View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Waseda University Repository

Automatic identification of words with novel but infrequent senses*

Paul Cook and Graeme Hirst

Department of Computer Science
University of Toronto
{pcook,gh}@cs.toronto.edu

Abstract. We propose a statistical method for identifying words that have a novel sense in one corpus compared to another based on differences in their lexico-syntactic contexts in those corpora. In contrast to previous work on identifying semantic change, we focus specifically on infrequent word senses. Given the challenges of evaluation for this task, we further propose a novel evaluation method based on synthetic examples of semantic change that allows us to simulate differing degrees of sense change. Our proposed method is able to identify rather subtle simulated sense changes, and outperforms both a random baseline and a previously-proposed approach.

Keywords: Neologisms, semantic change, word senses, lexical semantics

1 New word senses

The meanings of words are not static, but can vary and change in a number of ways. In particular, words can undergo diachronic change—change over time—and come to be used in new senses. Contemporary examples of sense change can be seen in the following usages of *rock*, *sick*, and *text*, all of which correspond to relatively new senses of these words.

1. *Marvez has rocked the mullet* [hair-style] *for years as a style statement.* [*rock* = 'display with pride']
2. *LeBron has one of the sickest vertical leaps in the game, yet how many alley-oops have you seen him convert in his career?* [*sickest* = 'best']
3. *After she ignored the first few texts and phone calls, I gave up.* [*text* = 'text message']

Furthermore, new word senses are not necessarily frequent. For example, the above usages are taken from the enTenTen corpus,[1] a very large corpus containing a wide variety of text types, but the corresponding senses of these words appear to be rare in this corpus.

The identification of new word senses is an important task in lexicography, and is necessary to keep dictionaries up-to-date. But lexical semantic change has only recently been studied from a computational perspective, and only to a limited extent (e.g., Sagi et al., 2009; Cook and Stevenson, 2010; Gulordava and Baroni, 2011). Furthermore, despite the low frequency of many novel word senses, no work to date has specifically considered the identification of words with novel infrequent senses. In contrast, in this paper we focus specifically on this issue

The manual identification of novel senses is becoming increasingly difficult nowadays due to the vast quantities of text being produced (e.g., through online social media) that must be searched.

* We thank Afsaneh Fazly, Diana McCarthy, Suzanne Stevenson, and the members of the University of Melbourne Language Technology Group for their feedback on earlier versions of this work. This work was financially supported by MITACS. Paul Cook is now a McKenzie Postdoctoral Fellow in the Department of Computer Science and Software Engineering at the University of Melbourne.

Copyright 2011 by Paul Cook and Graeme Hirst

[1] http://trac.sketchengine.co.uk/wiki/Corpora/enTenTen

The methods we propose could form the basis for a lexicographical tool to aid in the identification of new word senses that are particularly difficult to find due to their relatively low frequency.

Evaluating approaches to identifying diachronic semantic change is difficult; indeed, most previous studies have relied on rather small datasets (e.g., Sagi et al., 2009; Cook and Stevenson, 2010) or human judges' intuitions about changes in meaning, which might not accurately reflect changes in meaning observed in corpora (e.g., Gulordava and Baroni, 2011). Therefore, taking inspiration from evaluation approaches in word sense disambiguation that make use of artificially-ambiguous words (e.g., Schütze, 1992; Gale et al., 1992), we propose evaluation methods that use synthetic examples of semantic change. Crucially, this enables us to to carefully control for the frequency of senses; this allows us to assess how rare a word sense may be, and yet still be identified by our method.

One further consideration is that methods for identifying semantic change must be applicable to relatively small corpora. Although there is a move towards using ever-larger corpora in computational linguistics, many historical corpora and corpora for specific time periods are rather small. For example, one year of text from the New York Times Annotated Corpus (Sandhaus, 2008) consists of roughly 50 million words, which is small compared to contemporary corpora which often contain billions of words. In this study, we therefore focus on relatively small corpora.

