Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View

Renfen Hu [1,2,♠] Shen Li [3,♣] Shichen Liang [1,2,♠]

♠ {irishu, shichen}@mail.bnu.edu.cn
♣ shen@deeplycurious.ai

[1] Institute of Chinese Information Processing, Beijing Normal University
[2] UltraPower-BNU Joint Laboratory for Artificial Intelligence, Beijing Normal University
[3] DeeplyCurious.ai

Abstract

Diachronic word embeddings have been widely used in detecting temporal changes. However, existing methods face the meaning conflation deficiency by representing a word as a single vector at each time period. To address this issue, this paper proposes a sense representation and tracking framework based on deep contextualized embeddings, aiming at answering not only what and when, but also how the word meaning changes. The experiments show that our framework is effective in representing fine-grained word senses, and it brings a significant improvement in word change detection task. Furthermore, we model the word change from an ecological viewpoint, and sketch two interesting sense behaviors in the process of language evolution, i.e. sense competition and sense cooperation.

1 Introduction

The meanings of words continuously change over time, reflecting complicated processes in language and society (Kutuzov et al., 2018). With the rapid development of language representation learning, word embeddings have been widely introduced into diachronic linguistic studies. By training and comparing word embeddings of different time epochs, one can capture the semantic drift of words (Kim et al., 2014), learn diachronic analogies between terms (Szymanski, 2017), as well as discover the statistical laws of semantic change (Hamilton et al., 2016). Furthermore, this kind of method has gained fruitful results in broader social science studies, e.g. tracing armed conflicts (Kutuzov et al., 2017), gender and ethnic stereotypes (Garg et al., 2018) and social attitudes (Jaidka et al., 2018).

It is well known that word meaning can be represented with a range of senses. However, existing methods only assign one embedding to a word for a time period, thus they face challenges in representing senses and tracking the change of them. Given the word embeddings, one can tell the coarse-grained change of the word from one time to another, e.g. the word *gay*'s nearest neighbors in the vector space move from *cheerful* and *flaunting* to *homosexual* and *lesbian*. But these word representations are not able to show which sense has changed, which sense is stable, and how they may interact with each other.

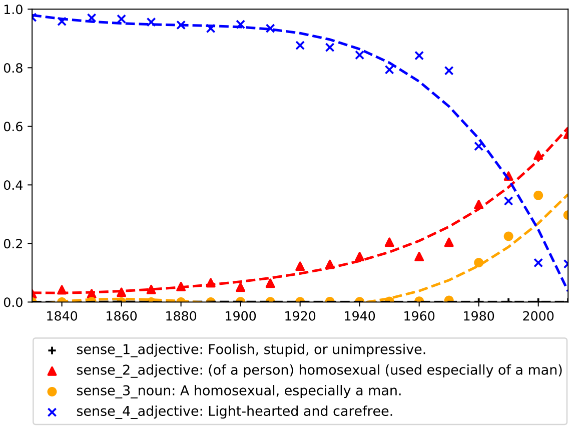


Figure 1: The evolvement of four senses for word *gay*. Two important phenomena: (1) competition between sense_2 and sense_4; (2) cooperation between sense_2 and sense_3.

Recently, an increasing boom on large-scale pre-trained language models e.g. ELMo and BERT have attracted considerable attention in the field of NLP (Peters et al., 2018; Devlin et al., 2018). These models can ideally capture complex characteristics of word use, and how they vary across linguistic contexts, i.e. a word with different contexts can yield different representations.

Inspired by the above works, this paper proposes to use deep contextualized embeddings to represent and track word senses. Figure 1 shows that our method can trace the fine-grained senses of a word in a smooth process, i.e. change does not

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3899–3908
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

happen at a time point, but continuously throughout the process. We further model the evolvement from an ecological viewpoint, and propose that senses can compete and cooperate just like groups of organisms. The contribution of this paper is as following:

- We construct an efficient sense representation method using the pre-training language model BERT and data from Oxford dictionary. This method can precisely learn and identify fine-grained senses, and achieves a high accuracy of 93.8% in a sense identification task.
- Based on the sense representation, we detect in depth the trend of word senses in 200 years of texts. In evaluation, our method brings a significant improvement on word meaning change task.
- Interestingly, we further model the word change from an ecological viewpoint, and introduce two important sense behaviors in language evolution, i.e. sense competition and sense cooperation.

