Detecting Reliable Novel Word Senses: A Network-Centric Approach

Abhik Jana
Indian Institute of Technology
Kaharagpur, India
abhik.jana@iitkgp.ac.in

Animesh Mukherjee
Indian Institute of Technology
Kaharagpur, India
animeshm@gmail.com

Pawan Goyal
Indian Institute of Technology
Kaharagpur, India
pawang@cse.iitkgp.ac.in

ABSTRACT

In this era of Big Data, due to expeditious exchange of information on the web, words are being used to denote newer meanings, causing linguistic shift. With the recent availability of large amounts of digitized texts, an automated analysis of the evolution of language has become possible. Our study mainly focuses on improving the detection of new word senses. This paper presents a unique proposal based on network features to improve the precision of new word sense detection. For a candidate word where a new sense (birth) has been detected by comparing the sense clusters induced at two different time points, we further compare the network properties of the subgraphs induced from novel sense cluster across these two time points. Using the mean fractional change in edge density, structural similarity and average path length as features in an SVM classifier, manual evaluation gives precision values of 0.86 and 0.74 for the task of new sense detection, when tested on 2 distinct time-point pairs, in comparison to the precision values in the range of 0.23-0.32, when the proposed scheme is not used. The outlined method can therefore be used as a new post-hoc step to improve the precision of novel word sense detection in a robust and reliable way where the underlying framework uses a graph structure. Another important observation is that even though our proposal is a post-hoc step, it can be used in isolation and that itself results in a very decent performance achieving a precision of 0.54-0.62. Finally, we show that our method is able to detect the well-known historical shifts in 80% cases.

CCS CONCEPTS

• **Computing methodologies** → **Semantic networks**; **Lexical semantics**; *Temporal reasoning*;

KEYWORDS

Novel Sense Detection; Distributional Thesaurus Network; Complex Network Measures

ACM Reference Format:
Abhik Jana, Animesh Mukherjee, and Pawan Goyal. 2019. Detecting Reliable Novel Word Senses:, A Network-Centric Approach. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3297280.3297376

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04…$15.00
https://doi.org/10.1145/3297280.3297376

1 INTRODUCTION

How do words develop new senses? How does one characterize semantic change? Is it possible to develop algorithms to track semantic change by comparing historical data at scale? In order to extract meaningful insights from these data, a very important step is to understand the contextual sense of a word, e.g., does the word 'bass' in a particular context refer to fish or is it related to music?

Most data-driven approaches so far have been limited to either word sense induction where the the goal is to automatically induce different senses of a given word in an unsupervised clustering setting, or word sense disambiguation where a fixed sense inventory is assumed to exist, and the senses of a given word are disambiguated relative to the sense inventory. However in both these tasks, the assumption is that the number of senses that a word has, is static, and also the senses exist in the sense inventory to compare with. They attempt to detect or induce one of these senses depending on the context. However, natural language is dynamic, constantly evolving as per the users' needs which leads to change of word meanings over time. For example, by late 20th century, the word 'float' has come up with the 'data type' sense whereas the word 'hot' has started corresponding to the 'attractive personality' sense.

1.1 Recent advancements

Recently, with the arrival of large-scale collections of historic texts and online libraries such as Google books, a new paradigm has been added to this research area, whereby the prime interest is in identifying the temporal scope of a sense [10, 14, 16, 25] which, in turn, can give further insights to the phenomenon of language evolution. Some recent attempts [5, 8, 11, 12, 15] also have been made to model the dynamics of language in terms of word senses.

One of the studies in this area has been presented by Mitra *et al.* [19] where the authors show that at earlier times, the sense of the word 'sick' was mostly associated to some form of illness; however, over the years, a new sense associating the same word to something that is 'cool' or 'crazy' has emerged. Their study is based on a unique network representation of the corpus called a distributional thesauri (DT) network built using Google books syntactic n-grams. They have used unsupervised clustering techniques to induce a sense of a word and then compared the induced senses of two time periods to get the new sense for a particular target word.

1.2 Limitations of the recent approaches

While Mitra *et al.* [19] reported a precision close to 0.6 over a random sample of 49 words, we take another random sample of 100 words separately and repeat manual evaluation. When we extract the novel senses by comparing the DTs from 1909-1953 and 2002-2005, the precision obtained for these 100 words is as low as 0.32. Similarly if we extract the novel senses comparing the DTs of 1909-1953 with 2006-2008, the precision stands at 0.23. We then explore another unsupervised approach presented in Lau *et al.* [16] over the same Google books corpus[1], apply topic modeling for sense induction and directly adapt their similarity measure to get the new senses. Using a set intersecting with the 100 random samples for Mitra *et al.* [19], we obtain the precision values of 0.21 and 0.28, respectively. Clearly, none of the precision values are good enough for reliable novel sense detection. This motivates us to devise a new approach to improve the precision of the existing approaches. Further, being inspired by the recent works of applying complex network theory in NLP applications like co-hyponymy detection [13], evaluating machine generated summaries [20], detection of ambiguity in a text [4], etc. we opt for a solution using complex network measures.

