Extracting a Causal Network of News Topics

Eduardo Jacobo Miranda Ackerman

Computer Networks, Dresden University of Technology,
Nöthnitzer Str. 46, Dresden, Germany
eduardo.miranda@inf.tu-dresden.de

Abstract. Because of the abundance of online news, it is impossible for users to process all the available information. Tools are needed to help process this information. To mitigate this challenge we propose generating a network of causally related news topics to help the user understand and navigate throughout the news. We assume that by providing the causes or effects of a news topics, the user will be able to relate current news to past news topics that the user knows about, or that the user will discover past news topics as currently relevant. Also, the additional context will facilitate the understanding of the current news topic.

To generate the causal network, information is extracted from several distributed news sources while maintaining important journalistic features such as source referencing and author attribution. We propose ranking different causes of an event, to provide a more intuitive summary of multiple causal relations.

To make the network easily understandable, news topics must be represented in a format that can be causally related therefore, a news topic model is proposed. The model is based on the phrases used by online news sources to describe an event or activities, during a limited timeframe. To maintain usability the results must be provided in a timely manner from streaming sources and in an easy to understand format.

Keywords: Topic Detection, Causal Relation Extraction, Information Overload, Automatic News Organization.

1 Introduction

The overwhelming amount of online news presents a challenge called news information overload, to mitigate this challenge we propose a system to generate a causal network of news topics. The causal network has some advantages over term-overlap clustering techniques. First, the news topics are put into a context the user may already know, thus facilitating understanding. If the causally related topics are unknown to the user, these novel topic may now be of interest, therefore novel interesting material may be presented. Also, it may be easier for a user to navigate the news on a causally related network instead of a traditional list-column layout.

To generate the aforementioned network several components are necessary. These include a component to extract causal relations from news article text and defining a news topic in a way that can be discovered in causally-related

P. Herrero et al. (Eds.): OTM 2012 Workshops, LNCS 7567, pp. 33–42, 2012.
© Springer-Verlag Berlin Heidelberg 2012

text. All this while maintaining journalistic features such as source reference and author attribution, to allow the user to verify the information.

To illustrate the proposed system, consider a user interested in reading the news. The user finds a novel news topic, namely *The Dodd-Frank Act*. To get a quick intuition about the topic the user generated a causal network, with the proposed system, for *The Dodd-Frank Act*. The generated causal network shows that *The Dodd-Frank Act* was created in response to *The Late-2000s Financial Crisis*, that was in turn a cause by *The Subprime Mortgage Crisis*. Because the user already knows about some of the causally related topics beforehand, the novel topic gets an interesting context. The causally related topics the user does not know about have already past, but now they become newly interesting, because of the novel context. For the above illustration the causal chain of news topics is as follows: *The Subprime Mortgage Crisis* caused *The Late-2000s Financial Crisis* caused *The Dodd-Frank Act*.

2 Related Works

To generate a causal chain we need to extract this information from natural text such as news articles. To present this information in a clear and understandable format it is necessary to find a topic representation that is coherent within a causal relation, a topic such as *Sports* is often too coarse grained to provide a coherent causal relation. Other topic representations, such as term vectors, are also not well suited because a causal relation between vectors is not readily understandable. In this section we present three types of works: Works that mitigate news information overload by generating a network of news, works that extract causal relations from natural text, and works that generate topic models.

2.1 Network of News

Incident Threading for News Passages [1], is a work that builds upon Topic Detection and Tracking [21]. In it they present a method of generating a network of events about a single news topic. The network is generated by clustering passages at different thresholds, one threshold for clustering passages and a lower one for linking clusters. The links are based on term overlap, though they are given a direction based on temporal features. A work that focuses on causal relations is Causal Network Construction to Support Understanding of News [2]. In this work causal relations are extracted based on "clue phrases" in Japanese such as "Tame" (translated: for the sake of), and the topics are represented by keywords extracted from the title or body of news article. A dictionary of Japanese case frames is used to extract the keywords and causal phrases, a relevance metric is given to the keyword vector that represents a topic, this metric is used to discard causal relations thus reducing the complexity of the resulting graph. The keyword vector that represents a topic is not intuitively coherent therefore the results may hinder understanding. To find a focus on coherence we review Connecting the Dots Between News Articles [3]. It presents

