Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings

Terrence Szymanski
Insight Centre for Data Analytics
University College Dublin
terrence.szymanski@insight-centre.org

Abstract

This paper introduces the concept of temporal word analogies: pairs of words which occupy the same semantic space at different points in time. One well-known property of word embeddings is that they are able to effectively model traditional word analogies ("word w_1 is to word w_2 as word w_3 is to word w_4") through vector addition. Here, I show that temporal word analogies ("word w_1 at time t_α is like word w_2 at time t_β") can effectively be modeled with diachronic word embeddings, provided that the independent embedding spaces from each time period are appropriately transformed into a common vector space. When applied to a diachronic corpus of news articles, this method is able to identify temporal word analogies such as "*Ronald Reagan* in 1987 is like *Bill Clinton* in 1997", or "*Walkman* in 1987 is like *iPod* in 2007".

1 Background

The meanings of utterances change over time, due both to changes within the linguistic system and to changes in the state of the world. For example, the meaning of the word *awful* has changed over the past few centuries from something like "awe-inspiring" to something more like "very bad", due to a process of semantic drift. On the other hand, the phrase *president of the United States* has meant different things at different points in time due to the fact that different people have occupied that same position at different times. These are very different types of changes, and the latter may not even be considered a linguistic phenomenon, but both types of change are relevant to the concept of temporal word analogies.

I define a temporal word analogy (TWA) as a pair of words which occupy a similar semantic space at different points in time. For example, assuming that there is a semantic space associated with "President of the USA", this space was occupied by Ronald Reagan in the 1980s, and by Bill Clinton in the 1990s. So a temporal analogy holds: "*Ronald Reagan* in 1987 is like *Bill Clinton* in 1997".

Distributional semantics methods, particularly vector-space models of word meanings, have been employed to study both semantic change and word analogies, and as such are well-suited for the task of identifying TWAs. The principle behind these models, that the meaning of words can be captured by looking at the contexts in which they appear (i.e. other words), is not a recent idea, and is generally attributed to Harris (1954) or Firth (1957). The modern era of applying this principle algorithmically began with latent semantic analysis (LSA) (Landauer and Dumais, 1997), and the recent explosion in popularity of word embeddings is largely due to the very effective *word2vec* neural network approach to computing word embeddings (Mikolov et al., 2013a). In these types of vector space models (VSMs), the meaning of a word is represented as a multi-dimensional vector, and semantically-related words tend to have vectors that relate to one another in regular ways (e.g. by occupying nearby points in the vector space). One factor in word embeddings' recent popularity is their eye-catching ability to model word analogies using vector addition, as in the well-known example $king + man - woman = queen$ (Mikolov et al., 2013b).

Sagi et al. (2011) were the first to advocate the use of distributional semantics methods (specifically LSA) to automate and quantify large-scale studies of semantic change, in contrast to a more traditional approach in which a researcher inspects

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 448–453
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2071

a handful of selected words by hand. While not using a VSM approach, Mihalcea and Nastase (2012) used context words as features to perform "word epoch disambiguation", effectively capturing changes in word meanings over time. And several recent papers have combined neural embedding methods with large-scale diachronic corpora (e.g. the Google Books corpus) to study changes in word meanings over time. Kim et al. (2014) measured the drift (as cosine similarity) of a word's vector over time to identify words like *cell* and *gay* whose meaning changed over the past 100 years. Kulkarni et al. (2015) used a similar approach, combined with word frequencies and changepoint detection, to plot a word's movement through different lexical neighborhoods over time. Most recently, Hamilton et al. (2016) employed this methodology to discover and support two laws of semantic change, noting that words with higher frequency or higher levels of polysemy are more likely to experience semantic changes.