1.3 Our proposal and the encouraging results

We propose a method based on the network features to reduce the number of false positives and thereby, increase the overall precision of the method proposed by Mitra *et al.* [19]. In particular, if a target word qualifies as a 'birth' as per their method, we construct two induced subgraphs of those words that form the cluster corresponding to this 'birth' sense from the corresponding distributional thesauri (DT) networks of the two time points. Next we compare the following three network properties: (i) the edge density, (ii) the structural similarity and (iii) the average path length [27, 29] of the two induced subgraphs from the two time points. A remarkable observation is that although this is a small set of only three features, for the actual 'birth' cases, each of them has a significantly different value for the later time point and are therefore very discriminative indicators. In fact, the features are so powerful that even a small set of training instances is sufficient for making highly accurate predictions.

Results: Manual evaluation of the results by 3 evaluators shows that this classification achieves an overall precision of 0.86 and 0.74 for the two time point pairs over the same set of samples, in contrast with the precision values of 0.32 and 0.23 by the original method. Note that we would like to stress here that an improvement of **more than double** in the precision of novel sense detection that we achieve has the potential to be the new stepping stone in many NLP and IR applications that are sensitive to novel senses of a word.

1.4 Detecting known shifts

Further we also investigate the robustness of our approach by analyzing the ability to capture known historical shifts in meaning. Preparing a list of words that have been suggested by different prior works as having undergone sense change, we see that 80% of those words get detected by our approach. We believe that the ability to detect such diachronic shifts in data can significantly enhance various standard studies in natural language evolution and change.

[1]http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html, we use 'triarcs' dataset from 'English All'

1.5 Impact

We stress that our work could have strong repercussions in historical linguistics [1]. Besides, lexicography is also expensive; compiling, editing and updating sense inventory entries frequently remains cumbersome and labor-intensive. Time specific knowledge would make the word meaning representations more accurate. A well constructed semantic representation of a word is useful for many natural language processing or information retrieval systems like machine translation, semantic search, disambiguation, Q&A, etc. For semantic search, taking into account the newer senses of a word can increase the relevance of the query result. Similarly, a disambiguation engine informed with the newer senses of a word can increase the efficiency of disambiguation, and recognize senses uncovered by the inventory that would otherwise have to be wrongly assigned to covered senses. Above all, a system having the ability to perceive the novel sense of a word can help in automatic sense inventory update by taking into account the temporal scope of senses.

2 RELATED WORK

Our work broadly classifies under data-driven models of language dynamics. One of the first attempts in this area was made by Erk [6], where the author tried to model this problem as an instance of outlier detection, using a simple nearest neighbor-based approach. Gulordava and Baroni [10] study the change in the semantic orientation of words using Google book n-grams corpus from different time periods. In another work, Mihalcea *et al.* [18] attempted to quantify the changes in word usage over time. Along similar lines, Jatowt and Duh [14] used the Google n-grams corpus from two different time points and proposed a method to identify semantic change based on the distributional similarity between the word vectors. Tahmasebi *et al.* [25] attempted to track sense changes from a newspaper corpus containing articles between 1785 and 1985. Efforts have been made by Cook *et al.* [3] to prepare the largest corpus-based dataset of diachronic sense differences. Attempts have been made by Lau *et al.* [17] where they first introduced their topic modeling based word sense induction method to automatically detect words with emergent novel senses and in a subsequent work, Lau *et al.* [16] extended this task by leveraging the concept of predominant sense. The first computational approach to track and detect statistically significant linguistic shifts of words has been proposed by Kulkarni *et al.* [15]. Recently, Hamilton *et al.* [12] proposed a method to quantify semantic change by evaluating word embeddings against known historical changes. In another work, Hamilton *et al.* [11] categorized the semantic change into two types and proposed different distributional measures to detect those types. Attempts have also been made to analyze time-series model of embedding vectors as well as time-indexed self-similarity graphs in order to hypothesize the linearity of semantic change by Eger *et al.* [5]. A dynamic Bayesian model of diachronic meaning change has been proposed by Frermann *et al.* [8]. Recently, researchers have also tried to investigate the reasons behind word sense evolution and have come up with computational models based on chaining [21]. Researchers also attempt to apply dynamic word embeddings as well to detect language evolution [23, 30], analyze temporal word analogy [24].

We now describe the two baselines that are relevant for our work.

Baseline 1: Mitra *et al.* [19] The authors proposed an unsupervised method to identify word sense changes automatically for nouns.

Datasets and graph construction: The authors used the Google books corpus, consisting of texts from over 3.4 million digitized English books published between 1520 and 2008. The authors constructed distributional thesauri (DT) networks from the Google books syntactic n-grams data [9]. In the DT network, each word is a node and there is a weighted edge between a pair of words where the weight of the edge is defined as the number of features that these two words share in common. A snapshot of the DT is shown in leftmost image of Figure 1. To study word sense changes over time, they divided the dataset across eight time periods; accordingly DT networks for each of these time periods were constructed separately.

Sense change detection: The Chinese Whispers algorithm [2] is used to produce a set of clusters for each target word by decomposing its neighbourhood in the DT network. The hypothesis is that different clusters signify different senses of a target word. The clusters for a target word 'float' is shown in the right image of Figure 1. The authors then compare the sense clusters extracted across two different time points to obtain the suitable signals of sense change. Specifically, for a candidate word w, a sense cluster in the later time period is called as a 'birth' cluster if at least 80% words of this cluster do not appear in any of the sense clusters from the previous time period. The authors then apply multi-stage filtering in order to obtain meaningful candidate words.