The remaining part of this paper is organized as following. After introducing the related work in Section 2, we will describe our sense representation model and how to track senses in 200 years in Section 3. In Section 4, we analyze the sense behaviors from an ecological viewpoint, and sketch two interesting phenomena: sense competition and cooperation. At last, we draw conclusions and propose future work in Section 5.

2 Related Work

2.1 Diachronic Word Embeddings

Neural word embeddings have been widely used in diachronic linguistic studies. The basic idea is to train word embeddings on different time-sliced corpora and then compare them over time. Kim et al. (2014) firstly use neural embeddings to capture the change of word meaning. Their method initializes the vectors with the data of the previous year. Kulkarni et al. (2015) and Hamilton et al. (2016) train the embeddings independently and then use a mapping method to align them for comparison. Bamler and Mandt (2017) propose to use dynamic word embeddings trained jointly over all times periods. Instead of modeling lexical change via time series, Rosenfeld and Erk (2018) represent time as a continuous variable and model a word's usage as a function of time. Yin et al. (2018) propose global anchor method for detecting linguistic shifts and domain adaptation.

However, the above methods could only assign one neural embedding to a word at each time period, which cannot model the change of the word senses. To address this problem, we propose to conduct a sense-level diachronic study with deep contextualized word embeddings, and detect in depth not only what and when, but also how the word meaning changes.

2.2 Diachronic Sense Modeling

Existing works on sense modeling mainly exploit topic modeling and clustering methods. Lau et al. (2012) and Cook et al. (2014) propose to detect novel senses by comparing a reference corpus and a focus corpus with topic modeling. Wijaya and Yeniterzi (2011) firstly try to track word senses with K-means clustering and the Topic-Over-Time algorithm. Mitra et al. (2014) identify the sense birth, death, join and split based on clustering of a co-occurrence graph. Frermann and Lapata (2016) present a dynamic Bayesian model to track the prevalence of senses, and further model language change as a smooth, gradual process. Tang et al. (2016) attempted to cluster the contexts to find senses, and to classify the senses into different change types. Tahmasebi and Risse (2017) exploit curvature clustering algorithm to induce word senses and track the change of them.

Although these studies have made great progress in novel sense detection and diachronic sense tracking, they may have two disadvantages in sense modeling: (1) It is arbitrary and difficult to select the number k of the clusters or topics, and there are few works explaining the reason of the setting. (2) The "senses" induced from clusters or topics require huge amount of human analysis to interpret or additional mappings to an external sense inventory. Thus, the discussion is usually limited to a few cases.

2.3 Learning Sense and Contextual Embeddings

Pilehvar and Collier (2016); Camacho-Collados and Pilehvar (2018) address the meaning conflation deficiency of existing methods representing a word as a single vector, as it may have negative impacts on accurate semantic modeling. For example, *rat* and *screen* are pulled towards each other in the vector space for their similarities to two different senses of *mouse*.

To solve this problem, there are a line of works making extensions of the Skip-gram model to learn sense-specific embeddings (Neelakantan et al., 2014; Liu et al., 2015; Qiu et al., 2016; Lee and Chen, 2017). In addition, knowledge bases e.g. Wordnet are introduced into representation (Chen et al., 2014, 2015; Faruqui et al., 2014; Johansson and Pina, 2015; Rothe and Schütze, 2015).

Recently, it has attracted considerable attention by constructing unsupervised contextual representations with language models. Melamud et al. (2016) represent the context of a target word with the output embedding of a multi-layer perceptron built on top of a Bi-LSTM language model. Peters et al. (2018) show that their language model ELMo can implicitly disambiguate word meaning with their contexts. Devlin et al. (2018) propose bidirectional encoder representations from Transformers (BERT). It is fine-tuned with just one additional output layer, and achieves state-of-the-art results for a wide range of tasks. In this study, we propose to learn sense representations following Devlin et al. (2018)'s work since it can yield deep and effective contextual representations on both sentence and token level.

3 The Framework

3.1 Sense Representation

In this paper, we build fine-grained sense representations with deep contextualized word embeddings, i.e. represent each sense as a distinguished sense embedding. We directly adopt the fine-grained senses defined by lexicographers. Comparing with existing diachronic sense studies, our method does not rely on human interpretations or mappings to dictionary definitions.