While the study of word meanings over time using diachronic text corpora is a relatively niche subject with little commercial applicability, it has recently gained attention in the broader computational linguistics community. A 2015 SemEval task was dedicated to Diachronic Text Evaluation (Popescu and Strapparava, 2015); while systems submitted to the task successfully predicted the date of a text using traditional machine learning algorithms (Szymanski and Lynch, 2015), none of the submissions employed distributional semantics methods. Also in 2015, the president of the ACL singled out recent work (cited in the previous paragraph) using word embeddings to study semantic change as valuable examples of "more scientific uses of distributed representations and Deep Learning" in computational linguistics (Manning, 2015).

The present work is inspired by this line of research, and is a continuation on the same theme with a twist: whereas past work has investigated specific words whose meanings have changed over time, the present work investigates specific meanings whose words have changed over time.

2 Method for Discovering Temporal Word Analogies

In general, work using diachronic word embeddings to study semantic change follows a common procedure: first, train independent VSMs for each time period in the diachronic corpus; second, transform those VSMs into a common space; and finally, compare a word's vectors from different VSMs to identify patterns of change over time. This paper follows a similar methodology.

Algorithm 1 Calculating temporal word analogies

$V_A \leftarrow Word2Vec(CorpusA)$
$V_B \leftarrow Word2Vec(CorpusB)$
$T \leftarrow FitTransform(V_B, V_A)$
$V_{B^*} \leftarrow ApplyTransform(T, V_B)$
$vec_1 \leftarrow LookupVector(V_A, word_1)$
$vec_2 \leftarrow NearestNeighbor(vec_1, V_{B^*})$
$word_2 \leftarrow LookupWord(V_{B^*}, vec_2)$

The general process of calculating a TWA is given in Algorithm 1. For example, if Corpus A is the 1987 texts, and Corpus B is the 1997 texts, and $word_1$ is *reagan*, then $word_2$ will be *clinton*. Crucially, it is only by applying the transformation to the B vector space that it becomes possible to directly compare the B^* space with vectors from the A space. The Python code used in this paper to implement this method is available to download.[1]

1. Training Independent VSMs: The data used in this analysis is a corpus of New York Times articles spanning the years 1987 through 2007, containing roughly 50 million words per year. The texts were lowercased and tokenized, and common bigrams were re-tokenized by a single pass of *word2phrase*. A separate embedding model was trained for each year in the corpus using *word2vec* with default-like parameters.[2] This resulted in 21 independent vector space models, each with a vocabulary of roughly 100k words.

2. Aligning Independent VSMs: Because training word embeddings typically begins with a random initial state, and the training itself is stochastic, the embeddings from different runs (even on the same dataset) are not comparable with one another. (Many properties of the embedding space, such as the distances between points, are consistent, but the actual values of the vectors are random.) This poses a challenge to work on diachronic word embeddings, which requires the ability to compare the vectors of the same word in different, independently-trained, VSMs. Previous work has employed different approaches to this problem:

[1] https://github.com/tdszyman/twapy.

[2] Example parameters: CBOW model, vector size 200, window size 8, negative samples 5.

Non-random initialization. In this approach, used by Kim et al. (2014), the values for the VSM are not randomly initialized, but instead are initialized with the values from a previously-trained model, e.g. training of the 1988 model would begin with values from the 1987 model.

Local linear regression. This approach, used by Kulkarni et al. (2015), assumes that two VSMs are equivalent under a linear transformation, and that most words' meanings do not change over time. A linear regression model is fit to a sample of vectors from the neighborhood of a word (hence "local"), minimizing the mean squared error, i.e. minimizing the distance between the two vectors of a given word in the two vector spaces. A potential drawback of this approach is that it must be applied separately for each focal word.

Orthogonal Procrustes. This approach, used by Hamilton et al. (2016), is similar to linear regression in that it aims to learn a transformation of one embedding space onto another, minimizing the distance between pairs of points, but uses a different mathematical method and is applied globally to the full space.