a system that generates a coherent chain of stories given an initial and final story. The system is compared to a baseline system as well as Google Timeline and Event Threading [4], a precursor of Incident Treading [1]. The system does not generate a network per se, but it does link news articles as though they were news topics. Another work that generates chains of news topics is Topic Chains for Understanding a News Corpus [5], in this work Latent Dirichlet Allocation (LDA) [6] is used to generate topics and several techniques are used to cluster topics, clusters are segregated into different time periods and links are generated between similar clusters from consecutive time periods, the result is a sequence of news topics that change over time. Other systems such as Unified Analysis of Streaming News [7] focus on scalability aspects of generating links between news topics.

2.2 Causal Relation Extraction

To better explain causal relation extraction, we classify the works into two types: explicitly causal and implicitly causal. Explicit causation is characterized by a causal marker in the phrase, for example "because" in the sentence "he fell because it was wet". An implicit causal relation may not have the causal markers for example "he fell and went to the hospital". We focus on explicit causation because it is less ambiguous therefore easier to understand. An example of explicit causal relation extraction is given [2], another example is given in [13], where causal taxonomies [14,15] are used to define causal markers and patterns, to extract causal relations.

There are several approaches to finding implicit causal relations. In [12] verb pairs are assigned a probability of being causally related, for example a sentence that contains the verbs "slip-fall" is likely a causal sentence. Noun pairs can also be used to extract causation, in [16] the noun pair "HIV-AIDS" is used to extract causal markers. There are explicit-implicit hybrid approaches [17], and there are also methods that extract several types of semantic relations, such as [18,19]. We propose to rank causal relation between news topics, by comparing occurrence probabilities similar to [12], using the temporal order features of news topics instead of screenplays. The proposed approach would be comparable to the method presented in [20]. And the results would be useful for a user to asses the validity of the extracted information.

2.3 Topic Models

Many works represent a topic as a weighted term vector, a popular model is Latent Dirichlet Allocation (LDA) [6], this method is useful to classify a large collection of documents into a set of topics. The top ranking words in a topic vector provide an intuition about the topic concept, but the vector format is not well suited for causal relations. A more comprehensive topic model has been proposed for the task of automatic multi-document summarization. In Topic Themes for Multi-document Summarization [8] phrases are selected based on LDA and part-of-speech-patterns, the phrases are used to represent a topic and

key information about the topic. One drawback of this approach is that it does not consider multiple phrases referring to the same topic. Having a collection of phrases that represent a single news topic makes it possible to reduce the result set of a network of news topics, thus making it easier for the user to understand.

3 Research Hypotheses

The proposed system is based on several hypotheses:

Hypothesis 1. Causal relations between news topics help the user understand the news. We know that clustering news articles into topics reduces the result set thus reducing the cognitive load for selecting a news topic and we know that other semantic relations, such as temporal order, also known as a timeline, add information but still facilitate understanding, we must evaluate if the additional information from causal relations also facilitates understanding.

Hypothesis 2. Causal relations between news topics can be coherently extracted from a single phrase or sentence. This is important to provide understandable results to the user, to be able to directly reference the source of the information and to reduce processing requirements.

Hypothesis 3. Because of the abundance of online news sources, causal relation information is available for almost all popular news topics. Although we know there is an increasing amount of online news, we need to define the scope of news topics for which causal relation information is generated, namely news topics for which there is an abundance of information.

In order to evaluate these hypotheses, a system is proposed that will automatically generate a causal network of news topics. For the system to function properly the gaps presented in the related works section need to be filled, these are: The development of a topic model that is well suited for establishing causal relation and a method to rank explicit causal relations between news topics. The technical contributions of this work will be the methods to fill the aforementioned gaps and the proposed system to mitigate information overload.

4 Material

Currently an experimental prototype is under development, it uses the Palladian toolkit[1], to extract information from online news article accessed trough a web search engine. The system can use a static collection of documents as a corpus, for example the Reuters corpus[2] with access via a search engine such as Lucene[3]. The online version is preferred because it reduces local processing requirements.

[1] `http://palladian.ws/` accessed 1.6.2012
[2] `http://about.reuters.com/researchandstandards/corpus/` accessed 1.6.2012
[3] `http://lucene.apache.org/core/` accessed 1.6.2012

Redirect link information is extracted from Wikipedia with the help of JWPL[4] a library that provides structured access to Wikipedia.