We discuss some related work in Section 2, and then present our model for identifying words with differing senses in Section 3. In Sections 4 and 5 we empirically evaluate our model on synthetic examples of semantic change created from Senseval data and near synonyms. We then offer some concluding remarks in Section 6.

2 Related work

Sagi et al. (2009) and Cook and Stevenson (2010) focus on identifying specific types of diachronic change—widening and narrowing, and amelioration and pejoration, respectively—and exploit properties of these phenomena in their methods for identifying them. Gulordava and Baroni (2011) consider the identification of diachronic changes in meaning from an *n*-gram database, but in contrast to Sagi et al. and Cook and Stevenson, do not focus on specific types of semantic change. Others have studied differences in meaning between dialects and domains, instead of over time. Peirsman et al. (2010) consider the identification of lectal markers—words typical of one dialect versus another, either because of their marked frequency or sense—in Belgian and Netherlandic Dutch. McCarthy et al. (2007) consider the identification of predominant word senses in corpora, focusing on differences between domains. This method can be applied to not only identify the words that differ in predominant sense in two corpora, but also the specific predominant senses of those words. Nevertheless, none of these studies has specifically considered the identification of words with novel infrequent senses, the focus of this study.

Given the challenges of evaluating methods for identifying semantic change—namely a lack of suitable resources—we propose the use of synthetic examples of semantic change for evaluation. Gaustad (2001) showed that evaluations using pseudowords (artificially-ambiguous words) can over-estimate the accuracy of a word sense disambiguation system on real data. Gaustad suggests this is because word senses are typically related (i.e., words are polysemous) whereas pseudowords are usually created from words with distinct senses. Nakov and Hearst (2003) and Otrusina and Smrz (2010) propose methods for constructing more-realistic pseudowords by taking into account information about lexical categories, and lexical or distributional information, respectively. In this work we propose a new use for pseudowords—evaluating methods for identifying semantic change—and attempt to address concerns about the use of pseudowords, such as those raised by Gaustad, by creating our pseudowords (discussed in Sections 4 and 5) from real word senses and words with related senses.

3 Model

The input to our method is two corpora which represent different text varieties, e.g., different time periods. The output is a set of words—in this study, either nouns or verbs—that are hypothesized by the method to have different senses in one of the corpora compared to the other.

We consider a statistical model similar to one that Peirsman et al. (2010) used for automatically identifying lectal markers. This approach assumes that usages of different senses of a word will occur in different contexts, and that the aggregated contexts of a word in two corpora will differ if the senses of that word differ in those corpora. The model is based on a distributional representation of meaning that draws on work on automatically clustering similar words (Lin, 1998) that has been incorporated into tools used by lexicographers to identify word senses (Kilgarriff and Tugwell, 2002). Specifically, this method measures the similarity of two lexico-syntactic representations of the aggregated contexts of a target word; these two representations would typically come from different corpora representing, for example, different time periods. The lexico-syntactic representations capture the association of a target word with dependency triples, and the similarity between two target word representations is determined with a number of metrics. We propose some variations to Peirsman et al.'s model—specifically a novel association measure (Section 3.1) and similarity metric (Section 3.2)—that are found to improve its performance (discussed in Section 5).

3.1 Association measures

We use the following information-theoretic association measure proposed by Lin (1998):

$$I(w_1, r, w_2) = \log \left(\frac{||w_1, r, w_2|| \; ||*, r, *||}{||w_1, r, *|| \; ||*, r, w_2||} \right) \tag{1}$$

where w_1, r, and w_2 are a head, dependency relation, and dependent, respectively; $||\cdot||$ is the frequency of some tuple (e.g., w_1, r, w_2); and $*$ refers to any item (e.g., $w_1, r, *$ is a tuple with head w_1, relation r, and any dependent).

