Cross-lingual Trends Detection for Named Entities in News Texts with Dynamic Neural Embedding Models

Andrey Kutuzov
University of Oslo
Postboks 1080 Blindern 0316, Oslo, Norway
andreku@ifi.uio.no
Elizaveta Kuzmenko
National Research University Higher School of Economics
Moscow, Russia
eakuzmenko_2@edu.hse.ru

Abstract

This paper presents an approach to detect real-world events as manifested in news texts. We use vector space models, particularly neural embeddings (prediction-based distributional models). The models are trained on a large 'reference' corpus and then successively updated with new textual data from daily news. For given words or multi-word entities, calculating difference between their vector representations in two or more models allows to find out association shifts that happen to these words over time. The hypothesis is tested on country names, using news corpora for English and Russian language. We show that this approach successfully extracts meaningful temporal trends for named entities regardless of a language.

1 Introduction

We propose an approach to track changes happening to real-world entities (in our case, countries) with the help of constantly updated distributional semantic models. We show how one can train such models on new textual data arriving daily and draw conclusions about events based on changes in word vectors induced by new contexts. In other words, subtle *semantic shifts* which the words undergo over time, influenced by real-world events, are detected by the presented method.

Detecting semantic shifts can be of use in a variety of linguistic applications. First, this method can be of help in the problem of automatically monitoring events through the stream of texts [AGK01]. Detected semantic shifts can potentially be used as additional features in the algorithms aimed at extracting the course of events. Without unsupervised approaches, it is impossible to process all the continuously generated data. This is the primary motivation factor for our research. Second, the developed approach can be used to study language shift and compare temporal corpora slices. This language area is traditionally studied by linguists, who put a lot of efforts into describing semantic shifts with the help of dictionaries, corpora and sociolinguistic research. At the same time, it is impossible to grasp all the language vocabulary and describe every lexical shift manually. Distributional semantic models facilitate this task.

The approaches to events detection and modeling of language shifts have a lot in common. First techniques employed various frequency metrics [JS09] and shallow semantic modeling [KNR15], [HBB10]. With the emergence of distributive semantic models detection of semantic shifts acquired new potential, as it was shown that word embeddings significantly improve the performance of algorithms [KARPS15].

The rest of the paper is organized as follows. In Section 2 we introduce the basics of prediction-based vector models of semantics. Section 3 describes the

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

principles of comparing such models, trained on pieces of text which follow each other in time. Specifics of our datasets are covered in Section 4, followed by the description of experimental setting in Section 5. Section 6 evaluates the results and in Section 7 we conclude.

2 Distributed Semantic Models

Vector space models (VSMs) are well established in the field of computational linguistics and have been studied for decades (see [TP$^+$10], [Reh11]). Essentially, a model is a set of words and corresponding vectors, which are produced from typical contexts for a given word. The most widespread type of contexts is other words co-occurring with a given one, which means that the set of all possible contexts generally equals the size of the vocabulary of the corpus. The dimensionality of the resulting *count model* can be reduced with well-known techniques like Principal Components Analysis (PCA) or Singular Value Decomposition (SVD). But in turn, this effectively forbids online training (continuously updating the model with new data), because after each update one has to perform computationally expensive dimensionality reduction over the whole co-occurrence matrix.

To overcome this, we employ a type of VSMs called *prediction-based models*: particularly, Continuous Bag-of-Words (CBOW) algorithm ([BDV03], [MSC$^+$13])[1]. Predictive models rather approximate co-occurrence data, instead of counting it directly, and show a promising set of properties. Using them, one directly learns dense lexical vectors (*embeddings*). Vectors are initialized randomly and then, as we move through the training corpus with a sliding window of a pre-defined width, gradually converge to values maximizing the likelihood of correctly predicting lexical neighbors. Such models as a rule use artificial neural networks to train; this is why they are sometimes called *neural models*.

For our task, it is important that predictive models can be updated with new co-occurrence data in a quite straightforward way. As already said, this is usually not the case with count models which demand computationally expensive calculations each time a new text is added.