Baseline 2: Lau *et al.* [16]: The authors proposed an unsupervised approach based on topic modeling for sense induction, and showed novel sense identification as one of its applications. For a candidate word, Hierarchical Dirichlet Process [26] is run over a corpus to induce topics. The induced topics are represented as word multinomials, and are expressed by the top-N words in descending order of conditional probability. Each topic is represented as a sense of the target word. The words having highest probability in each topic represent the sense clusters. The authors treated the novel sense detection task as identifying those sense clusters, which did not align with any of the recorded senses in a sense repository. They used Jensen-Shannon (JS) divergence measure to compute alignment between a sense cluster and a synset. They computed JS divergence between the multinomial distribution (over words) of the topic cluster and that of the synset, and converted the divergence value into a similarity score. Similarity between topic cluster t_j and synset s_i is defined as,

$$sim(t_j, s_i) = 1 - JS(T \parallel S) \tag{1}$$

where T and S are the multinomial distributions over words for topic t_j and synset s_i, respectively, and $JS(X \parallel Y)$ is the Jensen-Shannon divergence for the distribution X and Y. Since we define novel senses while comparing sense clusters across two time points, we use the same JS measure to detect novel sense of a target word. A sense cluster in the newer time period denotes a new sense ('birth') only if its maximum similarity with any of the clusters in older time period is below a threshold, which we have set to 0.35 based on empirical observation.

3 PROPOSED NETWORK-CENTRIC APPROACH

Mitra *et al.* [19] selected 49 candidate 'birth' words from a total of 2789 candidate 'birth' words while comparing 1909-1953 DT with the 2002-2005 DT for manual evaluation; 31 words were found to be true positives and 18 words were false positives. We first study these 49 candidate 'birth' words and show that network features can be useful to discriminate the true positives from the false positives. For each of these candidate words w, we take the 'birth' cluster from 2002-2005, which is represented by a set of words S. According to our hypothesis, if the words in set S together represent a new sense for w in 2002-2005 which is not present in 1909-1953, the network connection among these words (including w) would be much more strong in the 2002-2005 DT than the 1909-1953 DT. The strength of this connection can be measured if we construct induced subgraphs of S from the two DTs and measure the network properties of these subgraphs; the difference would be more prominent for the actual 'birth' cases (true positives) than for the false 'birth' signals (false positives). Note that by definition, the nodes in an induced subgraph from a DT will be the words in S and there will be an edge between two words if and only if the edge exists in the original DT; we ignore the weight of the edge henceforth. Thus, the difference between the two subgraphs (one each from the older and newer DTs) will only be in the edge connections. Figure 2 takes one true positive ('register') and one false positive ('quotes') word from the set of 49 words and shows the induced subgraphs obtained by a subset of their 'birth' clusters across the two time points. We can clearly see that connections among the words in S is much stronger in the newer DT than in the older one in the case of 'registers', indicating the emergence of a new sense. In the case of 'quotes', however, the connections are not very different across the two time periods. We choose three *cohesion indicating* network properties, (i) the edge density, (ii) the structural similarity and (iii) the average path length, to capture this change.

Let $S = \{w_1, w_2, \ldots, w_n\}$ be the 'birth' cluster for w. Once we construct a graph induced by S from the DT, these network properties are measured as follows:

Edge Density (ED): ED is given by

$$ED = N_a/N_p \tag{2}$$

where N_a denotes the number of actual edges between $w_1, w_2, \ldots, w_n$ and N_p denotes the maximum possible edges between these, i.e., $\frac{n(n-1)}{2}$.

Structural Similarity (SS): For each pair of words (w_i, w_j) in the cluster S, the structural similarity $SS(w_i, w_j)$ is computed as:

$$SS(w_i, w_j) = \frac{N_c}{\sqrt{deg(w_i) * deg(w_j)}} \tag{3}$$

where N_c denotes the number of common neighbors of w_i and w_j in the induced graph and $deg(w_k)$ denotes the degree of w_k in the induced graph, for $k = i, j$. The average structural similarity for the cluster S is computed by averaging the structural similarity of all the word pairs.

Average Path Length (APL): To compute average path length of S, we first find the shortest path length between w and the words w_i, in the induced graph of S. Let spl_i denote the shortest path

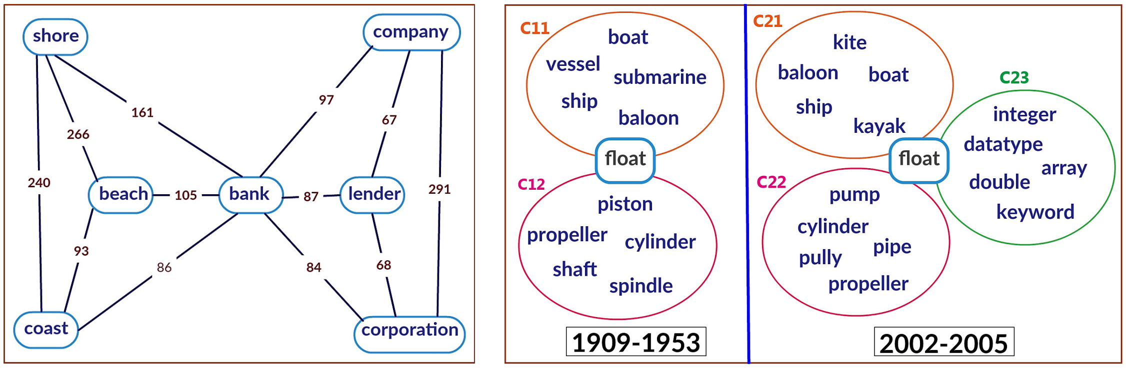

Figure 1: Left image is a sample snapshot of the Distributional Thesaurus Network from the time period 2002-2005 where each node represents a word and the weight of the edge is defined as the number of context features that these two words share in common. Right image shows Chinese Whisper clusters for the target word 'float' extracted from Google books syntactic n-gram corpus of both the time periods (1909-1953 and 2002-2005). A new sense of the word 'float' has emerged with the 'programming' related new cluster (C23) in 2002-2005.