In some cases I is not an appropriate association measure. For small corpora, counts for many dependency triples will be very low and hence unreliable. To avoid these data sparseness problems we therefore consider a second, and much simpler, association measure. Joanis et al. (2008) found that the frequency of a verb occurring in specific syntactic relations, as well as the frequency with which that verb co-occurs with specific prepositions, are useful features in verb classification. Our conditional probability-based association measure (cprob), captures information similar to that used by Joanis et al., and is calculated as below:

$$\text{cprob}(w_1, r, w_2) = \begin{cases} \dfrac{||w_1, r, w_2||}{||w_1, *, *||} & \text{if } r = \text{preposition or particle} \\ \\ \dfrac{||w_1, r, *||}{||w_1, *, *||} & \text{otherwise} \end{cases} \tag{2}$$

Prepositions and particles both often indicate the meaning of a verb. Because these parts-of-speech are frequent, and other dependents are ignored, this association measure can be estimated from smaller corpora more accurately than I.

3.2 Similarity metrics

We refer to the usages of a given word w in a corpus C as w_C. We define the salient tuples T_C for a word w_C to be the set of all head, dependency relation, dependent tuples t in corpus C having w as head and frequency in C greater than a threshold, which we set to 5. Peirsman et al. (2010) find cosine to slightly outperform several other metrics—including the similarity metric proposed by

Lin (1998)—in a cross-varietal synonymy detection task, and use cosine in their experiments on identifying lectal markers; we therefore also consider cosine and define the similarity for usages of tuples in corpora A and B as follows:

$$\text{Cosine}(w_A, w_B) = \frac{\sum_{t \in T_A \cap T_B} A_A(t) * A_B(t)}{\sqrt{\sum_{t \in T_A} A_A(t)^2} * \sqrt{\sum_{t \in T_B} A_B(t)^2}} \tag{3}$$

where $A_A(t)$ and $A_B(t)$ correspond to the association—either I or cprob—computed for corpora A and B, respectively.

Cosine is a symmetrical similarity metric, but an asymmetrical metric may be more appropriate in some cases. For example, one method used by lexicographers to search for neologisms is to compare a corpus of recent texts to a reference corpus representing standard usage (O'Donovan and O'Neil, 2008). In this case, focusing on salient usages in the corpus of newer texts that are less salient, or unattested, in the reference corpus may be more appropriate for identifying novel senses. We therefore propose the following asymmetrical metric—Newness—in which R is considered to be a reference corpus, and N a corpus of newer texts:[2]

$$\text{Newness}(w_N, w_R) = \frac{\sum_{t \in T_N - T_R} A_N(t) + \sum_{t \in T_N \cap T_R} \max(A_N(t) - A_R(t), 0)}{\sum_{t \in T_N} A_N(t)} \tag{4}$$

The first part of the numerator ($\sum_{t \in T_N - T_R} A_N(t)$) focuses on tuples that are in N but unattested in R. The second part of the numerator ($\sum_{t \in T_N \cap T_R} \max(A_N(t) - A_R(t), 0)$) focuses on tuples that have stronger association in N than R; the max prevents tuples with stronger association in R than N from impacting the score. The denominator ensures the final score is in $[0, 1]$.

Cosine is a similarity metric—words that have similar usages in two corpora will receive high scores. However, Newness is a dissimilarity metric which assigns high scores to words that have novel usages in one corpus compared to a reference corpus. We use Cosine and Newness to produce a ranking of lemmas, and account for this difference between the measures by simply negating Cosine scores to reverse the rankings for Cosine.

3.3 Inter-corpus and intra-corpus similarity

To determine the extent to which the meanings of a word differ between two corpora, we could simply compute inter-corpus similarity for that word directly, using any of the above similarity metrics; e.g., to compute the difference in the meanings of *click* between corpora A and B using the Newness similarity metric, we could compute $\text{Newness}(click_A, click_B)$. However, as Peirsman et al. (2010) observe, it may also be important to take into account the extent to which the meanings of a word vary *within* a corpus. For example, suppose we observe that the computed difference in meaning is large for some word w between corpora A and B. Taking B as a reference corpus, suppose that we also observe that the computed difference in meaning for w between two random samples of usages from B alone is large. In this case we should not necessarily believe that w differs in meaning between A and B because, based on w's distribution in B, we expect its computed meaning to vary. However, if the computed difference in meaning of w between two random samples from B were small, then the observed computed difference for w between A and B might be more indicative of w having different senses in A and B.