For a sense s_j of word w_i, we can obtain its example sentences $\{Sent_1^{w_i s_j}, Sent_2^{w_i s_j}, ..., Sent_n^{w_i s_j}\}$ from a dictionary. After feeding them into a pre-trained language model, w_i's token representations $\{e_1^{w_i s_j}, e_2^{w_i s_j}, ..., e_n^{w_i s_j}\}$ can be retrieved from the final hidden layer of the model. The sense embedding $e^{w_i s_j}$ of s_j is computed by taking the average of $\{e_1^{w_i s_j}, e_2^{w_i s_j}, ..., e_n^{w_i s_j}\}$.

In the experiments, we choose the Oxford English dictionary since it has (1) a comprehensive record of word senses in different times and (2) a sufficient number of example sentences for each sense.

To select the target words for diachronic study, we firstly extract word frequency information from COHA, a genre balanced corpus containing English texts from 1810 to 2009[1]. Only words that appear at least 10 times a year for over 50 consecutive years are retained. After lemmatization, we totally retrieve 4881 words, including 15836 senses in Oxford dictionary. The sense definitions and example sentences are then extracted from the online version of Oxford dictionary[2].

We feed at most 10 sentences for each sense to the pre-trained BERT model (Devlin et al., 2018) as the inputs. We use the uncased Bert-Base model that has 12 layers, 768 hidden units, 12 heads and 110M parameters. The language model is trained on BookCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words) with Masked LM and Next Sentence Prediction tasks. With deep bidirectional architecture, BERT yields powerful language representations on both sentence and token level.

After feeding the sentences containing a target word with a specific sense, its token representations can be generated from the hidden layers of the pre-trained model. We only keep the token representations of the final hidden layer of the Transformer. After obtaining the token embeddings of the target word for the specific sense, we can represent the sense as a 768-dimentional embedding by averaging the token embeddings.

3.2 Sense Identification

After obtaining the sense representations of the target words, we can easily identify the sense of a word in a sentence with its contextual embedding. Given a new sentence $Sent_k$ that contains a target word w_i with m senses, we can feed it into BERT to get w_i's contextual embedding $e_k^{w_i}$, and compute the cosine similarities between the token embedding $e_k^{w_i}$ and the word sense embeddings $\{e^{w_i s_1}, e^{w_i s_2}, ..., e^{w_i s_m}\}$. The sense $s_{\hat{j}}$ that has the highest similarity score is selected as the belonging sense.

$$s_{\hat{j}} = \arg\max_{s_j} \frac{e^{w_i s_j} \cdot e_k^{w_i}}{\|e^{w_i s_j}\|_2 \, \|e_k^{w_i}\|_2} \quad (1)$$

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

[2] https://en.oxforddictionaries.com/

Sentences with the target word	Most similar sense
1. You'll be satisfied with less food, which **means** you'll consume fewer calories each time you sit down to eat.	*v.* Have as a consequence or result.
2. Anna wanted to know exactly what he **meant**, but she did not ask.	*v.* Intend to convey or refer to; signify.
3. The **mean** score for this question is 55.0 for those who did not receive bills from physicians and labs.	*n.* calculated as a mean; average.
4. They were n't necessarily fighting or being **mean** to each other constantly.	*a.* Unkind, spiteful, or unfair.
5. Do not bring thine eye to their small, **mean**, and plodding lives...	*a.* poor in quality and appearance; shabby.
6. This man is a **mean** motor scooter on the mound.	*a.* Very skillful or effective; excellent.
7. I left for work before the kid **crawled** out of bed.	*v.* Move forward on the hands and knees.
8. This beta search site **crawls** the web for product-related information, including data from the product maker, magazine articles.	*v.* systematically visit a number of web pages in order to create an index of data.

Table 1: Sense identification for word *mean* and *crawl*. The model performs well in detecting dated sense (*sent5*), infrequent sense (*sent6*), and new sense (*sent8*).

Table 1 gives several sentences that contain polysemous words *mean* or *crawl*. With our method, the senses can be precisely captured, even when the word is used in a dated sense e.g. *poor in quality*[3], an infrequent sense e.g. *skillful and excellent*, or a new sense as seen in sentence 8. It shows that our method based on contextual embeddings and Oxford dictionary is able to capture the word senses of different periods and frequencies effectively.

3.3 Sense Tracking

To track the sense evolvement, we use the 200 years of texts from COHA corpus. After preprocessing and POS tagging, we feed the sentences to BERT, and retrieve the token embedding if the lemmatized token[4] is one of the 4881 target words. Using the sense representations built via the above method, we can easily tag the sense for each token. Tang et al. (2016) suggest that a time series of word status data can be decomposed into a trend component and a random noise. We follow this idea to model the time series of sense status.

