Computational Perspectives on Semantic Change: Insights from Word2Vec, GloVe, and BERT

Course Essay

LT2314 Language Resources
Student: Stanislav Hakkarainen
Lecturer: Dana Dannélls

ABSTRACT

The study centers on conducting an experiment by detecting semantic changes in meanings of the same word. The experiment employs a dataset generated by ChatGPT containing ten words, four meanings for each word, and the approximate time stamps of each meaning. The experiment uses three models to calculate cosine similarities: Word2Vec, GloVe, and BERT. The results are placed on graphs to represent the findings and to highlight differences and similarities among models. The paper concludes with a discussion of the findings and possible ways of skewing the experiment.

INTRODUCTION

Language is a living organism. Once created, it keeps on evolving, generating new ways to change its form, accessibility, and applicability (Kim et al., 2014). The main agents of a language are words. Words are the ones that tend to change their meanings over time due to various sets of reasons. It can be historical, social, psychological causes, foreign influences, or simply the need for a new name (Steinmetz, 2008).

Illustratively, it is worth looking at the word "adamant". A contemporary speaker associates this word with "determined, unyielding, inflexible", while it was brought from the Latin adamás, meaning the diamond, something of the hardest quality (Steinmetz, 2008).

It becomes imperative to explore a language's changes over time using all the tools available at the time to see the language as an evolutionary phenomenon, to understand how malleable the language can be; how much it has changed over time, and, possibly, to attempt to predict its future changes.

This paper focuses on the detection of semantic change between the meanings of the same words to detect whether the words have changed their meanings or not, and if yes then possibly assess how drastic the change might be.

Three models have been chosen to be used in the aforementioned task: Word2Vec, GloVe, and BERT.

Additionally, the capacity of ChatGPT has been tested to generate words, and their meanings changed over time. The reasoning behind using the LLM to generate the words and the meanings was the idea to test whether the LLM could be used for such tasks as it has been trained on a large corpus of data from different time periods.

LITERATURE OVERVIEW

The interest in identifying semantic changes has made linguists ponder various approaches ever since the onset of the 19th century. In his paper, Marcin Grygiel contends, "It was at that time when many students of language first realized that sense alterations can no longer be treated as corruption or degeneration and tried to bring them into order and system" (Grygiel, 2005).

Looking at more contemporary attempts to detect semantic changes, as stated in the paper written by Rada Mihalcea and Vivi Nastase, "Until recently, such task would not have been possible because of the lack of large amounts of non-contemporary data. This has changed thanks to the Google books and Google Ngrams historical projects" (Mihalcea & Năstase, 2012). Researchers themselves used a supervised learning approach to predict the time a specific word might have belonged to.

In 2011, Derry Tanti Wijaya and Reyyan Yenitezi proposed a way to analyze the change that a word undergoes concerning the changes in the words that surround the chosen word. They implemented the Topics-Over-Time (TOT) and k-means clustering to detect clusters of topics that surround the word of their interest (Wijaya & Yeniterzi, 2011).

In their paper named "Behind the Times: Detecting Epoch Changes using Large Corpora", Octavian Popescu and Carlo Strappavara tackled the idea of using large corpora of chronologically ordered language to detect correlations between language usage and time periods (Popescu & Strapparava, 2013).

METHODOLOGY

For the sake of the experiment, ChatGPT was used to obtain words that we used over the course of the study. The LLM (Large Language Model) generated ten words, four meanings for each word, and four time stamps for each meaning to approximately pinpoint when in the timeline one of the four meanings was used (see Appendix A).

When generating words and meanings from ChatGPT, the following biases were considered:

1. Common Usage Bias: The model might be biased to generate words that are more commonly used in contemporary language.

2. Sensitivity to Input Wording: The word choice suggested by the model might be influenced by the wording and phrasing of the user's query.

3. Limited Polysemy Recognition: ChatGPT might not recognize the difference between various meanings of a word, specifically in cases of polysemy.

4. Incomplete Understanding: The model might not fully comprehend the nuances of certain words or their historical evolution, which might result in producing words based on surface-level patterns in the training data of the model.