A thorough investigation into the relative benefits of the different methods listed above would be a valuable contribution to future work in this area. In the present work, I take a global linear regression approach, broadly similar to that used by Kulkarni et al. (2015). However, I use a large sample of points to fit the model, and I apply the transformation globally to the entire VSM. Experiments showed that the accuracy of the transformation (measured by the mean Euclidean distance between pairs of points) increases as the sample size increases: using 10% of the total vocabulary shared by the two models (i.e. roughly 10k out of 100k tokens) produces good results, but there is little reason not to use all of the points (perhaps excluding the specific words that are the target of study, although even this does not make much practical difference in the outputs).

3. Solving Temporal Word Analogies: Once the independent VSMs have been transformed, it is possible to compare vectors from one VSM with vectors from another VSM. Solving a TWA is then simply a matter of loading the vector of word w_1 from the VSM V_A, and then finding the point in the transformed VSM V_2^* closest to that vector. That point corresponds to word w_2, the solution to the analogy. It is also possible to align a series of VSMs to a single "root" VSM, and thereby trace the analogues of a word over time. The results of applying this method are discussed next.

3 Example Outputs and Analysis

Using the New York Times corpus, each of the 20 VSMs from 1998 to 2007 were aligned to the 1987 VSM, making it possible to trace a series of analogies over time (e.g. "Which words in 1988, 1989, 1990 ... are like *reagan* in 1987?"). A set of illustrative words from 1987 was chosen, and their vectors from the 1987 VSM were extracted. Then, for each subsequent year, the nearest word in that years' transformed VSM was found. The outputs are displayed in Table 1.

One way to interpret Table 1 is to think of each column as representing a single semantic concept, and the words in that column illustrate the different words embodying that concept at different points in time. So column 1 represents "President of the USA" and column 2 represents "mayor of New York City". The outputs of the TWA system perfectly reflect the names of the people holding these offices; likely due to the fact that these concepts are discussed in the corpus with high frequency and a well-defined lexical neighborhood (e.g. *White House*, *Oval Office*, *president*).

While the other columns do not produce quite as clean analogies, they do tell interesting stories. The breakup of the Soviet Union is visible in the transition from *soviet* to *russian* in 1992, and later that concept (something like "geopolitical foe of the United States") is taken up in the 2000s by North Korea and Iran, two members of George W. Bush's "Axis of Evil" . Changes in technology can be observed, with the 1987 vector for *Walkman* (representing something like "portable listening device") passing through *CD player* and *MP3 player* ultimately to *iPod* in 2007. Cultural changes can also be observed: the TWA "*yuppie* in 1987 is like *hipster* in 2003", is validated by reports in the media (Bergstein, 2016).

It is not easy to pin a precise meaning to each of these columns, and the "right" answer to any given TWA is to some degree a subjective judgment. Any given entity may fill multiple roles at once: which role should be the focus of the analogy? Each vector in the VSM can be thought of as combining multiple components: for example, the vectors for *reagan* include a component having to do with Ronald Reagan himself (based on words

1987	reagan	koch	soviet	iran_contra	navratilova	yuppie	walkman
1988	reagan	koch	soviet	iran_contra	sabatini	yuppie	tape_deck
1989	bush	koch	soviet	iran_contra	navratilova	yuppie	walkman
1990	bush	dinkins	soviet	iran_contra	navratilova	yuppie	headphones
1991	bush	dinkins	soviet	iran_contra	navratilova	yuppie	cassette_player
1992	bush	dinkins	russian	iran_contra	sabatini	yuppie	walkman
1993	clinton	dinkins	russian	iran_contra	navratilova	yuppie	cd_player
1994	clinton	mr_giuliani	russian	iran_contra	sanchez_vicario	yuppie	walkman
1995	clinton	giuliani	russian	white_house	graf	yuppie	cassette_player
1996	clinton	giuliani	russian	whitewater	graf	yuppie	walkman
1997	clinton	giuliani	russian	iran_contra	hingis	yuppie	headphones
1998	clinton	giuliani	russian	lewinsky	hingis	yuppie	headphones
1999	clinton	mayor_giuliani	russian	white_house	hingis	yuppie	buttons
2000	clinton	giuliani	russian	white_house	hingis	yuppie	headset
2001	bush	giuliani	russian	iran_contra	capriati	yuppie	headset
2002	bush	bloomberg	russian	white_house	hingis	gen_x	mp3_player
2003	bush	bloomberg	russian	white_house	agassi	hipsters	walkman
2004	bush	bloomberg	north_korean	iran_contra	federer	gen_x	headphones
2005	bush	bloomberg	north_korean	white_house	roddick	geek	ear_buds
2006	bush	bloomberg	iranian	white_house	hingis	teen	headset
2007	bush	bloomberg	iranian	capitol_hill	federer	dads	ipod