Given a word w_i that has senses $\{s_1, s_2, ..., s_m\}$, the diachronic status of sense s_j is represented by

$$T(s_j) = \{P_{t_1}^{s_j}, P_{t_2}^{s_j}, ..., P_{t_y}^{s_j}\}, \tag{2}$$

where $P_t^{s_j}$ is defined as

$$P_t^{s_j} = \frac{N_t^{s_j}}{\sum\limits_{k=1}^{m} N_t^{s_k}}, \tag{3}$$

where $N_t^{s_j}$ is the number of tokens identified as sense s_j at time t.

According to (Brockwell et al., 2002), $T(s_j)$ can be decomposed as

$$T(s_j) = Tr(s_j) + Noise(s_j), \tag{4}$$

where $Tr(s_j)$ is the trend and $Noise(s_j)$ is a random noise.

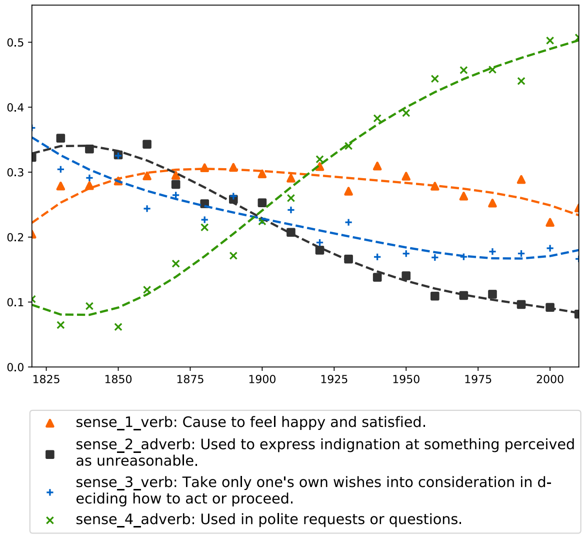


Figure 2: The evolvement of word *please*.

We conduct quartic polynomial curve fitting on account of the fluctuation. The noise $Noise(s_j)$ is thus striped, and we can get the trend $Tr(s_j)$ for further analysis. We set the time interval $\Delta t = 10$ since it has a robust performance in curve fitting.

With this method, we can clearly monitor the status of each individual sense, whether it is growing, decreasing, or unchanged. Figure 2 shows the fitting result of *please* which receives few attention from previous diachronic studies. It can be seen that sense_2 that expresses *indignation and unreasonable* is going down, sense_1 and sense_3 that use in verb meaning are relatively stable, and sense_4 that used in *polite requests or questions* has been consistently growing.

3.4 Evaluation of The Framework

To evaluate the sense representation and tracking methods, we conduct experiments on two tasks: (1) a synchronic sense identification task, and (2)

[3] This sense is labeled as "dated" in Oxford dictionary.

[4] We use the NLTK WordNet Lemmatizer.

a diachronic word meaning change task.

3.4.1 Word Sense Identification

To test the sense representation, we construct a dataset by randomly selecting another 2000 sentences from Oxford dictionary that have not been used in sense representation. Each test sentence contains at least a polysemous target word. Given the senses of the target word as candidates, the model needs to select the correct sense for the word in the sentence.

Considering the part of speech (POS) is a useful feature for word sense disambiguation tasks, we firstly do POS tagging for the sentences with NLTK. In the test, if the POS information is used, the model will limit the candidates to the senses with the same POS. Otherwise the model considers all the senses of this word being candidates. The test result is shown in Table 2.

System	Accuracy
Baseline	92.3%
Baseline + POS	93.8%

Table 2: Results of word sense identification task.

We can see that POS information does improve the accuracy. We further analyze the 124 bad cases of Baseline + POS system, and some examples are shown in Table 3.

Firstly, we find that in some cases the model predictions are not real mistakes. (1) the model prediction seems to be a better option in 16 cases, e.g. the sentence 1 in Table 3. (2) Given the context, the model prediction and the answer can be both reasonable for 3 cases, e.g. the sentence 2.

Secondly, for the remaining 105 real bad cases, the mistakes are mostly due to the following reasons. (1) The model prediction is a highly similar sense with the answer, or there is a meaning overlap between the two senses, e.g. sentence 3 in Table 3. (2) The model does not get a precise contextualized embedding from BERT since the text is short and can not provide sufficient information, e.g. sentence 4. It should be noted that in this case, the model also has a low confidence given the highest cosine similarity as 0.25.

