Tracing armed conflicts with diachronic word embedding models

Andrey Kutuzov
Department of Informatics
University of Oslo
andreku@ifi.uio.no

Erik Velldal
Department of Informatics
University of Oslo
erikve@ifi.uio.no

Lilja Øvrelid
Department of Informatics
University of Oslo
liljao@ifi.uio.no

Abstract

Recent studies have shown that word embedding models can be used to trace time-related (diachronic) semantic shifts for particular words. In this paper, we evaluate some of these approaches on the new task of predicting the dynamics of global armed conflicts on a year-to-year basis, using a dataset from the field of conflict research as the gold standard and the Gigaword news corpus as the training data. The results show that much work still remains in extracting 'cultural' semantic shifts from diachronic word embedding models. At the same time, we present a new task complete with an evaluation set and introduce the 'anchor words' method which outperforms previous approaches on this data.

1 Introduction

Several recent studies have investigated how distributional word embeddings can be used for modeling language change, and particularly lexical semantic shifts. This includes tracing perspective change through time, usually for periods equal to centuries or decades; see (Hamilton et al., 2016b) among others. One of the main problems in these studies is the lack of proper ground truth resources describing the degree and direction of semantic change for particular words. Unfortunately, there is no such manually compiled compendium of all the semantic shifts that English words underwent in the last two centuries. The problem is even more severe for studies using more fine-grained time units spanning days or years, rather than decades, like in (Kulkarni et al., 2015) or (Kutuzov and Kuzmenko, 2016): When trying to uncover subtle changes of perspective (for example, '*Trump*' moving towards being associated with '*president*' rather than '*millionaire*'), it is difficult to find gold standard annotations for rigorous evaluation of the proposed methods.

In this paper, we make use of a social science dataset which to the best of our knowledge has not been introduced in the NLP field before. This dataset is described in section 3 and comprises a manually annotated history of armed conflicts starting from 1946 up to now. Together with word embedding models trained on temporal slices of the *Gigaword* news corpus (Parker et al., 2011), this allows us to properly evaluate several methods for tracing semantic shifts. We monitor changes in the local semantic neighborhoods of country names, applying it to the downstream task of predicting changes in the state of conflict for 52 countries at the year-level. This is essentially a classification task with 3 classes:

1. Nothing has changed in the country conflict state year-to-year (class '**stable**');
2. Armed conflicts have escalated in the country year-to-year (class '**war**');
3. Armed conflicts have calmed down in the country year-to-year (class '**peace**').

The results of this evaluation provide some insights into the performance of current semantic shift detection techniques and describe the best combinations of hyperparameters. We also propose the 'anchor words' method and show that it outperforms previous approaches when applied to this classification task.

2 Related work

Significant results have already been achieved in employing word embeddings to study diachronic

Proceedings of the Events and Stories in the News Workshop, pages 31–36,
Vancouver, Canada, August 4, 2017.

language change. Hamilton et al. (2016a) proposed an important distinction between cultural shifts and linguistic drifts. They showed that global embedding-based measures, like comparing the similarities of words to all other words in the lexicon in (Eger and Mehler, 2016), are sensitive to regular processes of linguistic drift, while local measures (comparing restricted lists of nearest associates) are a better fit for more irregular cultural shifts in word meaning. We here follow this latter path, because our downstream task (detecting armed conflicts dynamics from semantic representations of country names) certainly presupposes cultural shifts in the associations for these country names (not a real change of dictionary meaning). Additionally, local neighborhood measures of change are more sensitive to nouns, which makes them even better for our purpose.

It is important to note that in (Hamilton et al., 2016b) and other previous work on the subject, proper names were mostly filtered out: their authors were interested in more global semantic shifts for common nouns. In contrast to this, for the practical task of monitoring news streams, we here make proper names (countries and other toponyms) our main target. We are mostly interested in what is happening to this or that named entity, not in whether there were subtle changes in the meaning of some common noun. Another difference between the previous work and ours is that our time span is much smaller: not decades but years.

3 Data description

In this section we provide some background on the conflict dataset that forms the basis of our experiments, and the modifications we have applied to extract the gold standard to evaluate diachronic embeddings models.

The UCDP/PRIO Armed Conflict Dataset[1] maintained by the Uppsala Conflict Data Program[2] and the Peace Research Institute Oslo[3] is a manually annotated geographical and temporal dataset with information on armed conflicts, in the time period from 1946 to the present (Gleditsch et al., 2002). It encodes both internal and external conflicts, where at least one party is the government of a state. The Armed Conflict Dataset is widely used in conflict research; thus, this can be the beginning of a fruitful collaboration between social scientists and computational linguists.