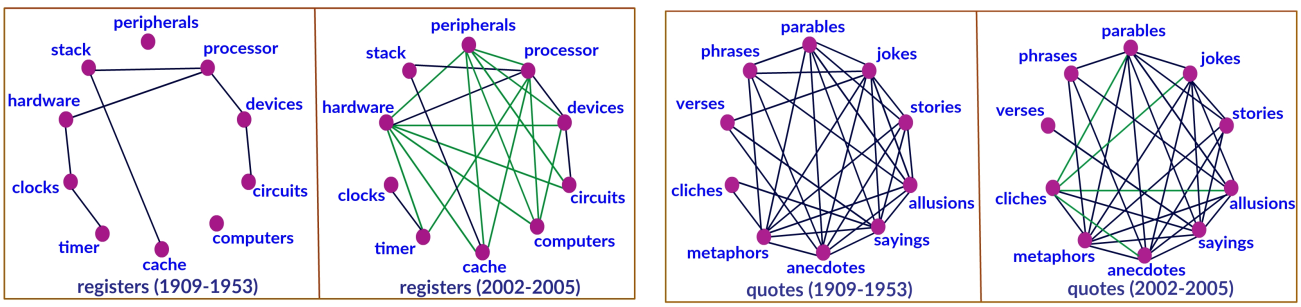

Figure 2: Induced subgraphs of the 'birth' clusters of 'registers' and 'quotes' for the two time periods (1909-1953 and 2002-2005). It shows that edge connections among the neighbours of 'registers' have increased significantly over time which leads to emergence of 'technology' related sense of 'registers' whereas the connections among the neighbours of 'quotes' are almost same over time, indicating non-emergence of any novel sense.

distance from w to w_i. The average path length is defined as:

$$APL = \sum_i spl_i/n \tag{4}$$

where n is the number of words in the cluster S.

Table 1 notes the values obtained for these network properties for the induced subgraphs of the reported 'birth' clusters for 'registers' and 'quotes' across the two time periods. The fractional changes observed for the three network properties show a clear demarcation between the two cases. Fractional change (Δ) of any network measure P is defined as,

$$\Delta(P) = (P(t_2) - P(t_1))/P(t_1) \tag{5}$$

where t_1 and t_2 are old and new timeperiods respectively. The change observed for the 'birth' cluster of 'registers' is significantly higher than that in 'quotes'[2].

We now compute these parameter values for all the 49 candidate cases. The mean values obtained for the true positives (TP) and false positives (FP) are shown in Table 2. The findings are consistent with those obtained for 'registers' and 'quotes'.

[2]As we have taken the 'birth' clusters from new time period (t_2), the words in the clusters are the direct neighbors of the target word always resulting in average path length of 1 in t_2

We, therefore, use the fractional changes in the three network properties over time as three features to classify the remaining candidate 'birth' words into true positives (actual 'birth') or false positives (false 'birth').

4 EXPERIMENTAL RESULTS

For experimental evaluation, we start with the 'birth' cases reported by Mitra *et al.* [19] – 2740 cases (after removing the 49 used in training) for 1909-1953 – 2002-2005 (T_1) and 2468 cases for 1909-1953 – 2006-2008 (T_2). We run Lau *et al.* [16] over these birth cases to detect 'novel' sense as per their algorithm. Separately, we also apply the proposed SVM classification model as a filtering step to obtain 'filtered birth' cases. This helps in designing a *comparative evaluation* of these algorithms as follows. From both the time point pairs (T_1 and T_2), we take 100 random samples from the birth cases reported by Mitra *et al.* [19] and get these manually evaluated. For the same 100 random samples, we now use the outputs of Lau *et al.* [16] and the proposed approach, and estimate the precision as well as recall of these.

To further evaluate the proposed algorithm, we perform two more evaluations. First, we take 60 random samples from each time point pair for computing precision of the 'filtered birth' cases. Secondly, we also take 100 random samples for each time point pair

Table 1: The network properties of the induced subgraphs of a true positive ('registers') and a false positive ('quotes') for the time periods 1909-1953 (t_1) and 2002-2005 (t_2).

Word	ED (t_1)	ED (t_2)	SS (t_1)	SS (t_2)	APL (t_1)	APL (t_2)	Δ (ED, SS, APL)
registers	0.108	0.546	0.076	0.516	1.9	1	4.045, 5.771, -0.9
quotes	0.858	0.833	0.835	0.622	1.72	1	-0.029, -0.255, -0.72

Table 2: Mean values of the network properties of the induced subgraphs of 31 true positives and 18 false positives for the time periods 1909-1953 (t_1) and 2002-2005 (t_2). The mean fractional changes (Δ) in network properties are significantly higher for the true positives (TP) as compared to the false positives (FP).

