Short-term meaning shift: an exploratory distributional analysis

Anonymous NAACL submission

Abstract

We investigate diachronic meaning shift that takes place in short periods of time (short-term meaning shift) and in an online community of speakers. We create a small dataset and use it to assess the performance of a standard model for meaning shift detection on short-term meaning shift, and find that this phenomenon poses specific difficulties for models based on the Distributional Hypothesis.

1 Introduction

Semantic change has received increasing attention in empirical Computational Linguistics / NLP in the last few years (Tang, 2018). Almost all studies so far have focused on meaning shift in long periods of time –decades to centuries. However, the genesis of meaning shift and the mechanisms that produce it operate at much shorter time spans, ranging from the online agreement on words meaning in dyadic interactions (Brennan and Clark, 1996) to the rapid spread of new meanings in relatively small communities of people (Del Tredici and Fernández, 2017). In this paper we focus on this latter phenomenon, that we call *short-term meaning shift*.

Short-term shift is usually hard to observe in standard language, such as the language of books or news, which has been the focus of long-term studies (Hamilton et al., 2016; Kulkarni et al., 2015), since it takes a long time for a new meaning to be widely accepted in the standard language. We therefore focus on the language produced in an online community of speakers, in which the adoption of new meanings happens at a much faster pace (Clark, 1996; Hasan, 2009).

We analyze the behavior of a standard distributional model of semantic change when applied to short-term shift, also creating a small dataset for this purpose.[1] Distributional models of semantic change are based on the hypothesis that a change in context of use mirrors a change in meaning. Our results show that this type of model successfully detects most meaning shifts, but that it overgeneralizes, since some contextual changes do not correspond to a meaning shift. We also show that this is a difficulty caused by the nature of short-term meaning shift, and propose to use contextual variability as a means to remedy it.

2 Related Work

Several methods have been proposed to investigate long-term meaning shift: common to all of them is the computation of time-related distributional representations for words in the vocabulary, and the sequential comparison of such representations in order to detect a drop in self-similarity, usually interpreted as a shift in meaning.

Among the most widely used techniques are Latent Semantic Analysis (Sagi et al., 2011; Jatowt and Duh, 2014), Topic Modeling (Wijaya and Yeniterzi, 2011; Rohrdantz et al., 2011), and simple co-occurence matrices of target words and context terms (Gulordava and Baroni, 2011; Xu and Kemp, 2015). More recently, researchers have used word embeddings computed using the skip-gram model by Mikolov et al. (2013). Since embeddings computed in different semantic spaces are not directly comparable, time related representation are usually made comparable either by aligning different semantic spaces (Kulkarni et al., 2015; Azarbonyad et al., 2017; Hamilton et al., 2016) or by initializing the embeddings at t+1 using those computed at t (Kim et al., 2014; Del Tredici et al., 2016; Phillips et al., 2017; Szymanski, 2017). We adopt the latter methodology.

Evaluation of semantic shift is difficult, due of

[1] Data and code will be made available upon publication.

sample	time bin	million tokens
Reddit$_{13}$	2013	~900
LiverpoolFC$_{13}$	2011-13	8.5
LiverpoolFC$_{17}$	2017	11.9

Table 1: Time bin and size of the datasets.

the lack of annotated datasets (Fremmann and Lapata, 2016). For this reason, evaluation is usually performed by manually inspecting the n words whose representation changes the most according to the model (Hamilton et al., 2016; Del Tredici et al., 2016; Kim et al., 2014). In this work, we introduce and make available a small dataset for short-term meaning shift, which allows for a more systematic evaluation and analysis and enables comparison in future studies.

3 Data

For this study, we focus on an online forum of football fans, namely the r/LiverpoolFC subreddit, one of the many communities hosted by Reddit.[2] A community such as that of football fans tends to be very active and interconnected, and hence is a good environment for observing linguistic innovations like meaning shift (Hamilton et al., 2017).