3 Introducing Temporal Dimension to Vector Models

Detecting semantic shifts which words undergo over time demands the ability to somehow compare reference ('baseline') and updated models, representing later periods of time.

[1]The well-known *word2vec* tool also implements SkipGram, which is another predictive algorithm. However, it is more computationally expensive, and we leave its usage for future work.

The idea of employing changes in distributional semantic models to track semantic shifts is not in itself new. [KCH$^+$14] proposed to detect language change with chronologically trained models. However, they used rather simplified measure of 'distance' between word vectors at different time slices, namely, raw cosine distance. We employ more sophisticated methods as described further. [POL10] developed an approach to the First Story Detection in Twitter posts. Their research is similar to ours in that it deals with streaming data. The authors explore the space of documents and compare new tweets to the existing ones. However, the algorithm is developed specifically for short texts like tweets, which differ radically from news pieces analyzed in the presented paper.

Updating a neural model with new texts (in addition to the base training corpus used for initial training) is technically straightforward. After that, we have two models M_1 and M_n, where the former is the 'baseline' reference model, and the latter is the updated one (or a sequence of n updated models, each corresponding to the next time period), probably bringing new semantic shifts. This dynamic model in a way tries to imitate human brain learning new things, gradually 'updating' its state with new input data every day.

What are the possible ways to extract these changes? Suppose there is a set S of named entities (organizations, locations or persons we are interested in). Initially in the model M_1, each element of S can be thought of as possessing a number of topical '*associates*' or '*nearest neighbors*': words with their respective vectors closest to this element vector, ranked by their closeness or similarity. The exact number of nearest neighbors we consider in the simplest case is defined arbitrarily (for example, 10 nearest words). As we update the model with new data, co-occurrence counts for the elements of S are gradually growing (the model sees them in new contexts). It means than in each successive model M_n learned vectors for elements of S can be different.

If contexts for these words remain pretty much the same throughout the training data, the list of associates (nearest neighbors) in M_n will also remain intact. However, if a word acquires new typical contexts or loses some previous ones, its neural embedding will change: a *semantic shift* happens. Accordingly, we will see a new list of associates. For example, the vector representation for the word *president* may change so that its nearest neighbor is the vector for the name of the actual president of a country, instead of the previous one.

In this way, lists of nearest neighbors can be compared across models trained on different corpora or across one and the same model after an incremental update (as in the presented research). Substantial

changes or *bursts* in such lists for the named entities we are interested in may signal that these entities have undergone or are undergoing semantic shifts, which in turn reflects real-world events. We dub this approach '*dynamic neural embedding models*'.

Sets of neighbors in different models can be compared in many ways. Approaches to this range from simple Jaccard index [Jac01] to complex graph-based algorithms. We test two methods:

1. *Kendall's τ coefficient* [Ken48], which measures similarity of item rankings in two sets. Intuitively, it is important to pay attention not only to raw appearance of some words in the nearest neighbors set, but also to their rankings in it.

2. *Relative Neighborhood Tree* (RNT), introduced by [CGS15]. It essentially produces a tree graph with the target word as its root, nearest neighbors as vertexes and similarities between them as weighted edges. We then select the immediate neighbors of the target word in this tree and rank them according to their cosine similarity to the target word. These rankings are then compared across models using the same *Kendall's τ*.

The reason behind the second method is that it theoretically allows a deeper analysis of nearest neighbors' sets structure. Obviously, the neighbors participate in similarity relations not only with the target word but also between themselves. These relations convey meaning as well, making it possible to find the most 'important' neighbors. Graph-based methods to analyze relations between words in distributional models were also used in [KWHdR15]; note, however, that the problem they deal with is inverse to ours – they attempt to trace changes in surface words for a stable set of concepts, while we attempt to trace semantic shifts (changes in underlying concepts for a stable set of words).

We hoped that this graph-supported 'pre-selection' would allow Kendall's τ to improve the performance of the model. However, these expectations failed and simple ranking turned out to be more efficient than graph-based methods; see Section 6.