The data from ChatGPT was extracted in a JSON format which consequently was converted into a CSV table. The only change that the data from the model underwent was the change in the time stamps. ChatGPT was not successful at producing exact years for each definition, instead, a century has been

given; thus, each century has been turned into a numeric value. For clarity, the time stamp for the first definition of the word "nice" was given as "13th century". The given format was transformed into the year "1200".

Speaking of the model parameters picked for the task, it is wise to talk about each model separately.

Word2Vec:

1. Vector Size: The following parameter defines the dimensionality of the word vectors. For this model, a vector size of 300 was picked. Since the dataset contains only ten words and the meanings of the words are rather petit, a smaller vector size was deemed as a reasonable decision.

2. Window Size: This parameter determines the maximum distance between the current and predicted words within a sentence. In this setting, the number of five was used.

3. Minimum Word Count: The parameter omits all the words with a total frequency lower than the specified value. Taking into account the fact that the "min_count" was set to one to count even the words that appear only once in the dataset.

4. Workers: This parameter was set to four to parallelize the training process and speed up computation (Gensim: Topic Modelling for Humans, n.d.).

Considering that the task was to capture the change of the meaning of a word, where the meaning is indicative of the word's changing semantics, the skip-gram was appraised to be more fitting for the task.

GloVe:
Unlike Word2Vec, the GloVe model does not have explicit parameters such as vector dimension or window size. The model's parameters are pre-determined during training. In the experiment's settings, the medium-sized model with 300-dimensional word vectors is used. The reason behind choosing the following pre-trained model lies in grasping the notion that the dataset for the experiment is rather small, which leads one to think that a medium-sized model will suffice.

It is also worth highlighting that GloVe vectors are based on co-occurrence statistics from large text corpora. They are pre-trained and might not be specific to the chosen experiment (Pennington, n.d.).

BERT:

For the sake of the experiment, the model is loaded with its default configuration for "bert-base-uncased". A common approach is selected for this model: the code uses the mean of the last hidden state of the BERT model for each token to obtain a single vector representation for the entire definition.

After each model produced cosine similarities for every word, the results were put onto a graph to visualize not only the progression of the cosine similarity change within one model but also to compare the results from the other two models and to observe whether the models would tend to follow a similar pattern in terms of detecting the changes in word meanings.

RESULTS

Having accumulated the results from all three models, the numbers were put onto a graph to detect and visualize semantic changes

between the first and the second meanings, as well as for the second/third and third/fourth meanings. Each graph represented the results of a singular word from Word2Vec, GloVe, and Bert.

WORD	SIM_1_2	SIM_2_3	SIM_2_3
Nice	0.365563	0.588406	1.000000
Girl	0.258767	0.336147	0.021395
Gay	0.021765	0.017710	0.000000
Awful	-0.035790	-0.042682	0.595105
Manufacture	0.723247	0.232880	0.422983
Villain	0.333942	0.262635	0.641178
Meat	0.280371	0.693011	0.616174
Silly	0.682940	0.583700	-0.068057
Cursor	0.630305	0.599630	1.000000
Guy	0.296224	0.056241	-0.028030

Table 1: Cosine Similarity Values Produced by Word2Vec.

WORD	SIM_1_2	SIM_2_3	SIM_3_4
Nice	0.724020	0.717571	0.729573
Girl	0.406771	0.486823	0.410270
Gay	0.264617	0.259576	0.387680
Awful	0.329916	0.346363	0.536652
Manufacture	0.954023	0.568576	0.658151
Villain	0.609213	0.634839	0.720194
Meat	0.615185	0.774768	0.828776
Silly	0.889003	0.771662	0.577265
Cursor	0.869057	0.710020	0.935544
Guy	0.707322	0.167815	0.499761

Table 2: Cosine Similarity Values Produced by GloVe.

WORD	SIM_1_2_	SIM_2_3	SIM_3_4
Nice	0.577930	0.565024	0.750605
Girl	0.684486	0.597035	0.613918
Gay	0.545578	0.624973	0.444866
Awful	0.654619	0.658465	0.645391
Manufacture	0.862474	0.737553	0.697798
Villain	0.638092	0.599023	0.740602
Meat	0.779187	0.739817	0.851496
Silly	0.853047	0.667896	0.634784
Cursor	0.882503	0.557948	0.777328
Guy	0.632681	0.573116	0.626453

Table 3: Cosine Similarity Values Produced by BERT.