We observe meaning shift in the period between 2013 and 2017. Following the typical approach in diachronic studies in NLP, we split the r/LiverpoolFC subreddit data into time bins. In order to enable a clear observation of short-term meaning shift, we define two, non-consecutive time bins: the first (t_1) contains data from 2011-2013 and the second (t_2) from 2017.[3] We also use a large sample of community-independent language for the initialization of the word vectors, namely, a random crawl from Reddit in 2013. For all three samples, we downloaded both textual content and time stamp of the posts (Table 2 shows sample sizes).[4]

Dataset. For evaluation and analysis, we create a dataset with positive and negative meaning shift examples based on the linguistic material produced in r/LiverpoolFC.

In order to create the dataset starting from the raw text downloaded from the subreddit, we initially leverage information about increase in frequency, which has been shown to positively correlate with meaning change (Wijaya and Yeniterzi, 2011; Kulkarni et al., 2015), and sample words with a significant increase in frequency between t_1 and t_2 (an increase in relative frequency is considered significant if it is at least two standard deviations above the mean).[5] Frequency increase is not a necessary condition for meaning shift to take place; however, given the positive correlation mentioned above, it is a reasonable start, since a random selection of words would contain very few positive examples.

This procedure yields ~200 words. The first author of the paper went through the list of words to identify cases of meaning shift, based on the analysis of the contexts of use in the r/LiverpoolFC corpus. We then collected a set of positive and negative examples (100 words in total), which includes 1) words that present a significant increase in frequency and were annotated as meaning shift by the first author (34 words), 2) words that present a significant increase in frequency but that were not annotated as meaning shift by the first author (33 words), and 3) words that keep a constant frequency between t_1 and t_2, and are not considered as examples of meaning shift by us (33 words).[6]

We then created the final dataset via an online survey, which we posted in the r/LiverpoolFC subreddit to recruit participants.[7] We provided participants with the 100 words together with randomly chosen examples from each time period (1 to 5 examples depending on the word). They were asked to label as many of the words as they wanted as 'shift' or 'no shift'. The order of presentation was randomized for each participant.

Overall, 26 members of r/LiverpoolFC participated in the survey, and each word in the dataset received between 5 and 12 judgements (average=8.8). The interannotator agreement, computed as Krippendorff's alpha, is 0.58 –a moderate level of agreement, as is common for semantic tasks (Artstein and Poesio, 2008). We assigned

[2] https://www.reddit.com.

[3] These choices ensure that the two datasets are approximately of the same size. The r/LiverpoolFC subreddit exists since 2009, but very little content was produced in 2009-2010.

[4] We used the Python package Praw, https://pypi.python.org/pypi/praw.

[5] We consider content words only, which we identify by using the external list of common words available at https://www.wordfrequency.info/free.asp

[6] Words in the three lists have overall absolute frequency included in range [50-500].

[7] Domain knowledge is needed for this task.

sample	time bin	million tokens
Reddit$_{13}$	2013	~900
LiverpoolFC$_{13}$	2011–13	8.5
LiverpoolFC$_{17}$	2017	11.9

Table 2: Time bin and size of the datasets.

a final label to each word in the dataset according to the majority of votes received by the redditors. This resulted in a final dataset including 21 cases of meaning shift and 76 of no shift (total 97 words).[8]

4 Experimental Setup

Data. We exploit user-generated language from an online forum of football fans, namely, the r/LiverpoolFC subreddit, one of the many communities hosted by the Reddit platform.[9] We focus on a short period of eight years, between 2011 and 2017. In order to enable a clearer observation of short-term meaning shift, we define two non-consecutive time bins: the first one (t_1) contains data from 2011–2013 and the second one (t_2) from 2017.[10] We also use a large sample of community-independent language for the initialization of the word vectors, namely, a random crawl from Reddit in 2013.[11] Table 2 shows the size of each sample.

Model. In the method proposed by Kim et al. (2014), word embeddings for the first time bin t_1 are initialized randomly; then, given a sequence of time-related samples, embeddings for t_i are initialized using the embeddings of t_{i-1} and further updated. If at t_i the word is used in the same contexts as in t_{i-1}, its embedding will only be marginally updated, whereas a major change in the context of use will lead to a stronger update of the embedding. The model makes embeddings across time bins directly comparable.

