Visualizing Trends of Key Roles in News Articles

Chen Xia[1]*, **Haoxiang Zhang**[1]*, **Jacob Moghtader**[2], **Allen Wu**[2], **Kai-Wei Chang**[1]
[1]University of California Los Angeles, [2]Taboola
kasinxc@cs.ucla.edu; haoxiangzhx@gmail.com;
{jacob.m, allen.wu}@taboola.com; kw@kwchang.net

Abstract

There are tons of news articles generated every day reflecting the activities of key roles such as people, organizations and political parties. Analyzing these key roles allows us to understand the trends in news. In this paper, we present a demonstration system that visualizes the trend of key roles in news articles based on natural language processing techniques. Specifically, we apply a semantic role labeler and the dynamic word embedding technique to understand relationships between key roles in the news across different time periods and visualize the trends of key role and news topics change over time.

1 Introduction

Nowadays, numerous news articles describing different aspects of topics are flowing through the internet and media. Underneath the news flow, key roles including people and organizations interact with each other and involve in various events over time. With the overwhelmed information, extracting relations between key roles allows users to better understand what a key person is doing and how he/she is related to different news topics. To understand the action of key roles, we provide a semantic level analysis using semantic role labeling (SRL). To measure the trend of news topics, a word vector level analysis is supported using dynamic word embeddings.

In our system, we show that a semantic role labeller, which identifies subject, object, and verb in a sentence, provides a snapshot of news articles. Analyzing the change of verbs with fixed subject over time can track the actions of key roles. Besides, the relationships between subjects and objects reflect how key roles are involved in different events. We implemented the semantic role analyzer based on the SRL model in AllenNLP, which formulates a BIO tagging problem (He et al., 2017) and uses deep bidirectional LSTMs to label semantic roles (Gardner et al., 2018).

On the other hand, word embeddings map words to vectors such that the embedding space captures the semantic similarity between words. We apply dynamic word embeddings to analyze the temporal changes, and leverage these to study the trend of news related to a key role. For example, President Trump is involved in many news events; therefore, he is associated with various news topics. By analyzing the association between "Trump" and other entities in different periods, we can characterize news trends around him. For example, in February 2019, "Trump" participated in the North Korea-United States Summit in Hanoi, Vietnam. The word embedding trained on news articles around that time period identifies "Trump" is closely associated with "Kim Jun Un" (the President of North Korea) and "Vietnam" (the country hosted the summit).

We create a system based on two datasets collected by Taboola, a web advertising company. 1) *Trump dataset* contains 20,833 English news titles in late April to early July 2018. 2) *Newsroom dataset* contains approximately 3 million English news articles published in October 2018 to March 2019. The former provides a controllable experiment environment to study news related to President Donald Trump, and the second provides a comprehensive corpus covering wide ranges of news in the U.S. Source code of the demo is available at https://bit.ly/32f8k3t and more details are in (Zhang, 2019; Xia, 2019).

2 Related Work

Various systems to visualize the transition of topics in news articles have been published. Kawai et al. (2008) detected news sentiment and visu-

*Equal contribution.

Proceedings of the 2019 EMNLP and the 9th IJCNLP (System Demonstrations), pages 247–252
Hong Kong, China, November 3 – 7, 2019.

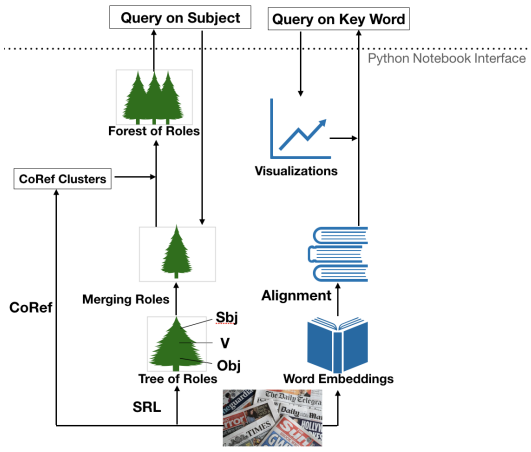

Figure 1: System Overview.