We operationalize the above intuition as follows: we randomly 2-partition the documents of B—the reference corpus—10 times. For a given word w, this allows us to compute 10 intra-corpus B similarities. However, corpus size can influence association measures.[3] In order to make

[2] In this case, following Lin (1998), we additionally restrict the salient tuples to those having positive association.

[3] I is known to assign high association to low frequency items; a given item will tend to have lower frequency in a smaller corpus.

comparisons between similar-size corpora, we randomly 2-partition the documents of A once.[4] This allows us to compute a total of 40 inter-corpus similarities for w. We then compute the difference between w's average inter-corpus and average intra-corpus similarity.

4 Synthetic examples of semantic change from Senseval data

In this section we use Senseval data to create synthetic examples of words which have undergone semantic change, and evaluate our model on these items. Specifically, we simulate words taking on a novel sense. We do so by dividing the usages for a given word into two parts such that one sense—the "novel" sense—is present in only one of the parts. We can further vary the frequency of the novel sense to simulate more or less drastic changes in sense. Crucially, using manually sense-annotated data allows us to create synthetic examples of semantic change based on real word senses; the resulting synthetic examples are constructed from related word senses, increasing our confidence that they are plausible examples of semantic change.

4.1 Data and preprocessing

We use the 32 verbs and 20 nouns in the data from the Senseval-3 English lexical sample task (Mihalcea et al., 2004) to create synthetic examples of semantic change. We restrict ourselves to the training portion of this data, which consists of from 26 to 266 manually sense-annotated usages of each word. We refer to the set of senses of each word w as S_w. We select the second most frequent sense, $s \in S_w$, ignoring instances which are assigned multiple senses or annotated as unknown. (We select the second most frequent sense because it tends to be moderately frequent, but does not account for the majority of usages.) We then partition the sense-tagged usages of w into three approximately equal-sized parts w_A, w_B, and w_C. In this discussion, w_A and w_B consist of instances whose corresponding manually-tagged senses are in $S_w - \{s\}$. They are used to simulate w *not* undergoing semantic change; the frequency of any given sense of w is approximately the same in both w_A and w_B. We refer to this as the "no change" condition.

By contrast, w_C consists of usages of w with any sense in S_w such that the percentage of usages of s is approximately r. These usages are used to simulate w undergoing semantic change, namely w acquires a novel sense; specifically, sense s is not present in w_A, but accounts for roughly r% of the usages in w_C. We refer to this as the "change" condition.

The usages of w are divided into w_A, w_B, and w_C in such a way that as much of the Senseval data as possible is used while still maintaining the appropriate ratio of senses in w_C and keeping the sizes of w_A, w_B and w_C approximately equal.

When a word changes in meaning by taking on a new sense, it typically goes through a period where it also maintains its original senses (Campbell, 2004, Chapter 9). For example, the predominant meaning of *gay* has changed from 'merry' to 'homosexual', but the 'merry' sense is still understood. Furthermore, *gay* has taken on another sense—often considered offensive—meaning roughly 'of poor quality'. We model this aspect of semantic change in our synthetic examples through the choice of r (the percentage of usages of the novel sense in the "change" condition). Values of r close to 100 simulate a word that has lost its original senses and taken on an entirely new meaning, while values of r closer to 0 correspond to a novel but relatively infrequent sense.