Speaking about trends and patterns in the results, Word2Vect exhibits inconsistencies in similarity scored across various words, which might indicate that the model might struggle with maintaining a consistent performance. Additionally, the model shows the signs of struggles with words that have nuanced meanings, as seen in words like “Silly” which gives a negative similarity to the last pair of meanings.

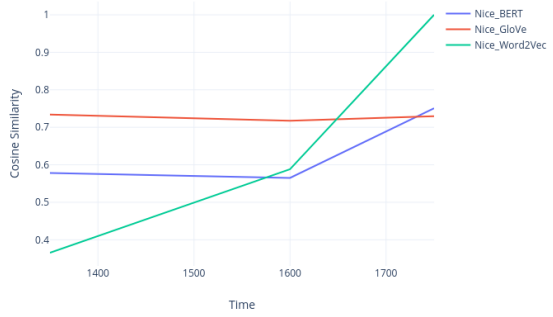
When it comes to GloVe, the model persists in performing well across various pairs of meanings, showcasing its reliability in capturing semantic relations. Moreover, GloVe seems to demonstrate robust semantic understanding by providing higher and more consistent similarity scores to nuanced words.

BERT consistently outperforms in capturing subtle distinctions and nuances in meanings, showcasing its precision in terms of capturing context. The model also provides secure and high similarity scores across various words, demonstrating that it can detect subtle changes in word meanings.

In terms of this paper, an examination of particular cases is warranted to elucidate the nuances of the study.

Nice:
Word2Vec demonstrates improvement in similarity scores for subsequent meanings, indicating a possible learning trend. Nonetheless, the drastic change in numbers, especially scoring 1.0 for the third and the fourth meanings and its initial inconsistency raises concerns regarding its ability to capture distinctions.

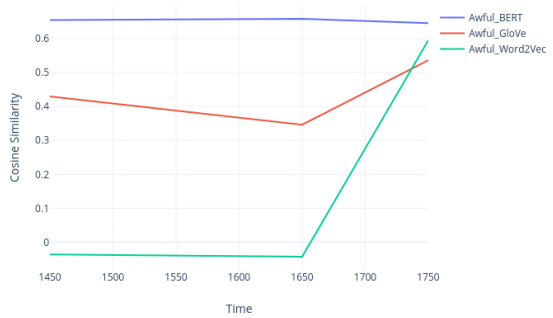
For GloVe and BERT, both models tend to consistently provide high similarity scores across different meanings, demonstrating a strong understanding of the word's various senses.



Graph 1: Semantic Change of Word "Nice".

Awful:
Word2Vec showcases uncertain results with varying scores, including a negative similarity. The following trend might indicate obstacles in accurately representing the different meanings of the word.

On the other hand, GloVe and BERT consistently offer high and positive similarity scores, proving their effectiveness in capturing nuanced meanings.



Graph 2: Semantic Change of Word "Awful".

Manufacture:

However, looking at another example, Word2Vec seems to demonstrate effective performance. The high similarity score for the first pair of meanings indicates that the model might effectively represent the commonalities. Additionally, moderate scores for subsequent pairs of meanings suggest an adequate understanding of the distinctions between the meanings.

Graph 3: Semantic Change of Word "Manufacture".

As proved by the examples, Word2Vec demonstrates limitations in maintaining consistency and struggles with nuanced distinctions in meanings. Nevertheless, for words with more straightforward and less ambiguous meanings, Word2Vec might still provide interpretable similarity scores. GloVe and BERT outperform in capturing subtle distinctions and providing robust semantic embedding consistency.

DISCUSSION

It is a prerogative to mention that there is no direct way of determining which model is better at detecting semantic changes in words. However, the model might be chosen based on the task at hand. If the task involved capturing subtle semantic differences, then a model that provides higher scores for truly similar meanings might be considered better, but in tasks requiring a deep understanding of context, models like BERT, which are context-aware, may be preferred.

Keeping this setting in mind, instead of determining an ultimately "better" model, one should pay close attention to the task itself and the outcomes of each model.