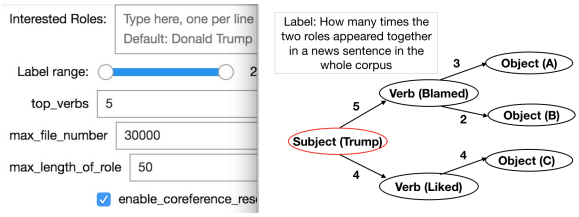

Figure 2: Tree Graph for Semantic Role Visualization.

alized them based on date and granularity such as city, prefecture, and country. Ishikawa and Hasegawa (2007) developed a system called T-Scroll (Trend/Topic-Scroll) to visualize the transition of topics extracted from news articles. Fitzpatrick et al. (2003) provided an interactive system called BreakingStory to visualize change in online news. Cui et al. (2010) introduced TextWheel to convey the dynamic natures of news streams. Feldman et al. (1998) introduced Trend Graphs for visualizing the evolution of concept relationships in large document collections. Unlike these works, our analysis focuses on the key roles in news articles. We extract semantic roles and word vectors from news articles to understand the action and visualize the trend of these key roles.

3 System Overview

To visualize the news trends, we apply semantic role analysis and word embedding techniques.

For semantic roles, we first construct a tree graph with subject as root, verbs as the first layer and objects as leaf nodes by extracting semantic roles with SRL (Gardner et al., 2018). Then we aggregate the tree graphs by collecting tree with the same subject and similar verb and object. Beyond applying simple string matching to identify same object and subject, we also apply a coreference resolution system (CoRef) to identify phrases refer to the same entity. As a result, we create a forest visualization where each tree represents the activities of a key role.

For word embeddings, we first train individual word vectors model for each month's data. However, there is no guarantee that coordinate axes of different models have similar latent semantics; therefore, we perform alignment algorithm to project all the word vectors into the same space. Once the embeddings are aligned, we are able to identify the shift of association between key roles and other news concepts based on their positions in the embedding space.

3.1 Visualization by Semantic Roles

Tree Graph for Semantic Roles We provide users with a search bar to explore roles of interest. For example, when searching for *Trump*, a tree graph is presented with *Trump* as root. The second layer of the tree is all of the verbs labeled together with subject *Trump*, e.g., *blamed* and *liked* in Figure 2. The edge label represents how many times two nodes, subject (e.g, *Trump*) and Verb (e.g., *liked*), appear together in a news sentence in the corpus. The edge label reflects the total number of semantic role combination in the given dataset, which depicts the importance of a news action.

Forest Graph for Semantic Roles In news articles, President Trump have different references, such as Donald Trump, the president of the United States, and pronoun "he" – a well-known task, called coreference resolution. When generating semantic trees, the system should not look only for *Trump* but also other references. To realize this, we preprocess the dataset with CoRef system (Lee et al., 2017) in AllenNLP (Gardner et al., 2018) and generate local coreference clusters for each news article. To obtain a global view, we merge the clusters across documents together until none of them shares a common role. A visualization demo for CoRef is also provided.

In Figure 3, the CoRef system clusters "*the Philladelphia Eagles*" with "*the Eagles*", and "*Hilary*" with "*Hilary Clinton*". The red nodes are center roles, which are representative phrases. For example, "*the Philladelphia Eagles*" and "*Hilary Clinton*" are the center roles of their corresponding cluster.

We use the following three rules to determine which phrases are center roles. If phrases are

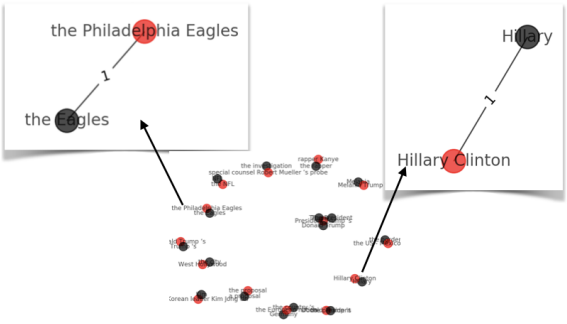

Figure 3: Coreference Resolution Clusters.

tied, the one with longest length will be selected: *LongestSpan* method selects the role with longest length. *WordNet* method marks spans not in the WordNet (Miller, 1998) as specific roles. *NameEntity* method marks roles in the name entity list generated by latent dirichlet allocation as specific ones. Both WordNet and NameEntity methods select the most frequent role as the center role.

