulder, Colorado.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of SemEval-2012*, pages 385–393, Montreal, Canada.

Lloyd Allison and Trevor I. Dix. 1986. A bit-string longest-common-subsequence algorithm. *Information Processing Letters*, 23(6):305–310.

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. UKP: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of SemEval-2012*, pages 435–440, Montreal, Canada.

Alberto Barrón-Cedeño, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism detection across distant language pairs. In *Proceedings of COLING*, pages 37–45, Beijing, China.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2012. Learning entailment relations by global graph structure optimization. *Computational Linguistics*, 38(1):73–111.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of ACL*, pages 496–501, Portland, Oregon.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

John A. Bullinaria and Joseph. P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, (3):510.

John A. Bullinaria and Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, 44:890–907.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of American Society for Information Science*, 41(6):391–407.

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

Tamer Elsayed, Jimmy Lin, and Douglas W. Oard. 2008. Pairwise document similarity in large collections with MapReduce. In *Proceedings of ACL-HLT*, pages 265–268, Columbus, Ohio.

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

Hagen Fürstenau and Mirella Lapata. 2012. Semi-supervised Semantic Role Labeling via structural alignment. *Computational Linguistics*, 38(1):135–171.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 1606–1611, Hyderabad, India.

Oren Glickman and Ido Dagan. 2003. Acquiring lexical paraphrases from a single corpus. In *Proceedings of RANLP*, pages 81–90, Borovets, Bulgaria.

Dan Gusfield. 1997. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of KDD*, pages 1406–1414, Beijing, China.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Taher H. Haveliwala. 2002. Topic-sensitive PageRank. In *Proceedings of WWW*, pages 517–526, Hawaii, USA.

Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides GM Petrakis, and Evangelos Milios. 2006. Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems*, 2(3):55–73.

Thomas Hofmann. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1):177–196.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of NAACL*, pages 57–60, NY, USA.

Thad Hughes and Daniel Ramage. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of EMNLP-CoNLL*, pages 581–589, Prague, Czech Republic.

Aminul Islam and Diana Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2):10:1–10:25.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of ROCLING X*, pages 19–30, Taiwan.

Adam Kilgarriff. 2001. English lexical sample task description. In *Proceedings of Senseval*, pages 17–20, Toulouse, France.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review*, 104(2):211.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of COLING*, pages 768–774, Montreal, Quebec, Canada.

Dekang Lin. 1998b. An information-theoretic definition of similarity. In *Proceedings of ICML*, pages 296–304, San Francisco, CA.

Bernardo Magnini and Gabriela Cavaglià. 2000. Integrating subject field codes into WordNet. In *Proceedings of LREC*, pages 1413–1418, Athens, Greece.

Ana G. Maguitman, Filippo Menczer, Heather Roinestad, and Alessandro Vespignani. 2005. Algorithmic detection of semantic similarity. In *Proceedings of WWW*, pages 107–116, Chiba, Japan.

Irina Matveeva, Gina-Anne Levow, Ayman Farahat, and Christiaan Royer. 2005. Terms representation with generalized latent semantic analysis. In *Proceedings of RANLP*, Borovets, Bulgaria.

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

Diana McCarthy. 2006. Relating WordNet senses for word sense disambiguation. In *Proceedings of the Workshop on Making Sense of Sense at EACL-06*, pages 17–24, Trento, Italy.

Rada Mihalcea and Dan Moldovan. 2001. Automatic generation of a coarse grained WordNet. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, USA.

Vivi Nastase and Michael Strube. 2013. Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85.

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 AAAI*, pages 108–114, Toronto, Canada.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost Word Sense Disambiguation performance. In *Proceedings of COLING-ACL*, pages 105–112, Sydney, Australia.

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

Martha Palmer, Hoa Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity - measuring the relatedness of concepts. In *Proceedings of AAAI*, pages 144–152, San Jose, CA.

Mohammad Taher Pilehvar and Roberto Navigli. 2013. Paving the way to a large-scale pseudosense-annotated dataset. In *Proceedings of NAACL-HLT*, pages 1100–1109, Atlanta, USA.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of WWW*, pages 337–346, Hyderabad, India.

Daniel Ramage, Anna N. Rafferty, and Christopher D. Manning. 2009. Random walks for text semantic similarity. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 23–31, Suntec, Singapore.

Reinhard Rapp. 2003. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, pages 315–322, New Orleans, LA.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pages 448–453, Montreal, Canada.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Gerard Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In *EMNLP-CoNLL*, pages 1005–1014, Prague, Czech Republic.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to rank answers to non-factoid questions from Web collections. *Computational Linguistics*, 37(2):351–383.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of RANLP*, pages 482–489, Borovets, Bulgaria.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Takelab: Systems for measuring semantic text similarity. In *Proceedings of SemEval-2012*, pages 441–448, Montreal, Canada.

Michael J. Wise. 1993. String similarity via greedy string tiling and running Karp-Rabin matching. In *Department of Computer Science Technical Report*, Sydney.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation rivaling supervised methods. In *Proceedings of ACL*, pages 189–196, Cambridge, Massachusetts.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Using Wiktionary for computing semantic relatedness. In *Proceedings of AAAI*, pages 861–866, Chicago, Illinois.