We implement the following steps:[12] First, we create word embeddings with the large sample Reddit$_{13}$, to obtain meaning representations that are community-independent.[13] We then use these embeddings to initialize those in LiverpoolFC$_{13}$, update the vectors on this sample, and thus obtain embeddings for time t_1. This step adapts the general embeddings from Reddit$_{13}$ to the LiverpoolFC community. Finally, we initialize the word embeddings for LiverpoolFC$_{17}$ with those of t_1, train on this sample and get embeddings for t_2.

The vocabulary is defined as the intersection of the vocabularies of the three samples (Reddit$_{13}$, LiverpoolFC$_{13}$, LiverpoolFC$_{17}$), and includes 157k words. For Reddit$_{13}$, we include only words that occur at least 20 times in the sample, so as to ensure meaningful representations for each word, while for the other two samples we do not use any frequency threshold: Since the embeddings used for the initialization of LiverpoolFC$_{13}$ encode community-independent meanings, if a word doesn't occur in LiverpoolFC$_{13}$ its representation will simply be as in Reddit$_{13}$, which reflects the idea that if a word is not used in a community, then its meaning is not altered within that community. We train with standard skip-gram parameters (Levy et al., 2015): window 5, learning rate 0.01, embedding dimension 200, hierarchical softmax.

Evaluation dataset. For evaluation and analysis, we create a small dataset of words to be annotated as positive or negative meaning shift examples by community members without linguistic background.[14] We initially leverage information about increase in frequency, which has been shown to positively correlate with meaning change (Wijaya and Yeniterzi, 2011; Kulkarni et al., 2015), and sample words with a significant increase in relative frequency between t_1 and t_2 (an increase is considered significant if it is at least two standard deviations above the mean).[15] Frequency increase is not a necessary condition for meaning shift to take place; however, given the positive correlation mentioned above, it is a reasonable starting point, as a random selection of words would contain very few positive examples. Our dataset is thus biased towards precision over recall.

This procedure yields ∼200 words. The first author of the paper went through the list of words to identify cases of potential meaning shift, based

[8] Three words were removed from the dataset: 'discord' and 'owls' due to the homonymy with proper names not detected by the authors during the implementation of the survey; 'tracking' because the chosen examples clearly mislead the judgements of the redditors.

[9] `https://www.reddit.com`.

[10] These choices ensure that the samples in these two time bins are approximately of the same size – see Table 2. The r/LiverpoolFC subreddit exists since 2009, but very little content was produced in 2009–2010.

[11] We used the Python package Praw for downloading the data, `https://pypi.python.org/pypi/praw`.

[12] We implement the model using the `gensim` library.

[13] Reddit$_{13}$ embeddings are initialized randomly.

[14] Domain knowledge is needed for this task.

[15] We consider content words only, which we identify by using the external list of common words available at `https://www.wordfrequency.info/free.asp`

on the analysis of the contexts of use in the r/LiverpoolFC data. By semantic shift, we mean change in the ontological type of what a word denotes (see examples in Section 5). We considered only new senses - i.e., not existing senses which increase in frequency- , both for monosemous and polysemous words. This resulted in 34 words. We added two types of counfounders: 33 words with a significant increase in frequency but not marked as meaning shift candidates and 33 words with constant frequency between t_1 and t_2, included as a sanity check.[16]

We then created an online survey, which we posted in the r/LiverpoolFC to recruit participants. The participants were shown the 100 words together with randomly chosen contexts of usage from each time period (1 to 5 contexts depending on the word). For feasibility, they were asked to label words as 'shift' or 'no shift', although semantic shift is better viewed as graded (see below). The order of presentation was randomized for each participant.[17] Overall, 26 members of r/LiverpoolFC participated in the survey, and each word in the dataset received on average 8.8 judgements. The final dataset includes 97 words.[18] Inter-annotator agreement, computed as Krippendorff's alpha, is $\alpha = 0.58$, a relatively low value but common in semantic tasks (Artstein and Poesio, 2008). We use the annotations to define a gradable *semantic shift index*, computed as the proportion of 'shift' judgements a word received in the survey.[19] The index ranges from 0 (no shift) to 1 (shift). As expected, all words with no frequency increase in t_2 have a shift index lower than 0.5, which validates our data selection method.