Word	ED (t_1)	ED (t_2)	SS (t_1)	SS (t_2)	APL (t_1)	APL (t_2)	Δ (ED, SS, APL)
TP	0.34	0.772	0.311	0.647	1.941	1	2.388, 4.654, -0.941
FP	0.576	0.828	0.574	0.681	1.828	1	0.747, 0.507, -0.828

for computing precision of our approach independently of Mitra *et al.* [19], i.e., the proposed approach is not informed of the 'birth' cluster reported by Mitra *et al.* [19], instead all the clusters in old and new time point are shown.

We perform all the evaluations manually and each of the candidate word is judged by 3 evaluators. These evaluators are graduate/post-graduate students, having good background in Natural Language Processing. They are unaware of each other, making the annotation process completely blind and independent. Evaluators are shown the detected 'birth' cluster from the newer time period and all the clusters from the older time period. They are asked to make a binary judgment as to whether the 'birth' cluster indicates a new sense of the candidate word, which is not present in any of the sense clusters of the previous time point[3]. Majority decision is taken in case of disagreement. In total, we evaluate the system for a set of as large as **520** words[4] which we believe is significant given the tedious manual judgment involved.

In this process of manual annotation, we obtain an inter-annotator agreement (Fleiss' kappa [7]) of **0.745**, which is *substantial* [28]. Table 3 shows three example words from T_1, their 'birth' clusters as reported in Mitra *et al.* [19] and the manual evaluation result. The first two cases belong to computer related sense of 'searches' and 'logging', which were absent from time point 1909-1953. On the other hand, the 'birth' cluster of 'pesticide' represents an old sense which was also present in 1909-1953. Similarly Table 4 shows manual evaluations results for 3 example cases, along with their novel sense as captured by Lau *et al.* [16].

Table 3: Example 'birth' clusters reported in Mitra *et al.* [19] and manual evaluation.

Word	'birth' cluster	Manual Evaluation
searches	folders, templates, syntax, formats, . . .	**Yes**, technology related sense
logging	server, console, security, service, . . .	**Yes**, technology related sense
pesticide	fertilizer, sediment, waste, . . .	**No**

Comparative evaluation: Only 32 and 23 words out of the 100 random samples from two time point pairs are evaluated to be actual 'birth's, respectively, thus giving precision scores of 0.32 and 0.23 for Mitra *et al.* [19]. Evaluation results for the same set of random samples after applying the approach outlined in Lau et al. [16] are presented in Table 5. Since the reported novel sense cluster can in principle be different from the 'birth' sense reported by the method of Mitra et al. [19] for the same word, we get the novel sense cases manually evaluated by 3 annotators (42 and 28 cases for the two time periods, respectively). Note that for these 100 random samples (that are all marked 'true' by Mitra et al. [19]), it is possible to find an upper bound on the recall of Lau et al. [16]'s approach automatically. While the low recall might be justified because this is a different approach, even the precision is found to be in the same range as that of Mitra et al. [19].

Table 4: Example novel senses as per Lau *et al.* [16] and manual evaluation.

Word	Novel sense	Manual Evaluation
stereo	system, player, computer, . . .	**Yes**, technology related sense
mailbox	email, pages, postal, . . .	**Yes**, technology related sense
acidification	acidosis, renal, distal, urinary, . . .	**No**

Table 6 presents the evaluation results for the same set of 100 random samples after using the proposed SVM filtering. We see that the filtering using SVM classification improves the precision for both the time point pairs (T_1 and T_2) significantly, boosting it from the range of 0.23-0.32 to 0.74-0.86. Note that, as per our calculations, indeed the recall of Mitra et al. [19] would be 100% (as we are taking random samples for annotation from the set of reported 'birth' cases by Mitra et al. [19] only). Even then Mitra et al. [19]'s F-measure ranges from 0.37-0.48 while ours is 0.67-0.68. Table 7 represents some of the examples which were declared as 'birth' by Mitra *et al.* [19] but SVM filtering correctly flagged them as 'false birth'. The feature values in the third column clearly show that the network around the words in the detected 'birth' cluster did not change much and therefore, the SVM approach could correctly flag these. Considering the small training set, the results are highly encouraging. We also obtain decent recall values for the two time point pairs, giving an overall F-measure of 0.67-0.68.

Table 5: Evaluation of the approach presented in Lau *et al.* [16] with accuracy for 100 random samples.

Time-point	Lau *et al.* [16]			
	# Novel senses	Precision	Recall	F-measure
T_1	1189	0.21	0.28	0.24
T_2	787	0.28	0.35	0.31

Further, we check if we can meaningfully combine the results reported by both the methods of Mitra *et al.* [19] and Lau *et al.* [16] for more accurate sense detection; and how does this compare with

[3] An anonymized sample evaluation page can be seen here: https://kwiksurveys.com/s/7TfSoYF2

[4] 100+100+60, per time point pair (T_1 and T_2)

Table 6: Evaluation of the SVM-based filtering with accuracy reported for 100 random samples.