The collection of the dataset started in the mid-1980s under the name *Conflict Data Project*, but has since then evolved constantly. In the autumn of 2003 the amount of work on conflict data collection led to a change in the name of the project and it was thus turned into the *Uppsala Conflict Data Program*.

An essential notion in the UCDP project is that of *armed conflict*, defined as 'a contested incompatibility concerning government and/or territory where the use of armed force between 2 parties results in at least 25 battle-related deaths' (Sundberg and Melander, 2013). Note that *armed force* here means the use of arms in order to promote the parties general position in the conflict, resulting in deaths. In turn, *arms* means any material means, e.g. manufactured weapons but also sticks, stones, fire, water etc. *Organized actor* can mean a government of an independent state, or a formally or informally organized group according to UCDP criteria [Ibid.].

The subset of the data that we employ is the *UCDP Conflict Termination dataset*.[4] It contains entries on starting and ending dates of about 2000 conflicts. We limited ourselves to the conflicts taking place between 1994 and 2010. We omitted the conflicts where both sides were governments (about 2% of the entries), for example, the 1998 conflict between India and Pakistan in Kashmir. The reason for this is that with these entries, distributional models have a hard time telling the name of the state (conflict actor) from the name of the territory (conflict location). Thus, we analyzed only the conflicts between a government and an insurgent armed group of some kind (these conflicts constitute the majority of the UCDP dataset anyway).

Another group of the omitted conflicts is where at least one of the sides was mentioned in the full *Gigaword* less than 100 times. The rationale for this decision was that these conflicts have too little contextual coverage in the corpus for our models to learn meaningful representations for them. These cases constitute about 1% of the entries.

In total, the resulting test set mentions 52 unique

[1] http://ucdp.uu.se/

[2] http://www.pcr.uu.se/research/ucdp/program_overview/about_ucdp/

[3] https://www.prio.org/Data/Armed-Conflict

[4] http://www.ucdp.uu.se/downloads/monadterm/ucdp-term-conf-2015.xlsx

locations and 673 unique armed conflicts. It also includes the UCDP intensity level of the conflict in the current year: 493 conflicts are tagged with the intensity level 1 (between 25 and 999 battle-related deaths), and 180 conflicts with the intensity level 2 (at least 1,000 battle-related deaths). For location–year pairs with no records in the UCDP dataset we assign the tag 0, indicating that there were no armed conflicts in this location at that time.

We then represented this data as a set of data points equal to the *differences* (δ) between the location's conflict state in the current year and in the previous year, 832 points in total (52 locations $\times$ 16 years). If there were several conflicts in the location in this particular year, we used the average of their intensities. As an example, for Congo, the transition from 2001 to 2002 was accompanied by the ending of armed conflicts. Thus, for the data point '*congo_2002*' we have $\delta = 0 - 1 = -1$. Then, there were no changes (each new δ has the value of 0) until 2006, when armed conflicts resumed with the intensity of 1. Thus, for the '*congo_2006*' data point, $\delta = 1 - 0 = 1$.

However, for practical reasons it is more useful to predict a human-interpretable class of the conflict state change, rather than a scalar value. A version of this test set was produced where δ values were transformed to classes:

$$class = \begin{cases} war & \text{if } \delta \geq 0.5 \\ peace & \text{if } \delta \leq -0.5 \\ stable & \text{otherwise} \end{cases}$$

The 'shifting' classes **War** and **Peace** constitute 10% and 11% of the data points respectively. Thus, they are minority classes and we are mostly interested in how good the evaluated models are in predicting them. Below we describe the evaluated approaches.

4 Evaluated approaches

For training distributional word embedding models, we employed the *Continuous Bag-of-Words*[5] algorithm proposed in (Mikolov et al., 2013), as implemented in the Gensim toolkit (Řehůřek and Sojka, 2010). This was chosen because it allows us to straightforwardly update the models incrementally with new data, unlike, for example, with *GloVe* (Pennington et al., 2014) or traditional PPMI+SVD matrices.

4.1 Representing time in the models

As we are dealing with temporal data, we experiment with different methods for representing chronological information in word embedding models. All *Gigaword* texts are annotated with publishing date, so it is trivial to compile yearly corpora starting from 1994. Then, we trained three sets of word embedding models, differing in the way they represent time:

1. yearly models, each trained from scratch on the corpora containing news texts from a particular year only (dubbed **separate** hereafter);

2. yearly models trained from scratch on the texts from the particular year and all the previous years (**cumulative** hereafter);

3. incrementally trained models (**incremental**).