We tag all sentences in our dataset with their part-of-speech using the TreeTagger (Schmid, 2004), and then parse the sentences using the MALT dependency parser (Nivre et al., 2007) with the provided pre-trained linear model for English.[5] For each dependency triple w_1, r, w_2 in the resulting parses we add a corresponding triple $w_2, r\text{-inverse}, w_1$. Then for each word w, we extract all dependency triples with head w, and compute the similarity between w_A and w_B, and w_A and

[4] Here we assume A and B are approximately equal in size. Differences in corpus size could be accounted for through a partitioning scheme that ensures that comparisons are made between equal-size corpus parts.

[5] `http://maltparser.org/`

Table 1: Average % accuracy over 32 verbs and 20 nouns using the Cosine and Newness similarity metrics for differing values of r—the percentage of usages of the novel sense in the "change" condition. Items marked * are significantly better ($p < .05$) than a random baseline.

	Average % accuracy			
	Verbs		Nouns	
r	Cosine	Newness	Cosine	Newness
100	85*	78*	82*	79*
80	79*	70*	81*	79*
60	75*	66*	77*	74*
40	67*	57*	71*	67*
20	58*	51	60*	54

w_C, using the cprob association measure.[6] In this first set of experiments we compute similarity directly—i.e., not taking into account intra-corpus similarity—due to the rather small number of tokens for each item in these experiments.[7]

4.2 Experimental setup and results

For each value of r in $\{20, 40, ..., 100\}$, we randomly repeat the process described in Section 4.1 100 times. For each word w and each trial, we compute the difference between the "change" and "no change" conditions using each similarity metric, e.g., for Newness we compute $\text{Newness}(w_A, w_B) - \text{Newness}(w_A, w_C)$. If this difference is positive, the method is scored as correct, and as incorrect otherwise. We compute the accuracy over the 100 trials for each word, and then compute the average accuracy over all verbs and nouns separately. In Table 1 we report the average accuracy over all verbs and nouns in our dataset for each similarity metric and value of r.

As expected, the accuracies are higher for experimental setups with higher values of r, i.e., words with more-frequent novel senses are easier to identify. For each similarity metric and each value of r, we compare the accuracies to a chance baseline of 50% accuracy using a one-sided one-sample Wilcoxon signed-rank test (a non-parametric alternative to a one-sample t-test). The differences are significant ($p < .05$) in all cases except those not marked with * in Table 1. This encouraging result demonstrates that—under the conditions of the present experiment—this method is better able to identify whether a given word has acquired a novel sense than a random baseline, even if that novel sense accounts for only 20% of the usages.

It is not clear how to statistically test the difference between Cosine and Newness over all values of r. Because the results using different values of r are based on the same usages of the experimental items, the paired differences obtained for different values of r are not independent. Therefore statistical tests such as the paired t-test and the Wilcoxon signed-rank test are not appropriate. We can, however, test the difference between the similarity metrics for a *specific* value of r. For the 32 verbs, the accuracy using Cosine is significantly better ($p < .05$) than that using Newness according to a two-sided paired Wilcoxon signed-rank test, for each value of r. For the 20 nouns, Cosine is significantly better than Newness in 1 case ($r = 20$).

5 Synthetic examples of semantic change from near synonyms

Many words have multiple related senses; i.e., many words are polysemous. Therefore, when a word takes on a novel sense, it is often related to its older senses. For example, the relatively recent sense of *post* referring to online message boards and social media is related to its older senses

[6] Because we are using a dataset of usages, as opposed to a corpus, we cannot compute the marginal frequencies required for I here.

[7] Although the similarity of w_A and w_B could be viewed as a measure of intra-corpus similarity, here we are using these samples to simulate two corpora in which w has the same set of senses.

referring to physical message boards. In this section we use this knowledge of semantic change to construct plausible synthetic examples of lexical semantic change based on near synonyms. By replacing varying numbers of usages of a given word in a corpus with a near synonym, we can simulate different degrees of sense change; i.e., we simulate different degrees of a word acquiring a novel, but related, sense. This corresponds to the view that semantic change is gradual, with words taking on new senses, and their original senses remaining or fading from usage over time. Moreover, these synthetic examples can be viewed as a type of *widening*, a common type of semantic change in which the use of a word is extended to additional contexts (Campbell, 2004). In contrast to the experiments in Section 4, here we have access to usages of our synthetic examples of semantic change in a corpus, and are able to calculate association measures such as I that require marginal frequencies; however, this comes at the expense of our synthetic examples of semantic change no longer being based on manually-identified word senses.