Merging Algorithms for Semantic Roles Finally, we use the following rule-based approach to merge trees with same referent subject by CoRef.

1) Merging Objects with the Same Verb To better visualize the semantic roles, we merge objects with similar meaning if they are associated with same verb. To measure the similarity, we generate bag-of-word representations with *TF-IDF* scores for each object. If the cosine similarity between the representations of two objects is larger than a threshold, we merge the two nodes. We then sum up the frequency weights on the edges of all merging objects to form a new edge.

2) Merging Verbs with the Same Subject Verbs like *believe*, *say* and *think* convey similar meanings. Merging such verbs can emphasize the key activities of the key roles. The similarity between verbs associated with the same subject is calculated by cosine similarity between word vectors using word2vec (Mikolov et al., 2013). In particular, we merge two verbs if their cosine similarity is larger than a threshold. By showing a certain range of edge labels, the system is also capable of filtering out verbs with extreme high or low frequency such as *say*, as these verbs carry less meaningful information.

Modifier, Negative and Lemmatization While our news analysis is mainly based on subject-verb-object relations, we also consider other semantic roles identified by the SRL model. For example, we include identification of modifier so that we can recognize the difference between "resign" and "might resign". We also add negation as an extra sentiment information. Verbs have different forms and tenses (e.g., win, won, winning). If we merge all verbs with the same root form, we can obtain a larger clusters and reduce duplicated trees. However, for some analysis, the tense of verbs are important. Therefore, we provide Lemmatizating as an option in our system.

3.2 Dynamic Word Embeddings

Dynamic word embeddings model align word embeddings trained on corpora collected in different time periods (Hamilton et al., 2016). It divides data into time slices and obtains the word vector representations of each time slice separately. To capture how the trends in news change monthly, we train a word2vec word embedding model on news articles collected in each month. We then apply the orthogonal Procrustes to align the embeddings from different time periods by learning a transformation $\mathbf{R}^{(t)} \in R^{d \times d}$:

$$\mathbf{R}^{(t)} = \arg\min_{\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}} \|\mathbf{W}^{(t)}\mathbf{Q} - \mathbf{W}^{(t+1)}\|,$$

where $W^{(t)} \in R^{d \times V}$ is the learned word embeddings of each month t (d is the dimension of word vector, and V is the size of vocabulary).

N-Gram To represent named entities such as 'white house' in the word embeddings, we treat phrases in news articles as single words. The max length of phrases is set as 4 to avoid large vocabulary size.

Absolute Drift Inspired by Rudolph and Blei (2018), we define a metric that is suitable to detect which words fluctuate the most relative to the key word w_k. Denote $cos(w_k, w_i, t)$ as the cosine similarities between the word w_i and the key word w_k at time t. For top n words close to w_k, calculate the absolute drift of each word w_i by summing the cosine similarity differences.

$$drift(w_i) = \sum_{t=2}^{T} |\cos(w_k, w_i, t) - \cos(w_k, w_i, t-1)|$$

After finding meaningful words that fluctuate the most, cosine similarities between these words and w_k of each month can be plotted to present possible useful interpretations.

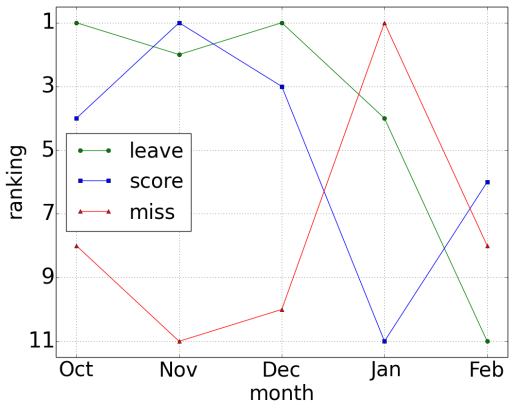

Figure 4: Action Tracking for *LeBron James*

rank	verbs for *LeBron James*	fixed main objects
1	miss	games
2	suffer	a groin strain injury
3	make	no fixed main objects
4	leave	Cleveland Cavaliers
5	lead	the team