We also find that the similarity scores indicating the model confidence have high correlations with the accuracy. Given the 902 cases that have similarities $\geq$ 0.8, the model accuracy increases to 98%. For the 44 cases with similarities $\geq$ 0.9, the accuracy is 100%. The experiment shows that it is very effective to use deep contextualized embeddings to represent word sense. With very few data (10 or less sentences for a sense), it yields reliable and precise sense representations. Using a very simple similarity measurement, the method achieves a high accuracy in the sense identification task. We believe it could serve as a good basis for the diachronic sense studies.

3.4.2 Word Meaning Change

For evaluation on the diachronic side, we conduct experiments on word meaning change task with the human rating dataset proposed by Gulordava and Baroni (2011).

The test set consists of 100 words taken from different frequency range. Five annotators are asked to label the change of each word from 1960s to 1990s on a 4-point scale (0: no change; 1: almost no change; 2: somewhat change; 3: changed significantly). The inter-annotator agreement is 0.51 (pairwise Pearson correlation, $p < 0.01$). We follow Frermann and Lapata (2016)'s work to quantify the word change via the novelty score defined by Cook et al. (2014).

Given a word w_i with m senses, the novelty score is calculated by

$$N(s_j) = \frac{p_f(s_j) + \alpha}{p_r(s_j) + \alpha}, \quad (5)$$

where s_j is one of the senses, $p_f(s_j)$ is the proportion of usages of s_j in the focus corpus, $p_r(s_j)$ is the proportion of usages of s_j in the reference corpus and α is a small parameter to avoid dividing by zero.

Further, we can calculate the score of word change w_i by

$$C(w_i) = \max\{N(s_1), N(s_2), ..., N(s_m)\}. \quad (6)$$

In the test, we select the data of 1960s from COHA as the reference corpus, and data of 1990s as the focus corpus. α is set to 0.01. After computing the novelty score for each word, we measure the correlation coefficient between the novelty scores and the average human ratings.

As shown in Table 4, the Pearson correlation score of our method is 0.52 ($p < 0.01$), and Spearman's ρ rank is 0.428 ($p < 0.01$), which achieve a significant improvement comparing with the existing studies. The test result further proves the effectiveness of our sense modeling method built on deep contextualized embeddings.

Sentence	Answer	Predict	Simi
1. Again, you'd expect that the most "important" words in a document, in terms of identifying what it's about, would be the ones most individually **freighted** with meaning.	Transport (goods) in bulk by truck, train, ship, or aircraft.	Be laden or burdened with.	0.89
2. He said a car had **just** managed to squeeze past the people carrier, and he had tried to do the same but in vain.	Barely; by a little.	Very recently; in the immediate past.	0.76
3. The move to establish the Pratas marine sanctuary must not be **separated** from the international movement to protect marine areas.	Divide into constituent or distinct elements.	Cause to move or be apart.	0.68
4. he paused **significantly**.	In a way that has a particular meaning.	In a sufficiently great or important way.	0.25

Table 3: Examples of bad cases in word sense identification task.

System	Corpus	Pearson	Spearman
Gulordava (2011)	Google Bigram	0.386	-
Fremann (2016)	COHA, DTE, CLMET3.0	-	0.377
Our method	COHA	0.52	0.428

Table 4: Results of word change task.

4 An Ecological View

Ecologists are interested in the dynamics of species populations over time (Odum and Barrett, 1971), while linguists focus on the language change. These two systems may share some commonalities, e.g. Nadas (1985) applied the Turing Formula (Good, 1953) which studies the population frequencies of species to word probabilities. In this study, after tracking the prevalence of word senses in 200 years, we find that senses can compete and cooperate just like ecological organisms. Of course, these behaviors are primarily determined by people who use it, learn it and transmit it to others (Haugen, 1971, 2001).

4.1 Sense Competition

A word is like an ecological population, and different senses are its subgroups. "Competition" exists between the senses. They do not compete for sunlight or food, but the dominance of the word. We can observe the **semantic** and **grammatical** change of words from the perspective of "competition".

Intuitively, word meaning changes gradually, and a significant change may take place at a time period when a dominant sense handing over to another one, usually referring to a semantic shift (Kulkarni et al., 2015). When the new dominant sense has different grammatical features, e.g. a different part-of-speech, we can observe a grammatical change. Thus, the sense competition for dominance may result in semantic and grammatical changes.