Time-point	SVM filtering			
	# birth cases	Precision	Recall	F-measure
T_1	318	**0.86**	0.56	0.68
T_2	329	**0.74**	0.61	0.67

Table 7: Example cases, which Mitra *et al.* [19] declared as true 'birth' but SVM filtering correctly filtered

Word	'birth' cluster	Δ(ED, SS, APL)
guaranty	acknowledgement, presumption, kind, . . .	0.11, -0.07, -0.5
troll	shellfish, salmon, bait, trout, tuna, . . .	-0.04, -0.18, -0.84
nightcap	supper, lunch, dinner, nap, luncheon, . . .	0.04, -0.17, -0.75

the SVM based filtering. Therefore, we filter the words, which are reported as 'birth' by both these methods and the reported 'birth' sense clusters have a non-zero overlap. Out of 2789 and 2468 cases reported as 'birth' by the method of Mitra *et al.* [19], we obtain 132 and 86 cases respectively as having an overlapping sense cluster with that obtained using Lau's method. Two such examples are shown in Table 8; both the senses look quite similar. Table 9 shows the accuracy results obtained using this approach. Only 6 and 2 words out of those 100 samples were flagged as 'birth' for the two time points T_1 and T_2 respectively. Thus, the recall is very poor. While the precision improves slightly for T_1 (4 out of 6 are correct), it is worse for T_2 (only 1 out of 6 words is correct). The results confirm that the proposed SVM classification approach works better than both the approaches, individually as well as combined.

Table 8: Example cases, which Mitra *et al.* [19] declared as 'birth' represent the same sense as obtained using Lau *et al.* [16] (T_1).

Word	'birth' cluster as reported in Mitra *et al.* [19]	Novel senses as obtained using Lau *et al.* [16]
burgers	rice, pizza, fries, drinks, entrees, desserts . . .	fries, orders, drinks, entrees, desserts . . .
semantic	syntactic, analytic, pragmatic, lexical, metaphoric . . .	syntactic, pragmatic, lexical, aspect, context . . .

Table 9: Evaluation of the intersection set while taking gold standard annotation of Mitra *et al.* [19].

time point	Precision	Recall	F-measure
T_1	0.67 (4/6)	0.13 (4/32)	0.22
T_2	0.5 (1/2)	0.043 (1/23)	0.08

Feature analysis: We therefore move onto further feature analysis and error analysis of the proposed approach. To validate the usefulness of all the identified features, we perform feature leave-one-out experiments. The results for T_1 are presented in Table 10 and 11. We see that F-measure drops as we leave out one of the features. While $\{ED, SS\}$ turns out to be the best for precision, $\{SS, APL\}$ gives the best recall. Table 12 provides three examples to illustrate the importance of using all the three features. For the word 'newsweek', using $\{ED, APL\}$ and for the word 'caring', using $\{ED, SS\}$ could not detect those as 'birth'. Only when all the three features are used, these cases are correctly detected as 'birth'. Edge density, on the other hand is very crucial for improving precision. For instance, when only $\{SS, APL\}$ are used, words like 'moderators' are wrongly flagged as 'true'. Such cases are filtered out when all the three features are used.

Table 10: Feature leave-one-out results (T_1).

Features used	Precision	Recall	F-measure
Δ(ED, SS)	0.85	0.53	0.65
Δ(ED, APL)	0.84	0.5	0.62
Δ(SS, APL)	0.81	0.56	0.66
Δ(ED, SS, APL)	**0.86**	0.56	**0.68**

Table 11: Feature leave-one-out results (T_2).

Features used	Precision	Recall	F-measure
Δ(ED, SS)	0.72	0.56	0.63
Δ(ED, APL)	0.73	0.6	0.66
Δ(SS, APL)	0.66	0.61	0.63
Δ(ED, SS, APL)	**0.74**	0.61	**0.67**

Table 12: Example cases to show the utility of all the features (T_1). The true positive cases like 'newsweek' and 'caring' get successfully detected whereas 'moderators' gets successfully detected as false positive if all the three features are considered together.

Word	'birth' cluster	Δ(ED, SS, APL)
newsweek	probation, counseling, . . .	0.82, 1.58, -1.3
caring	insightful, wise, benevolent, . . .	0.2, 0.13, -2.21
moderators	correlate, function, determinant, . . .	0.56, 0.44, -1.78

Extensive evaluation of the proposed approach: We first take 60 random samples each from the 'birth' cases reported by the SVM filtering for the two time point pairs, T_1 (from 318 cases) and T_2 (from 329 cases). The precision values of this evaluation are found to be **0.87** (52/60) and **0.75** (45/60) respectively, quite consistent with those reported in Table 6. We did another experiment in order to estimate the performance of our model for detecting novel sense, independent of the method of Mitra *et al.* [19]. We take 100 random words from the two time point pairs (T_1 and T_2), along with all the induced clusters from the newer time period and run the proposed SVM filtering approach to flag the novel 'birth' senses. According to our model, for T_1 and T_2 respectively, 16 and 15 words are flagged to be having novel sense achieving precision values of 0.54 and 0.62 on manual evaluation, which itself is quite decent. Note that, for some cases, multiple clusters of a single word have been flagged as novel senses and we observe that these clusters hold a similar sense.

Error analysis: We further analyze the cases, which are labeled as 'true birth' by the SVM but are evaluated as 'false' by the human evaluators. We find that in most of such cases, the sense cluster reported as 'birth' contained many new terms (and therefore, the network properties have undergone change) but the implied sense was already present in one of the previous clusters with *very few common words* (and therefore, the new cluster contained > 80% new words, and is being reported as 'birth' in Mitra *et al.* [19]). Two such examples are given in Table 13. The split-join algorithm proposed in Mitra *et al.* [19] needs to be adapted for such cases.

