Ontology Guided Purposive News Retrieval and Presentation

Abir Naskar
TCS Innovation Lab, India
abir.naskar@tcs.com

Rupsa Saha
TCS Innovation Lab, India
rupsa.s@tcs.com

Tirthankar Dasgupta
TCS Innovation Lab, India
dasgupta.tirthankar@tcs.com

Lipika Dey
TCS Innovation Lab, India
lipika.dey@tcs.com

Abstract

In this paper, we present a purposive News information retrieval and presentation system that curates information from News articles collected from multiple trusted sources for a given domain. A back-end domain ontology provides details about the concepts and relations of interest. We propose an attention based CNN-BiLSTM model to classify sentence tokens as ontology concepts or entities of interest. These entities are then curated and used to link articles to illustrate evolution of events over time and regions. Working systems are initiated with small annotated data sets which are later augmented with humans in the loop. It is easily customizable for various domains.

1 Introduction

News consumption is no more restricted to consuming a set of facts dished out by a specific agency. Readers are not only choosing the type of content they want to read but also how. Increasing interest in social statistics is also seeing News as a source of data for generating these statistics. Unlike social media, News from trusted sources is reliable. News presentation is therefore undergoing a sea-change. Along with a bird's eye view of global events, the ability to delve deep down into specific stories along various dimensions and also watch their evolution is necessary to enable systematic studies.

In this paper, we present an Ontology-guided News information retrieval and presentation system. The uniqueness of the proposed system lies in the use of a back-end domain ontology that specifies the entities and relations of interest in a domain, based on which, information components are extracted and classified using a deep neural network architecture. These concepts are used to create domain-specific "purposive indices" to aid concept-oriented Information retrieval rather than simple word-based retrieval.

A seed ontology of concepts along with a few instances of each concept is used to create annotated data to train a concept classification model. This is applied over a larger set of articles, the result of which is then validated through human evaluation and used subsequently to enhance the initial model. It is found that the proposed method takes much less time and effort to create annotated data sets for all situations that lack large labeled data needed to exploit deep-learning methods. Information components extracted from the News articles, are stored in indexed repositories for downstream analytics. The results are presented to the end-user through an innovative interactive interface that helps in consuming information at multiple levels of granularity.

2 Overview of proposed News Retrieval and presentation framework

Figure 1 presents an overview of the proposed News Information retrieval system. A number of News crawlers are deployed to collect News from dedicated and reliable sources. For each article its meta-data like

In: A. Aker, D. Albakour, A. Barrón-Cedeño, S. Dori-Hacohen, M. Martinez, J. Stray, S. Tippmann (eds.): Proceedings of the NewsIR'19 Workshop at SIGIR, Paris, France, 25-July-2019, published at http://ceur-ws.org