Even though there were trends of improvement in similarity scores for subsequent meanings, the inconsistency of Word2Vec's results might indicate a limitation in representing vague distinctions. GloVe and BERT, with their consistent scores, suggest a more stable representation of semantic relationships in the chosen words.

Speaking about factors influencing models' performances, training data is one of the first factors that arises. Considering the size of the training data for this experiment in particular, larger and more diverse datasets may lead to better model performance.

A potential limitation of the study is the use of a dataset generated by ChatGPT. As an LLM, it may noy fully grasp the indicates and shifts of certain words, resulting in surface-level patters influencing the generated words and affecting the accuracy of the results.

Another limitation is the use of century timestamps rather than exact year, which could lead to imprecise pinpointing of when each definition was used.

Generally, in models like Word2Vec and GloVe, the size of the context window influences how words are considered in relation to each other. Additionally, the choice of dimensionality affects model performance.

Finally, the characteristics of the language that is processed, including grammar, syntax, and the presence of specific linguistic features, can impact the outcome.

CONCLUSION

The key findings of the study highlight that given enough data, there is a way of developing a platform for evaluating semantic changes in word meanings by representing the cosine similarity scores on graphs.

Three models have been tested on a dataset generated by an LLM, and it has been concluded that, within the agreed settings of the experiment, there is no direct way of addressing the question of which model performs better.

One should also talk about the metric that has been chosen for this task and its plausible substitutes – cosine similarity.

Cosine Similarity is a widely used metric for measuring the similarity between vectors (Sidorov et al., 2014). Nonetheless, the approach has several setbacks.

Cosine Similarity numbers are sensitive to the magnitude of vectors. If the vector lengths of the meanings are drastically different, it can lead to misleading similarity scores. Additionally, the similarity only considers the angle between chosen vectors, not their direction. If two vectors with the same orientation but different magnitudes are compared, the trend might lead to a high yet untruthful cosine similarity (Søgaard et al., 2018; Turney & Pantel, 2010).

The similarity does not capture contextual information. Every dimension of the vector space is treated independently, disregarding the sequential or contextual relationships between meanings, which is crucial in measuring semantic changes (Turney and Pantel, 2010).

Moreover, by using the cosine similarity approach, one faces a challenge when dealing with polysemous or homonymous words. Lastly, the similarity sides with linear relationships between vectors. If the relationship between meanings is non-linear, which means that the connection between meanings is not well-represented by a straight line in a vector space or any linear transformation, cosine similarity may not accurately capture the semantic vagueness (Søgaard et al., 2018).

Additional or substitutional metrics should also be kept in mind.

For the sake of producing a more realistic picture, one could look into using Euclidean Distance, which measures the straight-line distance between two points in space (Danielsson, 1980).
Additionally, producing a word mover's distance (WMD), which originally measures the dissimilarity between two text documents as the minimum cumulative distance that words from one document need to pierce to match the distribution of words in the other document, could be an alternative approach to detect the changes in word meanings (Huang et al., 2016).

Lastly, one could turn to using a soft cosine similarity. The mentioned similarity supplements the original cosine similarity by considering a similarity measure that accounts for the relationship between words (meanings) in a more flexible manner by introducing a similarity matrix for incorporating semantic relationships (Sidorov et al., 2014).

To recapitulate, it is crucial to highlight plausible ways of improving the experiment.

Considering the fact that the experiment does not directly highlight a more suitable model to detect semantic shifts in word meanings, one could try and determine such a model by introducing the factor of human evaluation.

An alternative setting for the experiment could be bringing an individual to assess the meanings of the same word and give each pair of meanings a number that would correlate with their similarity or dissimilarity. By producing human numeric evaluations, it could become feasible to re-assess each model's performance to deduce which model generates similar numeric values.

Additionally, the dataset that has been used in the experiment could be enriched with more definitions of the same word and time stamps of the newly added definitions to portray a more advanced and refined image of the language change. Moreover, the definitions could be broadened by introducing more words in them to capture the slightest shifts between the meanings.

Another additional measure could be the use of different LLMs to produce additional datasets with the same words to determine whether the LLMS agree on the meanings of the words and how much the cosine similarity or any other metric values skew from the golden standard that is agreed upon in advance.