5 Results and Analysis

Our initial hypothesis, common to previous work, is that meaning shift is mirrored by a change in context of usage, which should be captured by an increased in cosine distance between the time-related vector representations of a word. The results of our experiment confirm this hypothesis: We find a positive correlation between cosine distance and semantic shift in our dataset (Pearson's r= 0.49, $p < 0.001$) - see Figure 1.

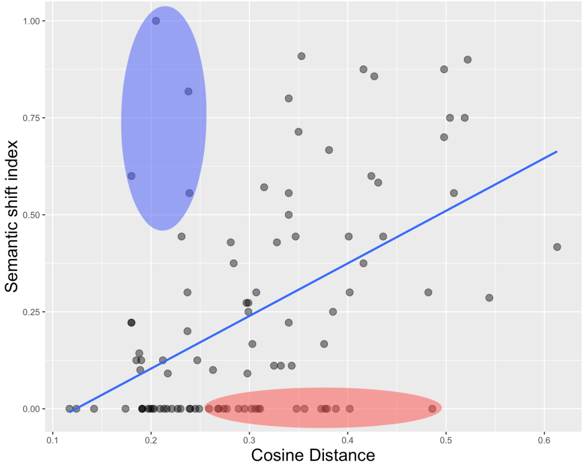

Figure 1: Semantic shift index vs. cosine distance for all words in the evaluation dataset (Pearson's r = 0.49, $p < 0.001$). Red ellipsis indicates false positives, blue ellipsis false negatives.

Among the words that undergo a strong shift are, for example, 'highlighter', which occurs in sentences like *'we are playing with the **highlighter** today'*, a metonymy used to talk about a kit in a colour similar to that of a highlighter, or 'lean', in *'I hope a **lean** comes soon!'*, due to players typically leaning on a Liverpool symbol when posing for a photo right after signing for the club. Particularly explanatory is the 'F5' example shown (in chronological order) in Table 3.[20] While 'F5' is initially used with its common usage of shortcut for refreshing a page (line 1), it then starts to denote the act of refreshing in order to get the lastest news about the possible transfer of a new player to LiverpoolFC (line 2). This use catches on and many redditors use it to express their tension while waiting for good news (3-5),[21] though not all members are aware of the new meaning of the word (6). After the player finally signed for the team, someone leaves the 'F5 squad' (7), and after a while, another member recalls the period in which the word was used (8).

Although the general tendency is in line with our expectations, we also find systematic deviations. First, *false positives*, that is, words that do not undergo semantic shift despite showing rela-

[16] All words have absolute frequency in range [50–500].

[17] Survey's instructions are in the supplementary material.

[18] Three words were discarded: 'discord' and 'owls' due to the homonymy with proper names not detected during survey's implementation; 'tracking' because the chosen examples clearly mislead the judgements of the redditors.

[19] The shift index is exclusively based on the judgements by the redditors, and does not consider the preliminary candidates selection done by one of us.

[20] The first example is from the June,18th, the last from the September,6th: such a short time span gives the idea of how quick is the meaning shift process considered in this work.

[21] Here the semantic change is accompanied by a change in the part of speech, and 'F5' becomes a denominal verb.

1.	after losing the F5 key on my keyboard,..
2.	F5 tapping is so intense now. I want him
3.	Don't think about it too much, man. Just F5.
4.	just woke up and thought it was f5 time
5.	this was a happy f5
6.	what is an F5?
7.	I'm leaving the f5 squad for now
8.	I made this during the f5 madness in the sub

Table 3: Example of use of 'F5'

tively large differences in context between t_1 and t_2 (red ellipsis in Figure 1; shift index=0, cosine distance>0.25). Manual inspection reveals that most of these "errors" are due to a referential effect: words are used almost exclusively to refer to a specific person or event, and so the context of use is narrowed down with respect to t_1. For instance, 'stubborn' is almost always used to talk about the coach of the team, who was not there in 2013 but only in 2017; 'entourage', for the entourage of one of the stars of the team; 'independence' for the political events of Catalunya. In all these cases, the meaning of the word stays the same, despite the change in context. In line with the Distributional Hypothesis, the model spots the change, but it is not sensitive to its nature. This is not a problem for long-term shift studies, because embeddings are built on a much larger number of occurrences and this makes them less sensitive to changes of referential nature like the one presented here. However, with smaller, community corpora, this problem clearly emerges.