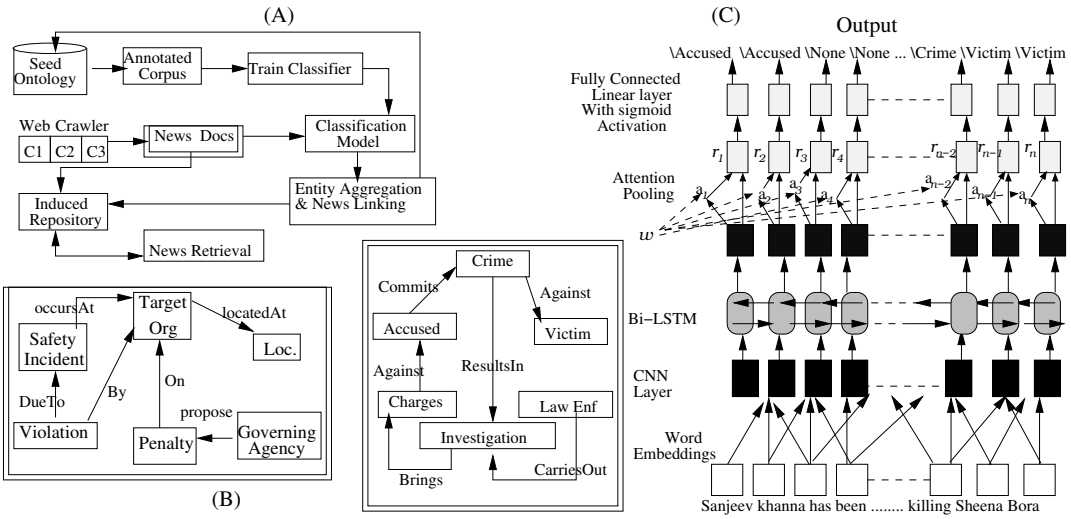

Figure 1: Overview of the Event and Entity Extraction Architecture(A). (B) depicts the example seed ontology structure. (C) explains the CRNN architecture.

source, date and time of publication, location, headline is stored along with the full article. These are all indexed using Solr. Exact or near duplicate articles are identified using Latent Semantic Hashing (LSH). Though only one copy is stored in the repository, total count for such articles are maintained. Each article is then passed through an information extraction and classification pipe-line that deploys component classifiers to detect ontology-components in sentences. The extracted elements are resolved and then used to create additional "purposive named-indices" for the documents. These elements are also used to link threads of the same News together.

The ontology is composed of a schema that contains different domain *concepts* (C) and a set of generic relations R between these concepts. Figure 1 illustrates a pair of ontologies for two different domains, namely, *crime* and *occupational health and safety*. Given an ontology that elucidates the basic components of a domain and their underlying relationships in a generic way, finding similar information components from vast collections of text still remains a challenging task, since the manifestation of these components in natural language can be extremely varied. For example, though the names of criminal and victim can be extracted as named-entities from text, establishing their roles unambiguously is not a simple task. Classifying an instance of a concept correctly requires deep contextual analysis. We propose the use of an attention-pooling convolutional Bi-LSTM neural network based architecture to do the task. Details of this is given in the next subsection.

Purposive indices for news articles are created using the ontology concepts extracted from text. For example crime News articles are indexed by criminal and victim names, section names, location of crime etc. News articles reporting safety incidents are similarly additionally indexed by incident names, location of incident, penalty incurred in different currency etc. Linking of related articles are done through the purposive indices only. It may be noted that all concepts for the same event may be obtained at once. New concepts can get associated as a single story unfolds over time. Similarly multiple isolated articles may get linked to each other at later stages. The purposive indices are also used to generate comparative and aggregate statistics over various dimensions. The visualization module enables the end-user to view News articles at various levels of granularity.

2.1 Ontology-guided Concept and Entity Detection using C-RNN Network

Convolutional neural networks (CNN) exploit local dependencies, while recurrent networks like Long Short Term Memory (LSTM) capture long-distance dependencies among features. The proposed Convolutional Bi-LSTM (C-RNN) model combines both the capabilities. For a given sentence, the network learns to assign ontology concept labels to each word. The input to the network is a sequence of word embeddings with 100 dimensions each.

A convolutional layer is first used to extract local n-gram features. All word embeddings are concatenated to form an embedding matrix $M \in R^{dX|V|}$. Where, $|V|$ is the vocabulary size and d is the embedding dimension. The matrix is divided into k regions. In each region, we apply convolution function represented by

$$Conv(x_{i:w}^{-}) = W.(x_{i:w}^{-}) + b \quad (1)$$

to calculate the output features. Where, W and b

are the weights that the network learns. We apply the same convolution operation repeatedly over the different matrix regions to get multiple output feature vectors. The output of the CNN layer is passed to the bidirectional LSTMs, which read it both backward and forward to take care of dependencies on the past neighbours as well as future long-distance dependencies. The Bi-LSTM layer is followed by an attention pooling layer over the sentence representations. Attention modules have been proved to boost accuracy for tasks like sentiment or activity detection by learning to focus more on certain linguistic elements over others, without increasing computational complexity. In our case, attention pooling achieves higher accuracy by learning the specific characteristics surrounding each concept in the form of weights associated to the output of the Bi-LSTM layer. This is represented as:

$$a_i = \tanh(W_a.h_i + b_a), \tag{2}$$

$$\alpha_i = \frac{e^{w_\alpha.a_i}}{\sum e^{w_\alpha.a_i}}, \tag{3}$$

$$O = \sum(\alpha_i.h_i). \tag{4}$$

Where W_a, w_α are weight matrix and vector respectively, b_a is the bias vector, a_i is attention vector for i-th sentence, and α_i is the attention weight of i-th sentence. The output of the attention layer is then passed to a fully connected linear layer with sigmoid activation. The mapping of the linear layer after applying the sigmoid activation function is given by

$$y = s(x) = sigmoid(w.x + b). \tag{5}$$

Where, x is the input vector, w is the weight vector, and b is bias value. Finally, the loss function is computed using the cross-entropy loss defined by

$$L = -\sum_{i=1}^{2} \bar{y_i} log(y_i). \tag{6}$$

Where $\bar{y}$ is the one-hot representation of the actual label for the input word. To avoid over-fitting, we apply dropout technique at each layer to regularize our model.

The model is initially built from a small annotated corpus, in which the instances of the ontology concepts are tagged by their respective labels. This model, when applied over a larger corpus yields new instances of each label, which are evaluated by humans and then accepted for next-level training if two out of three annotators simultaneously agree on the label. For repeated experiments on different domains, the inter-annotator agreement is found to be around 0.65, which is pretty high. This can be done multiple times, though we have restricted it to two times only.

As discussed earlier, classification of concepts is more complex than merely identifying named entities. It is imperative to also recognize the role played by the entity. Two such example sentences are presented here along with the concepts extracted -

Sanjeev Khanna has been taken into custody in Kolkata on charges of killing Sheena Bora - Concepts extracted are <Criminal, Sanjeev Khanna> and <Victim, Sheena Bora>. Kolkata is not labeled as any concept, correctly.

US Department of Labor's fines Heat Seal $95,000 for 15 health violations Concepts extracted - <Company, Heat Seal> and <Penalty, $95,000>.

2.2 Linking Articles to Indicate News Story Evolution

A link between two News articles is created only if they share purposive indices for specific concept classes. For example, for crime incidents, victim names and criminal names should overlap, while for safety incidents organization name and safety incident type should overlap.

The first challenge comes in the form of entity resolution since named entities are spelled differently in different sources. An edit distance based measure [MV93] is used to compare different entities and combine them if sufficiently similar. For example, an individual was variably referred to as "Mukherjee", "Mukerjea" and "Mookerjee" across various sources in our crime news database.

The second challenge comes from the fact that the sets of conceptual instances for a single story also evolve over time. For example, it is observed that for a long-drawn crime incident, new articles report new names as criminals or even victims, as new information pours in. A concept overlap threshold of 80% for specified types is used to link the articles into a single story.

The third challenge is due to the fact that mere overlap of the names of entities is not enough, even their corresponding roles need to be same, or at least similar, for us to consider them to refer to the same case.

Considering the above challenges, we propose a weighted similarity computation to determine the similarity of two articles. Highest weightage is given for candidates that are resolved to be similar and also belong to the same concept class. Candidates which are similar after resolution but are identified as instances of different classes are given a lower weightage. The final similarity measure is computed as a weighted sum of all candidate similarities. Two articles are considered similar if the similarity is above a user-defined threshold.