References:

1. Danielsson, P. (1980). Euclidean distance mapping. *Computer Graphics and Image Processing*, *14*(3), 227–248. https://doi.org/10.1016/0146-664x(80)90054-4
2. *Gensim: topic modelling for humans*. (n.d.). https://radimrehurek.com/gensim/models/word2vec.html
3. Grygiel, M. (2005). The methodology of analysing semantic change in historical perspective. *Studia Anglica Resoviensia*, *3*, 25–47. https://www.ur.edu.pl/files/ur/import/Import/2012/5/sar_v3_03.pdf
4. Huang, G., Guo, C., Kusner, M. J., Sun, Y., Weinberger, K. Q., & Sha, F. (2016). Supervised word mover's distance. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper_files/paper/2016/file/10c66082c124f8afe3df4886f5e516e0-Paper.pdf
5. Kim, Y., Chiu, Y., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal Analysis of Language through Neural Language Models. *arXiv Preprint arXiv:1405.3515*. https://doi.org/10.3115/v1/w14-2517
6. Mihalcea, R., & Năstase, V. (2012). Word Epoch Disambiguation: Finding how words change over time. *Meeting of the Association for Computational Linguistics*, 259–263. https://www.aclweb.org/anthology/P12-2051.pdf
7. Pennington, J. (n.d.). *GLOVE: Global Vectors for Word Representation*. https://nlp.stanford.edu/projects/glove/
8. Popescu, O., & Strapparava, C. (2013). Behind the Times: Detecting Epoch Changes using Large Corpora. *International Joint Conference on Natural Language Processing*, 347–355. https://aclanthology.org/I13-1040/
9. Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación Y Sistemas*, *18*(3). https://doi.org/10.13053/cys-18-3-2043
10. Søgaard, A., Ruder, S., & Vulić, I. (2018). On the Limitations of Unsupervised Bilingual Dictionary Induction. *arXiv Preprint arXiv:1805.03620*. https://doi.org/10.18653/v1/p18-1072
11. Steinmetz, S. (2008). *Semantic Antics: How and why words change meaning*. http://ci.nii.ac.jp/ncid/BA88094396
12. Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934
13. Wijaya, D., & Yeniterzi, R. (2011). Understanding semantic change of words over centuries. *Proceedings of the 2011 International Workshop on*

14. *DETecting and Exploiting Cultural diversiTy on the Social Web*. https://doi.org/10.1145/2064448.2064475

Appendix A:

Word	Guy	Cursor	Silly	Meat	Villain	Manufacture	Awful	Gay	Girl	Nice
First_definition	A man, person	A running messeng	Happy, fortunate	Any kind of food	A low-born rustic; a	To make by hand	Inspiring awe or fear	Happy, lively	Young person of either	Foolish, ignorant
First_year	18th century	17th century	15th century	7th century	14th century	16th century	14th century	14th century	14th century	13th century
Second_definition	A frightful figure	A runner, messeng er	Blessed, happy	Solid food, especiall y flesh	A man of ignoble pursuits	To make by machiner y	Frightful , ugly	Lively, brightly colored	A young unmarrie d woman	Wanton, dissolute
Second_year	19th century	18th century	16th century	14th century	15th century	18th century	17th century	19th century	15th century	16th century
Third_definition	Any person, fellow	A movable indicator on a computer screen	Innocent, harmless	Edible flesh of animals, especiall y mammal s	A wicked or evil person	To produce on a large scale	Deservin g of awe or reverenc e 18th century	Homose xual	Any woman	Reserved , shy
Third_year	20th century	20th century	17th century	16th century	16th century	19th century	18th century	20th century	17th century	18th century
Fourth_definition	A male friend or acquaint ance	A transpare nt slide used for marking on a	Lacking in good sense or judgment	The flesh of animals used as food	The antagoni st or antagoni st character	To create or produce	Extremel y bad or unpleasa nt	Derogato ry term for homosex uals	A female child	Pleasant, agreeabl e
Fourth_year	20th century	20th century	19th century	19th century	17th century	20th century	19th century	20th century	18th century	19th century