5.1 Data and experimental items

For these experiments we use the New York Times Annotated Corpus (Sandhaus, 2008) for the year 1990, a sample of approximately 50 million words of non-newswire text from the *New York Times*. We tag and parse this corpus using the TreeTagger and MALT parser as in Section 4.1.

We require a set of words to use in the creation of our synthetic examples of semantic change. These should be frequent words—so that we can accurately estimate our association measures—but not very frequent and highly polysemous words, such as light verbs, e.g., *give*, *take*, and *make*. We select the 1000 verbal lemmas and 1000 nominal lemmas in our corpus with frequency rank 101–1100 amongst verbal and nominal lemmas, respectively. From this set of words we then identify all words w such that the first WordNet (Fellbaum, 1998) synset of w has at least one lemma, near_synonym(w), that is not w and has frequency greater than 500; in the case that multiple lemmas satisfy these conditions, we randomly choose one.[8] For example, near_synonym(*propose*) is *suggest*. Of the 1000 verbs and nouns, 252 verbs and 214 nouns, respectively, satisfy these conditions. These items are used to create synthetic examples of semantic change. The remaining 748 verbs and 786 nouns are used as examples of words that do not undergo semantic change.

5.2 Synthetic example creation

We randomly partition the documents in our corpus into two parts, referred to as A and B. For a given word w, we refer to its usages in corpus parts A and B as w_A and w_B, respectively.

A synthetic example of semantic change is formed from w_A, a random sample of $(1-\frac{r}{100})\cdot|w_B|$ usages from w_B, and $\frac{r}{100}\cdot|w_B|$ usages of near_synonym(w) from corpus part B.

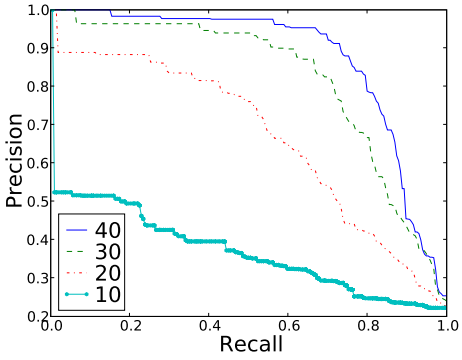
We consider values of r—the proportion of usages in corpus part B that correspond to a novel sense—in $\{10,20,30,40\}$. As an example, to form a synthetic example for *propose* with $r=20$, we use all usages of *propose* in A, a sample of 80% of the usages of *propose* in B, and a sample of usages of *suggest* in B of size equal to 20% of the number of usages of *propose* in B.

5.3 Experimental setup

In these experiments, we consider both of the association measures (I and cprob) and both of the similarity metrics (Cosine and Newness) discussed in Sections 3.1 and 3.2. We further consider similarity computed directly between two corpora, and taking into account intra-corpus similarity (Section 3.3).

We randomly repeat the process described in Section 5.2 5 times. For each of the 5 trials, we compute the similarity of each synthetic example of semantic change between the corpus parts (using both association measures, both similarity metrics, and both computing inter-corpus similarity directly and taking into account intra-corpus similarity). We similarly compute similarity for

[8] Although for a given word w we choose near_synonym(w) from the same synset, these words will typically differ somewhat in their usage and in the contexts in which they appear.



r	Top-100 % accuracy Nouns	Verbs
40	97	94
30	93	76
20	81	55
10	48	36
Baseline	25	21

Figure 1: Left: Interpolated precision–recall curves for identifying synthetic examples of nouns using the association measure I and Cosine similarity metric. Results for experiments in which the novel sense accounts for 10–40% of the usages of a noun are shown. Right: Top-100 % accuracy for identifying synthetic examples of nouns and verbs using the association measure I and Cosine similarity metric in experiments where the novel sense accounts for 10–40% of the usages of a word. The performance of a random baseline is also shown

the examples of words that do not change in meaning. We compute the average similarity of each item across the 5 trials. For both types of synthetic semantic change—nouns and verbs—we rank all experimental items—synthetic examples of semantic change, and examples of words that don't undergo semantic change—by average similarity. We then compute interpolated precision–recall curves for the synthetic examples of semantic change.