4 Data Description

We test our approach on lemmatized corpora of English and Russian news texts. The English corpus consists of *The Signal Media Dataset*[2], which contains 265,512 blog articles and 734,488 news articles from September 2015. The size of the corpus (after lemmatizing and removing stop words) is 222,928,287 words. We employ Stanford POS tagger [TKMS03] to extract lemmas and to assign each lemma a part-of-speech tag.

In order to test whether extracted semantic shifts are consistent across languages, we use a corpus of news articles in Russian published in September 2015 (unfortunately, not available publicly due to copyright restrictions). It contains about 500,000 texts extracted from about 1000 Russian-language news sites. The size of the corpus (after lemmatizing and removing stop-words) is 59,167,835 words. We employ Mystem [Seg03], a state-of-the art tagger for Russian to produce lemmas and part-of-speech tags.

5 Experimental setting

News texts from September 2015 do not seem to be a good training set alone. This is because such a corpus is inevitably limited in language coverage, lacking relations to events that happened earlier. Therefore, we first train a 'reference' or 'baseline' model which aims to mimic some background knowledge, which is then exposed to daily updates. For English, we used British National Corpus[3] (about 50 million words) to train this reference model, while for Russian it was the corpus of news articles published in the months preceding September 2015, precisely June, July and August (taken from the same source as the September articles). This corpus contains about 250 million words.

We acknowledge it is not quite correct to employ different types of corpora for 'reference' models in English and Russian. However, in a way, we compensate the quality and balance of BNC with the larger size of the reference corpus in Russian. In the future we plan to eliminate this inconsistency by using an analogous set of English news published in summer months or by employing Wikipedia dumps as reference corpora for both languages.

Both corpora were merged with same-language texts released in the first half of September 2015 (before 14th of September), in order to seed baseline models with some initial 'knowledge' of events and entities belonging to this month. Then, Continuous Bag-of-Words models were trained for both corpora, using negative sampling with 10 samples, vector size 300, symmetric window size 5 and 5 iterations. Words with frequency less than 10 were ignored during training.

After that, we successively updated these models with texts released in the following September time periods: 14th–15th, 16th–17th, 18th–20th, 21th–22th, 23th–24th, 25th–27th, and 28th–30th. Granularity of 2 or 3 days was chosen in order to enlarge the amount of data fed to models: for example, some one-day Russian corpora corresponding to weekends contained only

[2] `http://research.signalmedia.co/newsir16/signal-dataset.html`

[3] `http://www.natcorp.ox.ac.uk/`

several thousand words. For this reason, we additionally tried to include week-ends in 3-days periods, to make news stream more evenly distributed. As a result, average time period size in tokens was 18,774,000 for English data and 5,332,000 for Russian data.

We once again emphasize that our baseline models were not re-trained from scratch with new texts added from new corpora. Instead, we continued training the same model, gradually updating word vectors with new contexts. All interim states were saved as separate models, and in the end we had 8 successive models for each language.

We extracted English and Russian countries names from Wikipedia list of all world countries[4] and manually checked and normalized it, bringing all name variants to one lexeme. Then we filtered out the entities with frequency less than 30 per million words in either of our two reference corpora (English and Russian), producing a set CS of 36 frequent country names[5].

Finally, for each of the successive models, we found nearest neighbor sets for each entity in CS and compared them to the sets from the model state at the previous time period. Kendall's τ and Relative Neighborhood Tree (RNT) were used to compute similarity coefficients for each country within the given pair of models. This provided us with two lists of countries (for each language) ranked by their similarity to the same country in the 'previous' model. Supposedly, countries in which some major events happened during the last days have to position low in these lists, because their associations in news texts drifted towards the recent event or an opinion burst.

Let's illustrate how news texts and changes in the models reflect the real-life events by comparing 10 nearest associates for *Chile* in the English and Russian corpora. On the 16th of September 2015 there was an earthquake in Chile, and we can detect its 'echo' in the changes between our models for 14th–15th and 16th–17th of September (see Table 1).