The last type is most interesting: here we actually 'update' one and the same model with new data, expanding the vocabulary if needed. Our hypothesis was that this can help coping with the inherently stochastic nature of predictive distributional models. However, this turned out to be not entirely true (see Section 5).

4.2 Detecting and quantifying semantic shifts

Once the sets of models are there, one can detect semantic shifts in a given query word w_q (in our case, always a location name), with two major existing approaches:

1. align two models (current and previous year, M_{cur} and M_{prev}) using the orthogonal Procrustes transformation, and then measure cosine similarity between the w_q vectors in both models, as proposed in (Hamilton et al., 2016b);

2. alternatively, define a set of *anchor words* related to the semantic categories we are interested in, and then measure the 'drift' of w_q towards or away from these 'anchors' in M_{cur} compared against M_{prev}. This is the method we propose in this paper.

The first approach outputs one value of cosine similarity for each data point, representing the degree of the semantic shift, but not its direction. In

[5] *Continuous Skipgram* showed comparable but slightly worse results, thus we report only those for CBOW.

contrast, the *anchor words* method can potentially provide information about the exact direction of the shift. This can be quantified in two ways:

1. for each anchor, calculate its *cosine similarity* against w_q in M_{cur} and M_{prev} (dubbed **Sim** hereafter);

2. as above, but instead of using the cosine, find the *position of each anchor in the models' vocabulary* sorted by similarity to w_q; we normalize by the size of the vocabulary so that rank 1 means the the anchor is the most similar word to w_q while rank 0 means it is the least similar (we dub this approach **Rank**).

The selection of anchor words is further described in Section 5, but for now note that both methods produce two vectors $\vec{R_{prev}}$ and $\vec{R_{cur}}$, corresponding to the models M_{cur} and M_{prev}. Their size is equal to the number of the anchor words, and each component of these vectors represents the relation of w_q to a particular anchor word in a particular time period.

To compute the differences between these vectors, one can either:

1. calculate the *cosine distance between these 'second-order vectors'*, as described in (Hamilton et al., 2016a); we dub this **SimDist** or **RankDist**, depending on whether **Sim** or **Rank** was used;

2. element-wise *subtract* $\vec{R_{prev}}$ *from* $\vec{R_{cur}}$ to get the idea of whether w_q drifted towards or away from the anchors; we dub this **SimSub** or **RankSub**.

In the first case, the output is again one value, and in the second case it is the vector of diachronic differences, with the size equal to the number of the anchor words. These 'features' can then be fed into any classifier algorithm.

5 Results

To predict the actual 'direction' of the semantic shift (whether armed conflicts are escalating in the location or vice versa), one needs to perform classification into 3 classes: **war**, **peace** and **stable**.

To evaluate the approaches described in Section 4, we need a set of anchor words strongly related to the topic of armed conflicts. For this we adopted the list of search strings used within UCDP to filter the news texts for subsequent manual coding (Croicu and Sundberg, 2015): *kill*, *die*, *injury*, *dead*, *death*, *wound*, *massacre*. Additionally, an expanded version of this list was created, where every initial anchor word is accompanied with its 5 nearest associates (belonging to the same part of speech) in the *CBOW* model trained on the full *Gigaword*. This resulted in a set of 26 words (some nearest associates overlap).

Approach	Separ.	Cumul.	Increm.
Procrustes	0.15	0.24	0.29
	Basic word list		
SimDist	0.27	0.17	0.25
SimSub	0.31	0.26	0.26
RankDist	0.28	0.19	0.23
RankSub	0.26	0.22	0.21
	Expanded word list		
SimDist	0.25	0.18	0.23
SimSub	0.35	0.31	0.29
RankDist	0.24	0.20	0.28
RankSub	**0.36**	0.30	0.32

Table 1: Macro-F1 measure of predicting conflict state changes (ternary classification)

The classification itself was done using a one-vs-rest SVM (Boser et al., 1992) with balanced class weights. The features used were either the cosine distance between $\vec{R_{prev}}$ and $\vec{R_{cur}}$ (in the case of **SimDist** and **RankDist**) or the result of $\vec{R_{cur}} - \vec{R_{prev}}$ (in the case of **SimSub** and **RankSub**). In the first case we have only one feature, while in the second case the number of features depends on the number of the anchor words.

The results for CBOW, evaluated with 10-fold stratified cross-validation, are presented in Table 1 in the form of macro-averaged F1.

The labels for approaches are the same as in section 4. *Procrustes* is our baseline: it does not use any anchor words, only the cosine distances between w_q in aligned models.