5.4 Results

The left panel of Figure 1 presents interpolated precision–recall curves for identifying synthetic examples of nouns using the association measure I, Cosine similarity metric, and taking intra-corpus similarity into account, an experimental setup similar to one considered by Peirsman et al. (2010). Results for experiments using $r = 10$–40 (i.e., the percentage of usages corresponding to the novel sense) are shown. As in the experiments in Section 4.2, stronger instances of semantic change are easier to identify. Results using synthetic examples of verbs (not shown) are similar.

Because we expect the candidates returned by our method to be manually examined by a lexicographer, we are particularly interested in whether the top-ranked items are correct. We therefore also consider the accuracy of our method on the top 100 ranked items. Top-100 % accuracy for the same experimental conditions as above, and also for experiments using synthetic examples of verbs, is shown in the right panel of Figure 1; a random baseline is also shown.[9] In all cases, the accuracy is significantly better than the random baseline using a one-tailed binomial test ($p < .05$). For the rest of this study we focus on experiments with $r = 10$ because we are especially interested in cases where the novel sense is relatively infrequent.

We now consider whether the cprob association measure, and newly-proposed similarity metric Newness, are an improvement over the measures I and Cosine. The left panel of Figure 2 shows results on nouns using both association measures and similarity metrics—taking intra-corpus similarity into account—in experiments with $r = 10$. The methods using cprob outperform those using I. In terms of top-100 accuracy, cprob and Newness (93%), and cprob and Cosine (77%), are both significantly better than the best method using I (I and Cosine—48%) using a two-tailed binomial test ($p \ll .05$ in both cases). The best performance is achieved using cprob in combination with Newness; the top-100 accuracy of this method is significantly better than that of the next best

[9] The baseline is calculated as the number of synthetic examples of semantic change divided by the total number of items.

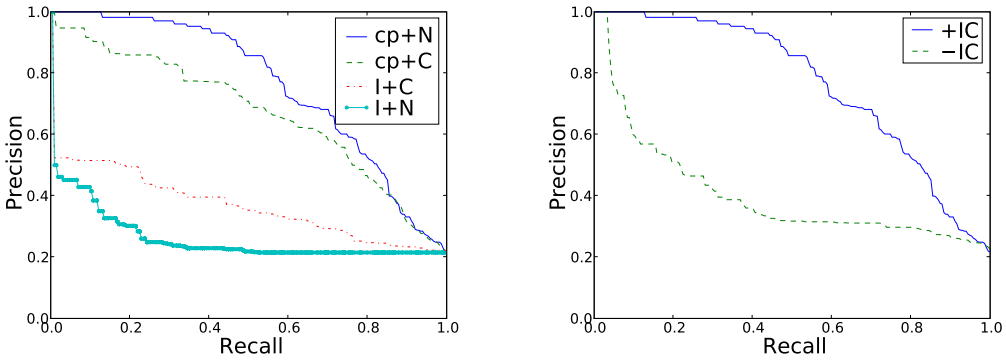


Figure 2: Left: Interpolated precision–recall curves for identifying synthetic examples of nouns with $r = 10$ using the association measures cprob (cp) and I, and the Cosine (C) and Newness (N) similarity metrics. Right: Interpolated precision–recall curves for identifying synthetic examples of nouns with $r = 10$ using the cprob association measure and Newness similarity metric, computing similarity based on intra-corpus similarity (+IC) and directly (−IC).

performing method—cprob and Cosine—using a two-tailed binomial test ($p \ll .05$). In the case of verbs (not shown) the methods using cprob are again significantly better than those using I; however, in this case the accuracy using cprob and Newness (72%) is not significantly different than that using cprob and Cosine (64%, $p > .05$).