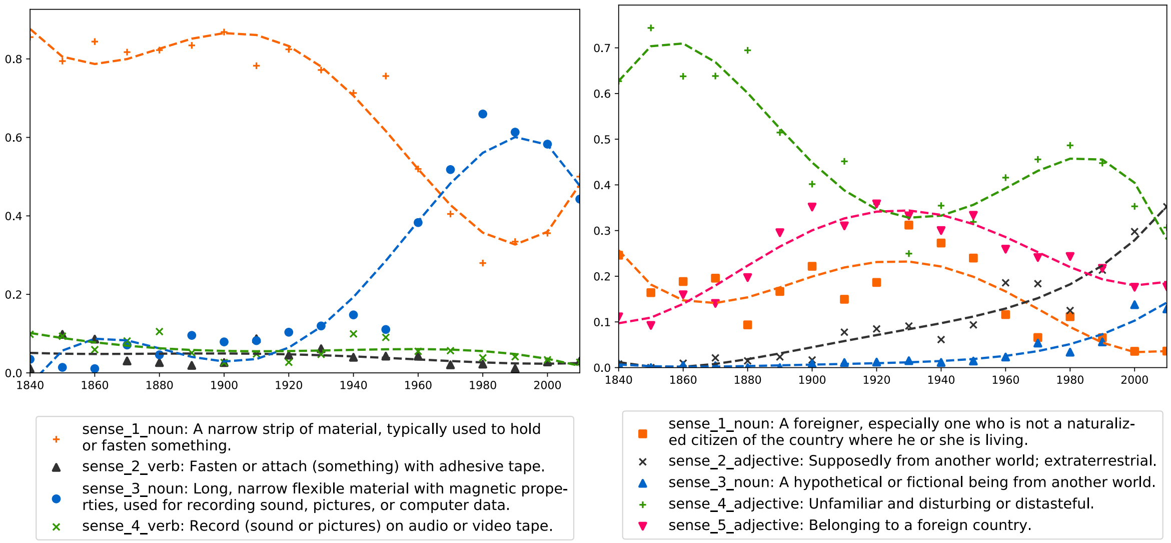
Figure 1 shows an example of semantic change for word *gay*. The adjective meaning of *homosexual* grows quickly in 20th century, and finally took the place of *light-hearted* to be the most dominant sense at the end of 1990s. Figure 2 illustrates both grammatical and semantic changes of word *please*, which is more and more frequently used as an adverb (*in polite requests or questions*) than as a verb.

Interestingly, the competition is not a monotonous process. As shown in Figure 3a, the *magnetic recording material* meaning of *tape* has a strong growth during 1920-1980, but degrades soon since 1990 because this material become dated in daily life. Then the dominant sense goes back to the *material for fastening things*.

In order to capture the trend of language evolvement, we track the senses of 3220 polysemous words with time interval $\Delta t = 10$. The tracking is based on polynomial curve fitting result. If the dominant sense changes from one to another, we count it as a word change. If the new dominant sense has a different part-of-speech from the old one, we count it as a grammatical change, otherwise a semantic change[5].

Among the 3220 words, 70.12% have a stable dominant sense, whereas 29.88% undergo a change of dominant sense for at least once, resulting in 1064 detected changes in which 69.26% are semantic changes, and 30.73% are grammatical changes. It indicates that the language system is mostly stable, and semantic change occupies

[5] It should be noted that the "semantic change" denoted here refers to a change of the semantic meaning, while the "grammatical change" may involve both changes of the POS and semantic meaning, e.g. the dominant sense of *please* changes from sense_1_verb (*cause to feel happy and satisfied*) to sense_4_adverb (*polite requests or questions*).



(a) *tape*

(b) *alien*

Figure 3: Examples of sense competition and cooperation. (a) *tape*: competition between sense_1 and sense_3; (b) *alien*: cooperation between sense_1 and sense_5, sense_2 and sense_3.