Table 1: Examples of words from 1987 and their analogues over time. Each column corresponds to a single point in vector space, and each row shows the word closest to that point in a given year.

relating to his personal attributes or names of family members) as well as a component having to do with the presidency (based on words like *president*, *veto* or *White House*). The analogies based on the 1987 *reagan* vector produce the names of other presidents over time (as in Table 1); however, if the 1999 *reagan* vector is used as the starting point, then 17 of the 20 analogies produced are either *reagan* or *ronald_reagan*. This illustrates how the vector from 1999 contains a stronger component of the individual man rather than the presidency, due to the change in how he was written about when no longer in office. Similarly, the *Iran Contra* vector can be viewed as a mixture of 'the Iran Contra crisis itself" and a more generic "White House scandal" concept. This second component causes Clinton-era scandals like *Whitewater* and *Lewinsky* to briefly appear in that space, while the first causes *Iran Contra* to continue to appear over time.

4 Evaluation

Standard evaluation sets of word analogies exist and are commonly used as a measure of the quality of word embeddings (but see Linzen (2016) for why this can be problematic and misleading). No data set of manually-verified TWAs currently exists, so a small evaluation set was assembled by hand: ten TWA categories were selected which could be expected to be both newsworthy and unambiguous, and values for each year in the corpus were identified using encyclopedic sources. When all pairs of years are considered, this yields a total of 4,200 individual TWAs. This data set, including the prediction outputs, is available online.

For comparison, a baseline system which always predicts the output w_2 to be the same as the input w_1 was implemented. (A system based on word co-occurrence counts was also implemented, but produced no usable output.[3]) Table 2 shows the accuracy of the embedding-based system and the baseline for each category. Accuracy is determined by exact match, so *mayor_giuliani* is incorrect for *giuliani*. Some categories are clearly much more difficult than others: prediction accuracy is 96% for "President of the USA", but less than 1% for "Best Actress". The names of Oscar Best Actress winners change every year with very little repetition, and it may be that an actress' role as an award winner only constitutes a small part of her overall news coverage.

Accuracy is a useful metric, but it is not necessarily the best way to evaluate TWAs. Due to the nature of the data (the U.S. President, for example, only changes every four or eight years), the baseline system works quite well when the time depth of the analogy ($\delta_t = |t_\alpha - t_\beta|$) is small. However,

[3]The co-occurrence matrices were expensive to construct due to the volume of data, and despite efforts to smooth the distributions, the analogy outputs were noisy, dominated by low-frequency tokens with relatively few non-zero components and high cosine similarities. But it is possible that with more careful engineering this approach could be effective.

	Baseline	Embeddings
CEO of Goldman Sachs	17.6	1.7
Governor of New York	44.8	62.4
Mayor of NYC	24.8	85.0
NFL MVP	1.9	1.4
Oscar Best Actress	1.0	0.5
President of the USA	40.0	96.4
Prime Minister of the UK	37.6	33.6
Secretary of State of USA	11.9	21.9
Super Bowl Champions	5.7	32.6
WTA Top-ranked Player	16.2	26.4
$\delta_t <= 5$ years	37.8	40.8
$\delta_t > 5$ years	6.9	32.7
Overall	20.1	36.2

Table 2: Analogy prediction accuracy.