Table 1: Verb Rankings for *LeBron James* in January

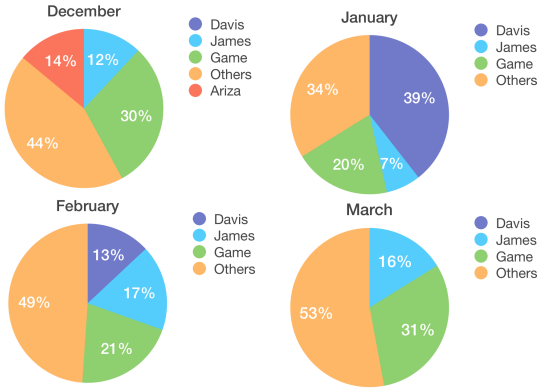

Figure 5: Breaking News Tracking on Trade Rumors.

4 Case Studies

4.1 Semantic Roles

Action Tracking on Verbs We apply semantic role labelling model to *newsroom dataset* collected from October 2018 to February 2019 on taxonomy: */sports/basketball* and search for subject *LeBron James*, a basketball player.

For each month, we generate the top frequent verbs from sentences where *LeBron James* is marked as the subject. We found that the top verbs include "Leave", "Score" and "Miss". Example sentences include: "LeBron James **leave** the Cleveland Cavaliers", "LeBron James **score** points" and "LeBron James **miss** games".

We further show the ranking of these verbs in different months in Figure 4. As results show the verb "leave" ranks at the top around October due to an earlier announcement that Lebron James will leave the Cavaliers. However, the frequency falls in January.

Meanwhile, news on *LeBron James **miss** games* ranked first and the verb "score" doesn't co-occur with LeBron James in January due to his injury.

To explain the absence, we list the top 5 frequent verbs are listed below. Verbs that occur with LeBron James only in December and January are colored in red.

From this analysis, we can see that *LeBron James* was suffering the groin strain injury in January, causing his absence of the game.

Breaking News Tracking on Objects We run our algorithm to analyze news article under the topic: */sports/basketball*, which has 75,827 peices of news title descriptions. We search *Lakers* as subject in every month and sum up all the label weights on the edges between verb and object.

$$W(V, o|S = s) = \sum_{v \in V} W(v, o|S = s), \quad (1)$$

where $W(v, o|S = s)$ denotes the weight on edges between all the verbs $v \in V$ and a specific object o under certain subject s.

We rank all objects based on Eq. (1) and the top 5 objects associated with the subject "Lakers" are: "Davis", "James", "Game", "Ariza", and "Others". We further show the pie chart to demonstrate the percentage of each object associated with "Lakers" in different months.

The purple part in Figure 5 shows that the number of news mentioning *Anthony Davis* and *Lakers* suddenly emerged and even beat *James* and *Lakers* in January but gradually decreased in February. The breaking news about Anthony and Lakers disappeared completely in March. The event happened in January and February was the trade rumors on *Davis*. After the trade deadlines, the topic eventually disappeared.

4.2 Dynamic Word Embeddings

2D Visualization The t-SNE embedding method (Maaten and Hinton, 2008) is used to visualize the word embeddings in two dimensions. First, given

Rank	Dec 2018	Jan 2019	Feb 2019	Mar 2019
1	los_angeles_lakers	los_angeles_lakers	los_angeles_lakers	los_angeles_lakers
2	lebron_james	**pelicans**	lebron_james	lebron_james
3	lonzo_ball	lebron_james	clippers	clippers
4	clippers	lonzo_ball	**pelicans**	kevin_durant
5	brandon_ingram	**anthony_davis**	boston_celtics	lonzo_ball
6	kevin_durant	cavs	kyle_kuzma	lebron
7	**anthony_davis**	boston_celtics	tobias_harris	giannis_antetokounmpo
8	raptors	rockets	**anthony_davis**	magic_johnson

Table 2: Top 5 Words closest to the Word 'lakers' in Each Month.

a word w that we are interested in, the nearest neighbors of w at different time periods are put together. Next, the t-SNE embeddings of these word vectors are calculated and visualized in a 2D plot.

Figure 6: Shifts of the Word 'Max'.