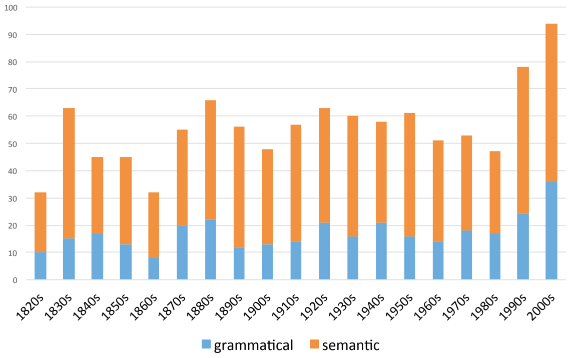


Figure 4: The counts of grammatical and semantic changes from 1820s to 2000s.

a larger proportion. From a diachronic perspective, Figure 4 shows that the counts of detected word changes are similarly distributed across the decades, while in 1990s and 2000s, senses are more active in competition.

4.2 Sense Cooperation

In addition to competition for selfish benefit, a group of organisms can also work together for common or mutual benefit in the evolution. Hamilton (1964) proposes that cooperation helps in transmitting underlying genes to future generations either for direct fitness (increasing personal reproductive successes) or for indirect fitness (increasing the reproductive successes of genetically similar relatives). In this study, we also find that similar senses are prone to cooperate to survive and compete with others.

Figure 1 gives us an intuitive example for word *gay*. The adjective sense *homosexual* has a relative: a noun sense of *homosexual man*. These two senses are not only very related in meaning, but also have highly consistent growth curve. In the competition, they cooperate to overtake sense_2 (*light-hearted and carefree*).

Based on the above analysis, we attempt to detect the cooperating senses automatically. We hypothesize that the cooperating senses should satisfy two conditions. Firstly, these senses should be similar or related in meaning. Secondly, they should grow or degrade in a similar trend. Starting from this hypothesis, we model the meaning similarity r with their sense embeddings, and the trend similarity c with Pearson correlation coefficient. In the case of *gay*, sense_2 and 3 are identified as relative senses which are cooperating in the competition because they have a high $r = 0.9565$ and $c = 0.8995$.

With the thresholds setting $r \geq 0.6$ and $c \geq 0.6$, we detect 490 pairs of relative senses that cooperate and also win in the competition against other senses, accounting for 31.67% of the changes. Table 5 lists the 10 words that has the highest mean value of r and c. It can be seen that the relative senses are highly similar in meaning or usages, and can be considered as a sense family.

We illustrate the cooperation between the senses and its role in language evolvement with an example word *alien*. As shown in Figure 3b, *alien* was

word	old dominant sense	new dominant sense	relative sense	r	c
lot (1890s)	*n.* A person's luck, situation, or destiny in life.	*pron.* A large number or amount; a great deal.	*d.* A great deal; much.	0.98	0.91
decline (1940s)	*v.* Politely refuse (an invitation or offer)	*n.* A gradual and continuous loss of strength, numbers, quality, or value.	*v.* (typically of something regarded as good) become smaller, fewer, or less; decrease.	0.99	0.88
alien (2000s)	*a.* Unfamiliar and disturbing or distasteful.	*a.* Supposedly from another world; extraterrestrial.	*n.* A hypothetical or fictional being from another world.	0.96	0.91
fancy (1940s)	*n.* A superficial or transient feeling of liking or attraction.	*a.* Elaborate in structure or decoration.	*a.* (of a drawing, painting, or sculpture) created from the imagination rather than from life.	0.94	0.92
review (1960s)	*v.* Write a critical appraisal of (a book, play, film, etc.) for publication in a newspaper or magazine.	*n.* A formal assessment of something with the intention of instituting change if necessary.	*v.* Assess (something) formally with the intention of instituting change if necessary.	0.98	0.88
gay (1990s)	*a.* Light-hearted and carefree.	*a.* (of a person) homosexual (used especially of a man)	*n.* A homosexual, especially a man.	0.96	0.90
desert (1940s)	*v.* Abandon (a person, cause, or organization) in a way considered disloyal or treacherous.	*n.* A waterless, desolate area of land with little or no vegetation, typically one covered with sand.	*a.* Like a desert.	0.96	0.90
exercise (1970s)	*v.* Use or apply (a faculty, right, or process)	*n.* Activity requiring physical effort, carried out to sustain or improve health and fitness.	*v.* Engage in physical activity to sustain or improve health and fitness.	0.98	0.88
abroad (1910s)	*d.* In different directions; over a wide area.	*d.* In or to a foreign country or countries.	*n.* Foreign countries considered collectively.	0.94	0.91
hit (1910s)	*v.* Reach (a particular level, point, or figure)	*v.* Bring one's hand or a tool or weapon into contact with (someone or something) quickly and forcefully.	*n.* An instance of striking or being struck.	0.99	0.86