Before the 16th of September, associates for *Chile* in both models were mostly the neighboring countries. However, after the earthquake things have completely changed: there was a strong bias towards this topic in news and blogs, and this is reflected in vectors for the word. 60% of English and 20% of Russian associates are now related to the event.

Kendall's τ coefficient between these two neighbors lists is as low as 0 (neighbors are completely replaced) for English and 0.56 for Russian. Average Kendall's τ for CS is 0.56 in the English models for the two days in question, with standard deviation 0.12. Thus, in the case of English, the change to the neighbors' set can be considered a significant burst, well above simple chance. In the case of Russian, Kendall's τ lies only 1 point below the average value of 0.57. It is obvious that Russian mass media paid less attention to the earthquake (they are more concerned with Michelle Bachelet, Chile's president), but the event is still reflected in the nearest neighbors set.

Table 1: Change in *Chile*'s neighbor set

14th–15th September		16th–17th September	
English	Russian	English	Russian
peru	бачелет	*quake*	аргентина
bolivia	аргентина	*earthquake*	бачелет (bachelet)
colombia	коста-рика	santiago	никарагуа
argentina	перчик	chilean	мексика
honduras	никарагуа	*tremor*	бельгия
brazil	швейцария	*tsunami*	исландия
ecuador	бельгия	*aftershock*	тунис
nicaragua	исландия	chileans	*магнитуда* (magnitude)
paraguay	аргентин	*temblor*	*землетрясение* (earthquake)
enchiladas	гватемала	kyushu	коста-рика

The next section describes how we employed cross-linguality of the data to evaluate the presented approach.

6 Cross-Lingual Evaluation of Events Detection

There is no 'golden standard' or ground truth which would allow to evaluate precision and recall of our events and associations extraction, and to tune hyperparameters of the algorithms. However, there is a way to indirectly estimate their performance in a kind of intrinsic evaluation.

We hypothesize that the better is an algorithm of detecting semantic shifts, the closer should be its results on model sequences trained on different language corpora. Obviously, national media focus on different topics, but this mostly concerns the domestic news. As for the world news, the worst scenario could be that a news story is not covered in national media of a particular country. However, such scenarios should be rare. In other cases, the perspective on a story can differ, but the 'burst' should remain the same[6].

Thus, English and Russian countries lists ranked by their 'burstiness' can be compared using Spearman's ρ

[4] `https://en.wikipedia.org/wiki/List_of_sovereign_states`

[5] Low-frequency country names bring in noise, because their vectors are susceptible to wild fluctuations when exposed to even a small amount of new contexts.

[6] Analyzing the degree to which the vision of events is different in national media is beyond the scope of the present research.

Table 2: 5 countries with most changed neighbors' sets (of total 36) between September 18–20 and 21–22

Rank	English	Russian (translated)
1	Italy	**Japan**
2	**Georgia**	Brazil
3	Malaysia	**China**
4	**Japan**	Spain
5	**China**	**Georgia**

[Spe04] for each time period. As there are 7 shifts from one time period to another, we use median of ρ values for these 7 cases as a tentative measure of algorithm's performance. The Table 2 gives an example of such country rankings for the changes between 18–20 and 21–22 of September. One can see that the top lists are highly similar, with 3 of 5 countries appearing in both (actual Sperman's ρ for the total lists of 36 countries between these periods is 0.5).

Overall results of applying this approach to the whole dataset using two our algorithms (with different sizes of nearest neighbors' sets to consider) are presented in the Table 3. We also applied it to a simple baseline method, where nearest neighbors are words which most frequently occurred in the window of 5 tokens to the right and to the left of the target entity in the given corpus.