We also analyze the 'false positive' cases, which are labeled as 'false birth' by the SVM filtering but are evaluated as 'true' by the human evaluators. Two such examples are given in Table 14. By looking at the feature values of these cases, it is clear that the network structure of the induced subgraph is not changing much, yet they undergo sense change. The probable reason could be that

Table 13: Example 'false positives' after SVM filtering (T_1). These words are flagged 'true birth' by SVM but manually evaluated as 'false'.

Word	'birth' cluster	Old cluster
aftercare	care, clinic, outpatient, . . .	treatment, therapy, hospitalization, . . .
electrophoresis	labeling, analysis, profiling, . . .	analysis, counting, procedure, . . .

the target word was not in the network of the induced subgraph in the old time point and enters into it in the new time point. Our SVM model is unable to detect this single node injection in a network so far. Handling these cases would be an immediate future step to improve the recall of the system.

Table 14: Example cases, labeled by SVM as 'false birth' but flagged as 'true birth' by annotators (T_1). The fractional change of the network measures is very low, leading to erroneous classification by SVM.

Word	'birth' cluster	Δ(ED, SS, APL)
baseplate	flywheel, cylinder, bearings, . . .	0.06, -0.08, -0.84
grating	beam, signal, pulse, . . .	0.2, -0.05, -0.88

5 DETECTION OF KNOWN SHIFTS

So far, we have reported experiments on discovering novel senses from data and measured the accuracy of our method using manual evaluation. In this section, we evaluate the diachronic validity of our method on another task of detecting known shifts. We test, whether our proposed method is able to capture the known historical shifts in meaning. For this purpose, we create a reference list L of **15** words that have been suggested by prior work [5, 11, 12] as having undergone a linguistic change and emerging with a novel sense. Note that, we focus only on nouns that emerge with a novel sense between 1900 and 1990. The goal of this task is to find out the number of cases, our method is able to detect as novel sense from the list L, which in turn would prove the robustness of our method.
Data: Consistent with the prior work, we use the Corpus of Historical American (COHA)[5]. COHA corpus is carefully created to be genre balanced and is a well constructed prototype of American English over 200 years, from the time period 1810 to 2000. We extract the raw text data of two time slices: 1880-1900 and 1990-2000 for our experiment.

Table 15: Example cases, from the training set for the experiment on detecting known shifts. Evaluation has been done by annotators

Word	'birth' cluster	Manual Evaluation
caller	phone, message, operator, customer, . . .	**Yes**, communication system related sense
courier	transport, purchase, company, delivery, . . .	**Yes**, marketing related sense
public	student, economist, general, . . .	**No**
richness	joy, happiness, stress, . . .	**No**

[5]https://corpus.byu.edu/coha

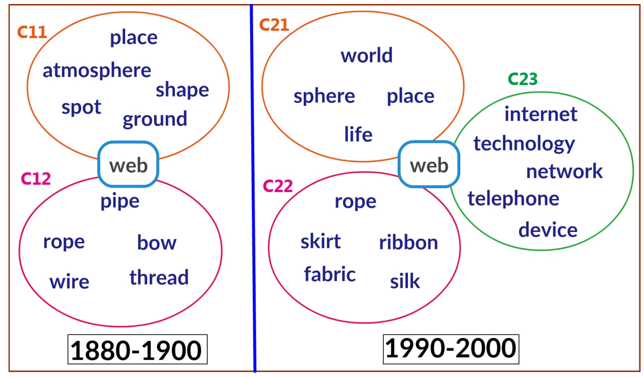

Figure 3: Chinese Whisper clusters for the target word 'web' extracted from COHA corpus for the time periods 1880-1900 and 1990-2000.

Table 16: Example cases from COHA corpus, having linguistic shifts as suggested by prior literature and correctly detected by our approach. The discriminative feature shows the network measure which has changed the most over time.

Word	'birth' cluster	Discriminative feature
virus	weapon, system, aircraft . . .	Δ(SSM)
cell	network, satellite, phone, . . .	Δ(SSM)
monitor	computer, TV, screen, . . .	Δ(ED)
axis	missile, fire, satellite, . . .	Δ(ED)
broadcast	TV, cable, service, . . .	Δ(APL)
check	wage, donation, fee, . . .	Δ(APL)
film	show, concert, script, . . .	Δ(ED)
focus	concern, ambition, . . .	Δ(APL)
major	university, discipline, . . .	Δ(APL)
program	project, database, testing, . . .	Δ(ED)
record	tape, card, disc, copy . . .	Δ(SSM)
web	Web, Internet, network . . .	Δ(ED)

Experiment details and results: We first construct distributional thesauri (DT) networks [22] for the COHA corpus at two different time points, 1880-1900 and 1990-2000. We apply Chinese Whispers algorithm [2] to produce a set of clusters for each target word in the DT network. The Chinese Whispers clusters for the target word 'web' are shown in Figure 3. Note that we have reported only some of the representative words for each cluster. Each of the clusters represents a particular sense of the target. We now compare the sense clusters extracted across two different time points to obtain the suitable signals of sense change following the approach proposed in Mitra *et al.* Mitra *et al.* [19]. After getting the novel sense clusters, we pick up 50 random samples, of which 25 cases are flagged as 'true birth' and the rest 25 cases are flagged as 'false birth' by manual evaluation. We use these 50 samples as our training set for classification using SVM. Some of the examples of this training set are presented in Table 15. We ensure that none of the words in the list L is present in the training set. Using this training set for our proposed SVM classifier, we are successfully able to detect **80%** of the cases (12 out of 15) from the list L as having novel sense. Table 16 presents all of these detected words along with the novel senses and the discriminative network feature. Our method is unable to detect three cases -'gay', 'guy' and 'bush'. For 'gay', since there is no sense cluster in the older time period with 'gay' being a noun, cluster comparison does not even detect the 'birth' cluster of 'gay'. The 'birth' sense clusters for 'guy', 'bush' in the new time period, as detected by split-join algorithm contain general terms like "someone, anyone, men, woman, mother, son" and "cloud, air, sky, sunlight" respectively. As the network around these words did not change much over time, our method found it difficult to detect. Note that even though COHA corpus is substantially smaller than the Google n-grams corpus, our approach produces promising result, showing the usability of the proposed method with not so large corpus as well.