Table 5: Examples of the cooperating senses that win in the competition.

mainly used as an adjective of *unfamiliar* meaning until the beginning of 20th century. After that, there are two groups of cooperation captured:

- With the increasing global communication at the end of the 19th century, sense_1 and sense_5 constituted a powerful family, in which one sense represents the noun meaning (*foreigner*), and the other one denotes the adjective (*belonging to a foreign country*).
- Since 1950s, with the exploration in the space, *alien* is used to refer to *extraterrestrial* and *hypothetical beings from another world*, i.e. the sense_2 and sense_3 which form a new sense family. They finally achieve the dominance of the word meaning via their cooperation.

It should be noted that just like groups of organisms, the cooperation does not only exist in growing senses, but also in stable and degrading senses. In addition, the competition can also take place between two relative senses, e.g. the dominant sense of word *heavily* changed from *with a lot of force or effort; with weight* to a more abstract meaning *to a great degree; in large amounts* in 1920s.

5 Conclusion and Future Work

This paper proposes a sense representation and tracking framework based on deep contextualized embeddings. With our method, we can find out not only what and when, but also how the word meaning changes from a fine-grained sense level. The experiment shows that our framework is effective in representing word senses and detecting word change. Furthermore, we model the word change from an ecological viewpoint, and sketch two interesting sense behaviors in language evolution, i.e. sense competition and sense cooperation.

Overall, our study sheds some light on diachronic language study with deep contextualized embeddings. The sense modeling data we built may serve as a basis for further and deeper analysis of linguistic regularities, as well as an important reference of sense granularities for lexicographers[6].

In addition to tracking the language evolvement in the history, we believe it is promising future work to use deep contextual embeddings in pre-

[6] We release the sense modeling data and a visualization tool at https://github.com/iris2hu/diachronic-sense-modeling.

dicting the future change or trend, as well as detecting novel senses that are not included in existing dictionaries.

Acknowledgments

The authors would like to thank Zhe Zhao for his helpful comments and discussions, Kristina Gulordava for sharing the dataset of word meaning change, and the anonymous reviewers for their feedback and suggestions. This work is supported by the Fundamental Research Funds for the Central Universities, National Social Science Fund of China (No. 18CYY029) and China Postdoctoral Science Foundation funded project (No. 2018M630095).

References

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR. org.

Peter J Brockwell, Richard A Davis, and Matthew V Calder. 2002. *Introduction to time series and forecasting*, volume 2. Springer.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2015. Improving distributed representation of word sense via wordnet gloss composition and context clustering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 15–20.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Irving J Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

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*, pages 67–71.

William D Hamilton. 1964. The genetical evolution of social behaviour. ii. *Journal of theoretical biology*, 7(1):17–52.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1489–1501.

Einar Haugen. 1971. The ecology of language. *Linguistic Reporter*.

Einar Haugen. 2001. The ecology of language. *The ecolinguistics reader: Language, ecology and environment*, pages 57–66.

Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. 2018. Diachronic degradation of language models: Insights from social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 195–200.

Richard Johansson and Luis Nieto Pina. 2015. Embedding a semantic network in a word space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1428–1433.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65.

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*, pages 625–635. International World Wide Web Conferences Steering Committee.

Andrei Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017. Tracing armed conflicts with diachronic word embedding models. Association for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.

Guang-He Lee and Yun-Nung Chen. 2017. Muse: Modularizing unsupervised sense embeddings. *arXiv preprint arXiv:1704.04601*.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

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 *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1020–1029.

Arthur Nadas. 1985. On turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1414–1416.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient nonparametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069.

Eugene Pleasants Odum and Gary W Barrett. 1971. *Fundamentals of ecology*, volume 3. Saunders Philadelphia.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

Mohammad Taher Pilehvar and Nigel Collier. 2016. De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1680–1690.

Lin Qiu, Kewei Tu, and Yong Yu. 2016. Context-dependent sense embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 183–191.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 474–484.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1793–1803.

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)*, volume 2, pages 448–453.

Nina Tahmasebi and Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *RANLP*, pages 741–749.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2016. Semantic change computation: A successive approach. *World Wide Web*, 19(3):375–415.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40. ACM.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. 2018. The global anchor method for quantifying linguistic shifts and domain adaptation. In *Advances in Neural Information Processing Systems*, pages 9434–9445.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*.