Table 3: Cross-lingual evaluation

Algorithm	Neighbors' set size	Median Spearman's ρ
Raw co-occurrences baseline	5	0.26 ($p = 0.12$)
	10	0.15
	100	0.06
CBOW and Kendall's τ	5	0.25
	10	0.25
	100	**0.28** ($p = 0.09$)
CBOW and Relative Neighborhood Tree	5	0.20
	10	0.16
	100	0.14

Kendall's τ consistently renders better results without additional selection of 'important' associates by a relative neighborhood tree (additionally, it is much faster). This once again raises questions about whether vector models can be efficiently processed with graph representations. Kendall's τ also outperforms the baseline approach: the margin is as small as two points, but it is supported by higher significance ($p < 0.1$).

Note that qualitative analysis of the baseline results shows that they are mostly inappropriate for any practical task. For the time period which is described in the Table 1, the baseline approach almost does not reveal any differences between neighbors sets: average Kendall's τ is 0.92 for English and 0.99 for Russian. Thus, if in the case of English the earthquake event is at least detected (we observe the emergence of 4 new related neighbors), in the case of Russian the neighbor set remained strictly the same. It seems that the raw co-occurrences approach suffers from overestimating the influence of the reference corpora, which are much larger than the daily updates. Dynamic neural embedding models overcome this problem.

Interestingly, the wider sets of neighbors taken into account results in better performance only for CBOW with Kendall's τ. For the baseline and for CBOW with RNT, increasing the size of processed neighbor sets actually results in poorer performance. The reason for this behavior in RNT can be that the algorithm begins to 'roam' in the graph attracting more far-away associates as immediate tree neighbors to the target word. In the baseline method it simply leads to much language-dependent noise, which semantically aware models filter out at the training stage.

7 Conclusions

We presented a method of detecting semantic shifts for countries in news texts with the help of dynamic neural embedding models. We explored the difference between entities' vector representations in the models from different temporal stages and discovered association shifts that happen to these words over time. This can be employed to trace trends and events in streaming news texts using a completely unsupervised approach.

We showed that distributional semantic models are rather efficient when detecting associations shifts and are in most cases language-independent. In our test sets, there is a statistically significant correlation between lists of 'semantically shifted' countries in English and Russian sequences of models for the same time period.

However, there is still room for improvement. First of all, some ways to evaluate semantic shifts extraction have to be developed (including creation of ground truth datasets). Additionally, we plan to test other ways of comparing neighbor sets and tune algorithms' hyperparameters. It would be also useful to improve the quality of corpora (e.g. eliminate more noise and stop words). Finally, we plan to experiment with using different algorithms or parameter sets for different languages: preliminary tests show promising results.

References

[AGK01] James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal summaries of

new topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 10–18, New York, USA, 2001.

[BDV03] Yoshua Bengio, Rejean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[CGS15] Amaru Cuba Gyllensten and Magnus Sahlgren. Navigating the semantic horizon using relative neighborhood graphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2451–2460, Lisbon, Portugal, September 2015.

[HBB10] Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent dirichlet allocation. In *Neural Information Processing Systems 23*, pages 856–864, Vancouver, Canada, 2010.

[Jac01] Paul Jaccard. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines.* Rouge, 1901.

[JS09] David Jurgens and Keith Stevens. Event detection in blogs using temporal random indexing. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 9–16, Borovets, Bulgaria, 2009.

[KARPS15] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Florence, Italy, 2015.

[KCH+14] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, page 61, Baltimore, USA, 2014.

[Ken48] Maurice George Kendall. *Rank correlation methods.* Griffin, 1948.

[KNR15] Manika Kar, Sérgio Nunes, and Cristina Ribeiro. Summarization of changes in dynamic text collections using Latent Dirichlet Allocation model. *Information Processing & Management*, 51(6):809–833, 2015.

[KWHdR15] Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1191–1200, New York, NY, USA, 2015. ACM.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

[POL10] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

[Reh11] Radim Rehurek. *Scalability of semantic analysis in natural language processing.* PhD thesis, Masaryk University, 2011.

[Seg03] Ilya Segalovich. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, pages 273–280. Citeseer, 2003.

[Spe04] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[TKMS03] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 NAACL-HLT Conference-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[TP+10] Peter Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.