6 CONCLUSION

In this paper, we showed how complex network theory can help improving the performance of otherwise challenging task of novel sense detection. This is the first attempt to apply concepts borrowed from complex network theory to deal with the problem of novel sense detection. We demonstrated how the change in the network properties of the induced subgraphs from a sense cluster can be used to improve the precision of novel sense detection significantly. Manual evaluation over two different time point pairs shows that the proposed SVM classification approach boosts the precision values from 0.23-0.32 to 0.74-0.86. Finally, from the experiments on the COHA corpus, we have also shown that our approach can reliably detect the words known to have sense shifts. We have made the human annotated data used in our experiments publicly available which could be used as gold standard dataset to validate systems built for novel sense detection[6].

This framework of precise novel sense detection of a word can be used by lexicographers as well as historical linguistics to design new dictionaries as well as updating the existing sense repositories like WordNet where candidate new senses can be semi-automatically detected and included, thus greatly reducing the otherwise required manual effort. Computational methods based on large diachronic corpora are considered to have huge potential to put a light on interesting language evolution phenomenon which can be useful for etymologists as well. In future, we plan to apply our methodology to different genres of corpus, like social network data, several product or movie reviews data which are becoming increasingly popular source for opinion tracking, to identify short-term changes in word senses or usages. These analyses would also provide insights on the evolution of language in a short span of time. Apart from that, we plan to extend our work to detect the dying senses of words; the senses which were used in the older texts, but are not being used in newer time anymore. Our ultimate goal is to prepare a generalized framework for accurate detection of sense change across languages and investigate the triggering factors behind language evolution as well.

REFERENCES

[1] David Bamman and Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. 1–10.

[2] Chris Biemann and Antal van den Bosch. 2011. *Structure discovery in natural language*. Springer Science & Business Media.

[3] Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel Word-sense Identification.. In *COLING*. 1624–1635.

[4] Beate Dorow, Dominic Widdows, Katarina Ling, Jean-Pierre Eckmann, Danilo Sergi, and Elisha Moses. 2004. Using curvature and markov clustering in graphs for lexical acquisition and word sense discrimination. *arXiv preprint cond-mat/0403693* (2004).

[6] https://tinyurl.com/ycj6ahud

[5] Steffen Eger and Alexander Mehler. 2016. On the Linearity of Semantic Change: Investigating Meaning Variation via Dynamic Graph Models. In *ACL*. 52–58.

[6] Katrin Erk. 2006. Unknown word sense detection as outlier detection. In *Proceedings of the main conference on Human Language Technology Conference of the NAACL*. 128–135.

[7] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[8] Lea Frermann and Mirella Lapata. 2016. A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics* 4 (2016), 31–45.

[9] Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, Vol. 1. 241–247.

[10] Kristina Gulordava and Marco 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*. 67–71.

[11] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the 2016 Conference on EMNLP*. 2116–2121.

[12] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *ACL*. Berlin, Germany, 1489–1501.

[13] Abhik Jana and Pawan Goyal. 2018. Network Features Based Co-hyponymy Detection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan.

[14] Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. 229–238.

[15] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*. 625–635.

[16] Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of ACL*. 259–270.

[17] Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of EACL*. 591–601.

[18] Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. 259–263.

[19] Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *ACL*. 1020–1029.

[20] Thiago Alexandre Salgueiro Pardo, Lucas Antiqueira, Maria das Graças Volpe Nunes, Osvaldo N Oliveira Jr, and Luciano da Fontoura Costa. 2006. Using complex networks for language processing: The case of summary evaluation. In *Communications, Circuits and Systems Proceedings, 2006 International Conference on*, Vol. 4. 2678–2682.

[21] Christian Ramiro, Mahesh Srinivasan, Barbara C Malt, and Yang Xu. 2018. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences* 115, 10 (2018), 2323–2328.

[22] Martin Riedl and Chris Biemann. 2013. Scaling to Large3 Data: An Efficient and Effective Method to Compute Distributional Thesauri.. In *EMNLP*. 884–890.

[23] Maja Rudolph and David Blei. 2018. Dynamic Embeddings for Language Evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. 1003–1011.

[24] Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 448–453.

[25] Nina Tahmasebi, Thomas Risse, and Stefan Dietze. 2011. Towards automatic language evolution tracking, a study on word sense tracking. In *Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn 2011), co-located with ISWC 2011*.

[26] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association* 101, 476 (2006).

[27] Ivana Turnu, Michele Marchesi, and Roberto Tonelli. 2012. Entropy of the degree distribution and object-oriented software quality. In *Proceedings of the 3rd International Workshop on Emerging Trends in Software Metrics*. 77–82.

[28] Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med* 37, 5 (2005), 360–363.

[29] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press.

[30] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 673–681.