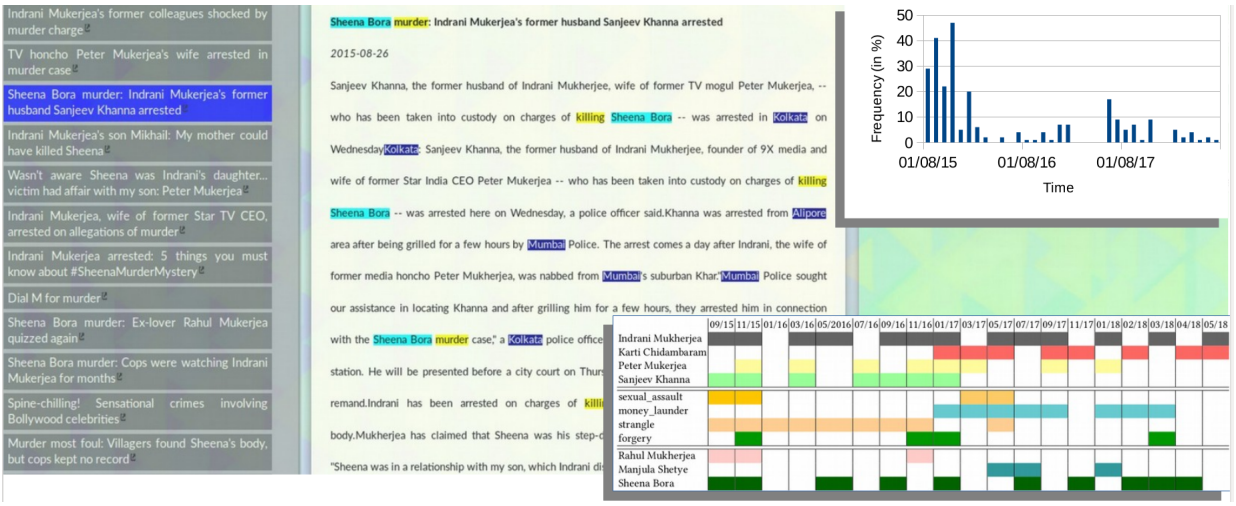

Figure 2: Illustration of the linked News articles. Selecting one article shows the extracted events and entities. Graph at the top right shows the distribution of a News over time. Figure at the bottom right shows evolution of entity types over time.

Once conceptually similar pairs of articles are found, a virtual group is created containing these along with a union of the conceptual entities contained in them. More articles are added to these by repeating the similarity computation of each article with an existing group. The system automatically links them and maintains them chronologically.

2.3 News Retrieval and Data Visualization

Apart from regular entity and concept based retrieval and tracking the evolution of a News, the proposed system also enables the user to explore the evolution of an incident across different dimensions like time and space. Figure 2 illustrates how News evolution is presented. The left most panel shows a series of crime incidents reporting an unsolved murder case gathered from across different time and sources. These have been linked together by the earlier algorithm using the purposive indices extracted by the c_RNN classifier. On selecting a particular crime news, the extracted entities and events are shown in the middle panel. The right top chart shows the temporal distribution of the reports over the period 2015 to 2018. The bottom right chart displays how crime entities have changed over time. While *Murder* is prevalent over the entire time-line, new crime incidents like *money laundering* has emerged as reported co-crimes at later stages. One can also explore presence or absence of similar incidents across different geographical regions. It can unearth regional affinities for certain kinds of acts. Visualizations also help in studying aberrations like different charges evoked for similar crimes or variability in penalty rates for similar safety incidents through canned analytics.

3 Experiments and Results

3.1 Data Collection

We have conducted experiments for two different domains using the ontology pair described earlier. Our collection consists of around 12000 crime-related news collected from the top 3 English news sources from each of four regions in the Indian subcontinent (north, south, east and west). The Occupational Health and Safety database has been created from approximately 4000 articles published by Occupational Safety and Health Administration (OSHA), United States Department of Labor, each of which detail various transgressions by organizations, and the actions taken against them.

3.2 Experiments

Each dataset is divided into 70%, 20% and 10% for training, validation and testing respectively. A Condition Random Field (CRF) model is trained as the baseline. This model uses part-of-speech(POS) and N-grams as features. Additionally, a number of other deep neural network based models such as BiLSTM, BILSTM with mean over time (MoT), BILSTM with Attention network, Convolution network (CNN), CNN with BILSTM+MoT were used to compare the results with the proposed CNN with BILSTM along with Attention Network. Each model is trained with three types of word embeddings GloVe(G), Word2Vec(w2v) and combination of GloVe and Word2Vec(G+w2v). Both w2v and a combination of Glove and w2v achieve similar performance for the proposed architecture, which is significantly better than the baseline and also others. F1 score for all the models are shown in Table 1.