In all the experiments so far in this section, we have taken intra-corpus similarity into account. We now examine the impact of this process on performance. The right panel of Figure 2 shows results using our best performing method—cprob in combination with Newness—for synthetic examples of nouns in experiments with $r = 10$. (Results using verbs—not shown—are similar.) The results taking intra-corpus similarity into account are much better than those for which similarity is computed directly. The top-100 accuracy using intra-corpus similarity (93%) is significantly better than that calculating similarity directly (47%) using a two-tailed binomial test ($p \ll .05$). These results confirm Peirsman et al.'s (2010) observation that it is important to take into account information about intra-corpus variation when identifying differences between corpora.

6 Conclusions

We have proposed a method for identifying words that are used in a novel sense in one corpus with respect to another. In contrast to previous work in this area, we focused specifically on infrequent novel senses, the identification of which is a challenge in lexicography due to the vast amount of text being produced nowadays. Our proposed method outperformed random baselines, even when evaluated on rather subtle changes in sense. Furthermore, the combination of a very simple association measure (cprob) and our newly-proposed asymmetrical similarity metric (Newness) outperformed methods using a standard association measure and symmetrical similarity metric.

Given the challenges of evaluation for this task—namely a lack of gold standard data—we further proposed the use of two different types of synthetic examples of semantic change to empirically assess performance on this task. Sense-tagged data was used to construct a small number of synthetic examples of semantic change based on real word senses, while near synonyms were used to build greater numbers of synthetic examples of semantic change.

Although we motivated this work in the context of identifying novel senses, these methods can be applied to any pair of comparable corpora to identifying words with different senses in those corpora. In our ongoing work we are applying our methods to pairs of corpora to manually assess its performance. Given the expense of this manual evaluation, synthetic examples of semantic

change remain attractive as a means for determining the strengths and weaknesses of approaches to this task, and for selecting approaches for further manual evaluation.

References

Campbell, L. 2004. *Historical Linguistics: An Introduction*. MIT Press, Cambridge, MA.

Cook, P. and S. Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 28–34, Valletta, Malta.

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

Gale, W. A., K. W. Church, and D. Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In Goldman, R., Norvig, P., Charniak, E., and Gale, B., editors, *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 54–60. AAAI Press, Menlo Park, CA.

Gaustad, T. 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 (ACL 2001) – Proceedings of the Student Research Workshop*, pages 61–66, Toulouse, France.

Gulordava, K. and M. Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, Scotland.

Joanis, E., S. Stevenson, and D. James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

Kilgarriff, A. and D. Tugwell. 2002. Sketching words. In Corréard, M.-H., editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 125–137. Euralex, Grenoble, France.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (ACL/COLING 1998)*, pages 768–774, Montreal, Canada.

McCarthy, D., R. Koeling, J. Weeds, and J. Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

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

Nakov, P. I. and M. A. Hearst. 2003. Category-based pseudowords. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 67–69, Edmonton, Canada.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryiğit, S. Kübler, S. Marinov, and E. Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

O'Donovan, R. and M. O'Neil. 2008. A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In *Proceedings of the 13th Euralex International Congress*, pages 571–579, Barcelona, Spain.

Otrusina, L. and P. Smrz. 2010. A new approach to pseudoword generation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1195–1199, Valletta, Malta.

Peirsman, Y., D. Geeraerts, and D. Speelman. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4):469–491.

Sagi, E., S. Kaufmann, and B. Clark. 2009. Semantic density analysis: Comparing word meaning across time and space. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece.

Sandhaus, E. 2008. The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia, PA.

Schmid, H. 2004. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Schütze, H. 1992. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.