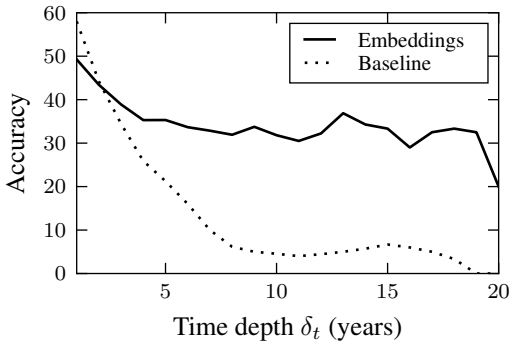

Figure 1: Accuracy as function of time depth.

as time depth increases, its accuracy drops sharply, while the embedding-based method remains effective, as illustrated in Figure 1. And even when the embedding-based system makes the "wrong" prediction, the output may still be insightful for data exploration, which is a more likely application for this method rather than prediction.

The analogies evaluated here have the benefits of being easy to compile and evaluate, but they represent only one specific subset of TWAs. Other, less-clearly-defined, types of analogies (like the *yuppie* and *walkman* examples) would require a less rigid (and more expensive), form of evaluation, such as obtaining human acceptability judgments of the automatically-produced analogies.

5 Conclusion

In this paper I have presented the concept of temporal word analogies and a method for identifying them using diachronic word embeddings. The method presented here is effective at solving TWAs, as shown in the evaluation, but its greater strength may be as a tool for data exploration. The small set of examples included in this paper illustrate political and social changes that unfold over time, and in the hands of users with diverse corpora and research questions, many more interesting analogies would likely be discovered using this same method.

The method presented in this paper is not the only way that TWAs could be predicted. If the VSMs could somehow be trained jointly, rather than independently, this would eliminate the need for transformation and the noise it introduces. Or perhaps it is sufficient to look at lexical neighborhoods, rather the vectors themselves. One limitation of the embedding approach is that it requires vast amounts of data from each time period: tens of millions of words are required to train a model with *word2vec*. This makes it impractical for many historical corpora, which tend to be much smaller than the corpus used here. If a simpler, count-based approach could be made to work, this might be more applicable to smaller corpora. A method which incorporates word frequency (Kulkarni et al., 2015) might be effective at identifying when one word drops from common usage and another word appears. And assigning a measure of confidence to the proposed TWAs could help automatically identify meaningful analogies from the vast combinations of words and years that exist.

In a domain where large quantities of real-time text are available, this method could potentially be applied as a form of event detection, identifying new entrants into a semantic space. And the same method described here could potentially be applied to other, non-diachronic, types of corpora. For example, given corpora of British English and American English, this methodology might be used to identify dialectal analogies, e.g. "*elevator* in US English is like *lift* in British English." Indeed, this general approach of comparing words in multiple embedding spaces could have many applications outside of diachronic linguistics.

Acknowledgments

I would like to thank all the participants at the Insight Centre for Data Analytics special interest group meeting on NLP at NUI Galway for their encouraging and insightful feedback. Thanks also to the anonymous ACL reviewers, for encouraging the addition of a quantitative evaluation. This work was partially supported by Science Foundation Ireland through the Insight Centre for Data Analytics under grant number SFI/12/RC/2289.

References

Rachelle Bergstein. 2016. Are hipsters the new yuppies? *forbes.com* October 12, 2016.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society.

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

Zellig Harris. 1954. Distributional structure. *Word* 10(23):146–162.

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.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211–140.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. pages 13–18.

Christopher D. Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics* 41(4).

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations Workshop*.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 746–751.

Octavian Popescu and Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. In *Current Methods in Historical Semantics*. De Gruyter Mouton.

Terrence Szymanski and Gerard Lynch. 2015. UCD: Diachronic text classification with character, word, and syntactic n-grams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.