Overall, one can see that more words in the anchor sets is beneficial, and using $\vec{R_{cur}} - \vec{R_{prev}}$ (**Sub**) is almost always better than $\cos(\vec{R_{cur}}, \vec{R_{prev}})$ (**Dist**). As for the using of either cosine similarities (**Sim**) or ranks (**Rank**) as $\vec{R}$ values, there does not seem to be a clear winner. We also tried to concatenate similarities and ranks

Class	Precision	Recall	F1
Peace	**0.13** (0.06)	**0.29** (0.06)	**0.18** (0.06)
Stable	0.80 (0.79)	0.58 (**0.82**)	0.67 (**0.80**)
War	**0.17** (0.12)	**0.33** (0.08)	**0.22** (0.10)

Table 2: Detailed performance of the best model (results of weighted random guess in parenthesis)

to produce the feature vector of size 52. However, this did not improve the classifier performance.

It is interesting that the best results are shown by the **separate** models: at least for this particular task, it does not make sense to employ schemes of updating the models with new data or concatenating new corpora with the previous ones. It seems that the models trained from scratch on yearly corpora are more 'focused' on the events happening in this particular year, and thus are more useful.

Note that for the Procrustes alignment baseline it is vice versa: separate models are the worst choice for alignment, probably because they are too different from each other (each initialized independently and with independent collection of training texts). Anyway, the anchor words approach outperforms the Procrustes alignment baseline in all types of models. Hamilton et al. (2016b) report almost perfect accuracy for the Procrustes transformation when detecting the direction of semantic change (for example, the meaning of the word '*gay*' moving away from '*happy*' and towards '*homosexual*'). However, our task and data is different: the time periods are much more granular and we attempt to detect subtle associative drifts (often pendulum-like) rather than full-scale shifts of the meaning.

Table 2 provides the detailed per-class performance of the best model (**separate** CBOW with the expanded word list, using differences in anchor ranks as features). In parenthesis, we give the performance values for the stratified random guess baseline. Detecting stability breaks seems to be more difficult than detecting the 'no changes' state. The performance for the '**war**' and '**peace**' minority classes is far from ideal. However, it is significantly better than chance.

6 Conclusion

In this paper, we evaluated several approaches for extracting diachronic semantic shifts from word embedding models trained on texts from different time periods. We have focused on time spans equal to one year, using the Gigaword news collection as the training corpus. As the gold standard for testing, we adapted a dataset from the field of conflict research provided by the UCDP and containing manually annotated data about the dates of armed conflicts starting and ending all over the world. Thus, we applied diachronic word embedding models to the task of predicting the events of conflicts escalating or calming down in 52 geographical locations, spanning over 16 years (1994–2010)[6].

The conclusion is that tracing actual real-world events by detecting 'cultural' semantic shifts in distributional semantic models is a difficult task, and much work is still to be done here. The approaches proposed in the previous work – mainly for large-scale shifts observed over decades or even centuries – are not very successful in this more fine-grained task. Our proposed '*anchor words*' method outperforms them by large margin, but its performance is still not entirely satisfactory, achieving a macro F1 measure of 0.36 on the task of ternary classification ('stable', 'escalating', 'calming down').

We plan to further study ways to improve the performance of diachronic word embedding models in the area of armed conflicts and other types of events. If successful, these techniques can be used to semi-automate the labor-intensive process of manually annotating the social science data, as well as to mine news text streams for emerging events and trends. It can also be interesting to trace differences in diachronic representations relative to the source of the training texts (for example, the NYT newspaper against the Xinhua news agency).

References

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, pages 144–152.

Mihai Croicu and Ralph Sundberg. 2015. UCDP georeferenced event dataset codebook version 4.0. *Journal of Peace Research* 50(4):523–532.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating mean-

[6]The test set is available here: http://ltr.uio.no/~andreku/armedconflicts/ucdp_conflicts_1994_2010_testset.tsv.

ing variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 52–58.

Nils Petter Gleditsch, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand. 2002. Armed conflict 1946-2001: A new dataset. *Journal of peace research* 39(5):615–637.

L. William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2116–2121.

L. William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. 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)*. Association for Computational Linguistics, pages 1489–1501.

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*. Florence, Italy, pages 625–635.

Andrey Kutuzov and Elizaveta Kuzmenko. 2016. Cross-lingual trends detection for named entities in news texts with dynamic neural embedding models. In *First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016)*. Technical University of Aachen, pages 27–32.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26* pages 3111–3119.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition LDC2011T07. Technical report, Linguistic Data Consortium, Philadelphia.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta, pages 45–50.

Ralph Sundberg and Erik Melander. 2013. Introducing the UCDP georeferenced event dataset. *Journal of Peace Research* 50(4):523–532.