Table 1: Results demonstrating F1 Scores for each model corresponding to two domain, *Crime* and *OSHA*

	Crime			OSHA		
	G	W2V	G+W2V	G	W2V	G+W2V
BiLSTM	67	70	64	64	65	70
BiLSTM-MoT	66	69	68	63	66	72
CNN	69	72	68	67	67	63
BiLSTM-att	71	71	70	68	69	64
CNN+BiLSTM-MoT	72	75	75	69	70	67
CNN+BiLSTM-att	74	76	**76**	70	71	**73**
CRF		58			61	

3.3 Hyper-parameters

For CNN, we keep the window size as 3 and number of filters as 30. For BiLSTM, state size is 200 with initial state value =0.0. We use a dropout rate of 0.05, batch size=10, initial learning rate of 0.01, decay rate of 0.05 and gradiant clipping of 5.0.

3.4 Results

Throughout all the target classes, the performance of the CNN-BiLSTM model has been found to be better than the others. The performance of combined local and global embedding word2vec method for learning word embeddings [MCCD13] have been observed to be very effective in capturing solely contextual information. It has also been observed that, combining both the W2V and GloVe embeddings surpasses the performance of models using the individual embeddings. Overall, the performance of the CNN-BiLSTM-att model along with combined W2V-GloVe embedding is higher than the rest of the existing models.

4 Related Works

While there has been a growing body of research in extracting structured information from texts, neural network based ontology guided News event extraction and story evolution is still in its nascent stage. Most of the existing methods are limited to event and named entity extractions and not into identifying granular level of entity role identification. Supervised learning with different flavours of LSTM or CNN [GCW$^+$16, MB16] are used for entity classification. Distant supervision involving some amount of annotated data and an initial knowledge source has been proposed to develop models in [ZNL$^+$09, NZRS12, CBK$^+$10, MBSJ09, SSW09, NTW11]. The unsupervised approach requires hand crafted rules pertaining to the information to be extracted [Hea92, SW13, JVSS98, HZW10, MB05].

5 Conclusion

In this paper we have proposed an Ontology-guided News information retrieval system using a Convolutional Bi-LSTM Network for Concept detection. Concept-based linking is used to link related articles to present News evolution and event distribution across regions. We have also illustrated how deep learning methods can be deployed with small volumes of annotated data. We intend to extend the proposed methods to work with all kinds of legal documents and also incorporate predictive technologies to predict activities or events.

References

[CBK$^+$10] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3. Atlanta, 2010.

[GCW$^+$16] Jiang Guo, Wanxiang Che, Haifeng Wang, Ting Liu, and Jun Xu. A unified architecture for semantic role labeling and relation classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1264–1274, 2016.

[Hea92] Marti A Hearst. Direction-based text interpretation as an information access refinement. *Text-based intelligent systems: current research and practice in information extraction and retrieval*, pages 257–274, 1992.

[HZW10] Raphael Hoffmann, Congle Zhang, and Daniel S Weld. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 286–295. Association for Computational Linguistics, 2010.

[JVSS98] Yaochu Jin, Werner Von Seelen, and Bernhard Sendhoff. An approach to rule-based knowledge extraction. In *Fuzzy Systems Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998*

IEEE International Conference on, volume 2, pages 1188–1193. IEEE, 1998.

[MB05] Raymond J Mooney and Razvan Bunescu. Mining knowledge from text using information extraction. *ACM SIGKDD explorations newsletter*, 7(1):3–10, 2005.

[MB16] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*, 2016.

[MBSJ09] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[MV93] Andres Marzal and Enrique Vidal. Computation of normalized edit distance and applications. *IEEE transactions on pattern analysis and machine intelligence*, 15(9):926–932, 1993.

[NTW11] Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. Scalable knowledge harvesting with high precision and high recall. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 227–236. ACM, 2011.

[NZRS12] Feng Niu, Ce Zhang, Christopher Ré, and Jude Shavlik. Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 8(3):42–73, 2012.

[SSW09] Fabian M Suchanek, Mauro Sozio, and Gerhard Weikum. Sofie: a self-organizing framework for information extraction. In *Proceedings of the 18th international conference on World wide web*, pages 631–640. ACM, 2009.

[SW13] Fabian Suchanek and Gerhard Weikum. Knowledge harvesting in the big-data era. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 933–938. ACM, 2013.

[ZNL+09] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. Statsnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th international conference on World wide web*, pages 101–110. ACM, 2009.