On March 10 2019, the Boeing 737 MAX 8 aircraft crashed shortly after takeoff. After this fatal crash, aviation authorities around the world grounded the Boeing 737 MAX series. Figure 6 shows that dynamic word embeddings capture this sudden trend change. In particular, before March 2019 (from when the 'max_Mar19' embedding is obtained), the word 'max' was close to different people names. When the crash happened or afterwards, the word 'max' immediately shifts to words such as 'boeing', '737' and 'grounding'.

Top Nearest Nighbors Listing the top nearest neighbors (words that have highest cosine similarities with the key word) of the key word w inside a table also shows some interesting results. For example, Table 2 confirms with Figure 5 that breaking news of *Anthony Davis* and *Lakers* happened because of the trade rumors.

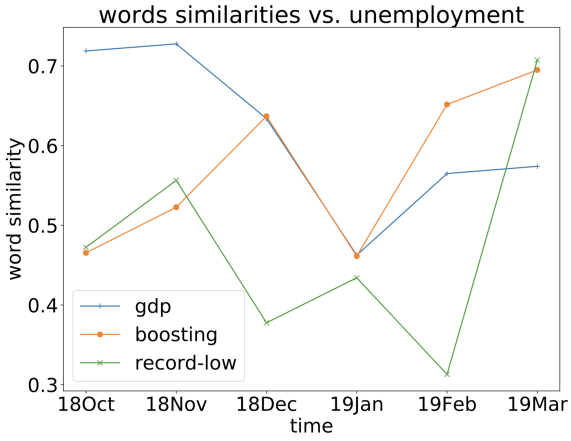

Figure 7: Cosine Similarities with 'Unemployment'.

Changing Words with Absolute Drift Figure 7 displays the cosine similarity changes with respect to 'unemployment'. One thing we can infer from this figure is that as the economy ('gdp') shows a strong signal ('boosting') in the first quarter of 2019, the unemployment rate reaches a 'record-low' position. According to National Public Radio, the first quarter's gross domestic product of U.S. grew at an annual rate of 3.2%, which is a strong improvement compared to the 2.2% at the end of last year. In addition, the Labor Department reported that 196,000 jobs were added in March, and the unemployment is near 50-year lows.

5 Conclusion

We presented a visualization system for analyzing news trends by applying semantic roles and word embeddings. We demonstrated that our system can track actions and breaking news. It can also detect meaningful words that change the most. Fu-

ture work will focus on adding entity linking to subjects, providing more semantic roles information. Also, we plan to work on qualitative assessment on the quality of the trends and other word embedding models like Glove(Pennington et al., 2014).

6 Acknowledgment

This work was supported in part by a gift grant from Taboola. We acknowledge feedback from anonymous reviewers and fruitful discussions with the Taboola team at Los Angeles.

References

Weiwei Cui, Hong Zhou, Huamin Qu, Wenbin Zhang, and Steven Skiena. 2010. A dynamic visual interface for news stream analysis. In *Proceedings of the first international workshop on Intelligent visual interfaces for text analysis*, pages 5–8. ACM.

Ronen Feldman, Yonatan Aumann, Amir Zilberstein, and Yaron Ben-Yehuda. 1998. Trend graphs: Visualizing the evolution of concept relationships in large document collections. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 38–46. Springer.

Jean Anne Fitzpatrick, James Reffell, and Moryma Aydelott. 2003. Breakingstory: visualizing change in online news. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, pages 900–901. ACM.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

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

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

Yoshiharu Ishikawa and Mikine Hasegawa. 2007. T-scroll: Visualizing trends in a time-series of documents for interactive user exploration. In *International Conference on Theory and Practice of Digital Libraries*, pages 235–246. Springer.

Yukiko Kawai, Yusuke Fujita, Tadahiko Kumamoto, Jianwei Jianwei, and Katsumi Tanaka. 2008. Using a sentiment map for visualizing credibility of news sites on the web. In *Proceedings of the 2nd ACM workshop on Information credibility on the web*, pages 53–58. ACM.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

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

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011.

Chen Xia. 2019. Extracting global entities information from news. Master's thesis, University of California, Los Angeles, California, US, 6.

Haoxiang Zhang. 2019. Dynamic word embedding for news analysis. Master's thesis, University of California, Los Angeles, California, US, 6.