A much smaller, but consistent group is that of *false negatives*, words that undergo semantic shift but are not captured by the model (blue ellipsis; shift index>0.5, cosine distance<0.25). These are cases of *extended* metaphor (Werth, 1994), that is, cases in which the metaphor is developed throughout the whole text produced by an author. Also in this case, the model "is right", in the sense that indeed the local context of the target words does not change in t_2. For instance, 'pharaoh' is the nickname of an Egyptian player who joined Liverpool in 2017 and is used in sentences like *'approved by our new **Pharaoh** Tutankhamun'*, *'our dear Egyptian **Pharaoh**, let's hope he becomes a God'*, and so on. Similarly, 'shovel', occurs in sentences like *'welcome aboard, here is your **shovel**'*, *'you boys know how to **shovel** coal'*: the team is seen as a train that is running through the season, and every supporter is asked to give its contribution, depicted as the act of shoving coal into the train boiler. Despite the metaphoric usage, the local context of these words is similar to the literal one, and so the model does not spot the meaning shift. We expect this to happen in long-term shift models, too, but we are not aware of results confirming this.

Contextual variability. As we have seen, the main issue for distributional models when dealing with short-term shift is contextual change due to referential aspects. We expect that in referential cases the context of use will be *narrower* than for words with actual semantic shift, because they are specific to one person or event. Hence, using a measure of *contextual variability* should help spot false positives. We define contextual variability as follows: for a target word, we create a vector for each of its contexts in t_2 by averaging the embeddings of the words occurring in it, and define variability as the average pairwise cosine distance between context vectors.[22]We then test whether contextual variability has explanatory power over cosine distance by fitting a linear regression model with these two variables as predictors and semantic shift index as dependent variable.[23] The results indicate that these two aspects are indeed complementary (contextual variability: β= 0.47, $p < 0.001$, cosine distance: β= 0.40, $p < 0.001$, adjusted R^2=0.44). While both shift words and referential cases change context of use in t_2, context variability captures the fact that only in referential cases words occur in a restricted set of contexts. The scatterplot in the supplementary material shows this effect visually. In future work, we plan to investigate more in depth the interplay between variability, cosine and semantic shift, both in short- and long-term meaning shift.

6 Conclusion

The goal of this preliminary study was to bring to the attention of the NLP community short-term meaning shift, an under-researched problem in the field. We hope that it will spark further research into a phenomenon which, besides being of theoretical interest, has potential practical implications for NLP downstream tasks concerned with user-generated text, as modeling how words meaning rapidly change in communities would allow to better understand what their members say.

[22] We consider the context as the five words occurring on the left and on the right of the target word.

[23] Contextual variability and cosine distance are not correlated in our data (Pearson's r= 0.18,$p > 0.05$).

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518. ACM.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *IWCS 2017-12th International Conference on Computational Semantics-Long papers*.

Marco Del Tredici, Malvina Nissim, and Andrea Zaninello. 2016. Tracing metaphors in time through self-distance in vector spaces. *arXiv preprint arXiv:1611.03279*.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *TACL*, 4:31–45.

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. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Proceedings of the eleventh International Conference on Web and Social Media*.

Ruqaiya Hasan. 2009. *Semantic variation: Meaning in society and in sociolinguistics*. Equinox London.

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*, pages 229–238. IEEE Press.

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. Association for Computational Linguistics.

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. ACM.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Lawrence Phillips, Kyle Shaffer, Dustin Arendt, Nathan Hodas, and Svitlana Volkova. 2017. Intrinsic and extrinsic evaluation of spatiotemporal text representations in twitter streams. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 201–210.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A Keim, and Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 305–310. Association for Computational Linguistics.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, pages 161–183.

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.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *arXiv preprint arXiv:1801.09872*.

Paul Werth. 1994. Extended metaphora text-world account. *Language and literature*, 3(2):79–103.

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.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *CogSci*.