To better understand the system architecture we first introduce the notion of news topic references (NTR). These are the phrases used to refer to a news topic, for example in a Wikipedia article about a news topic such as *"The Arab Spring"*, the title of that article is a NTR. News topic references are often found in the title of news articles or in the initial part of the body where a summary is provided, for example the title "After the Arab spring" or in the body "...As the Arab Spring remakes the...". We can also find multiple NTR that refer to the same news topic. For example the article titled "Tunisia Effect" in Wikipedia is a redirect page to the "Arab Spring", we can assume that both "Tunisia Effect" and "Arab Spring" refer to the same news topic. A feature of NTRs is that they are coherent in a causal relation, for example in the title "Arab Spring Causes Bankruptcy of Russia's Top Fruit Importer", two NTRs are causally related "Arab Spring" and "Bankruptcy of Russia's Top Fruit Importer", this example shows how different news topics can be causally related. Because most of the information is extracted from online sources, there are limitations to the information the system can provide, namely information for poorly documented news topics and distinction between opinion and fact.

5 Methods

To illustrate how the prototype system works we present Figure 1 and explain each step. In the figure the term "Alias", found in **Step 2, 8 and 9**, is used to refer to the collection of NTRs that refer to the same news topic, so the NTR "Arab Spring" is an alias of "Tunisia Effect" because both refer to the same news topic.

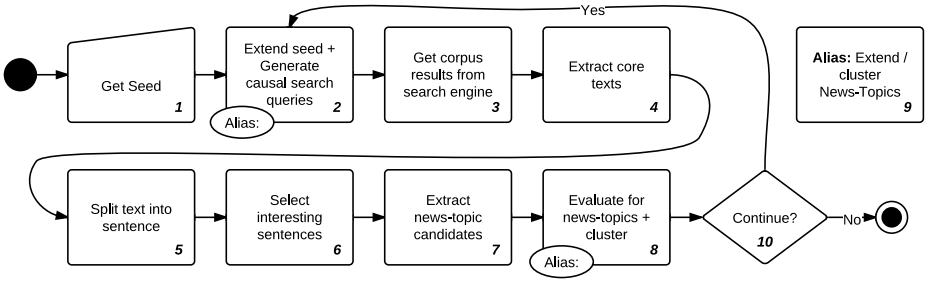

Fig. 1. Experimental prototype process flow

Step 1. The process begins by getting a seed NTR from the user, this will be the first node of the causal network. The heuristics used to validate the seed as an NTR are as follows:

[4] `http://www.ukp.tu-darmstadt.de/software/jwpl/` accessed 1.6.2012

Heuristics 1. The seed phrase is the title for a Wikipedia article that is classified as an event. We classify a Wikipedia article as event type article when it present several markers such as a time and place of happening, or the article contains section titles including timeline, aftermath, impacts, reaction, reponse; negative markers are also used such as title including the phrases "list of", "disambiguation", "band", "film" or "country"[5].

Heuristics 2. The seed phrase is found with semantic markers, such as "The", in the title of multiple news articles.

Heuristics 3. A collection of similar news articles is generated by using the seed phrase as a query. The search engine used may semantically enhance the query by normalizing the phrase, for example with stop word removal, word stemming and synonyms addition. The similarity metric used can be term vector based such as Jaccard's Coefficient or KL divergence [9], where the terms are entities referenced in the news articles. These heuristics are still under development.

Step 2. The initial seed is extended with causal markers such as "caused" and "led to" in a method similar to Hearst's in [10] to generate a query for retrieving relevant documents. The query is also extended with NTR aliases when they are available. We propose two methods to discover NTR aliases: First, when the seed NTR is the title of a Wikipedia article classified as an event, then the titles of the redirect pages are considered NTR aliases. Second, an experimental approach is based on the assumption that alias NTRs are used as search terms with correlated patterns, this is to say an NTR will be used as a search term with a similar popularity distribution to an alias of that NTR. Thus, by analyzing query patterns, like in Google Correlate[6], potential NTR aliases can be discovered.

Step 3. The extended query is used to retrieve relevant documents through the search engine. The top ranked results are selected for further processing. In this step the search engine can be based on a local corpus indexed by Lucene with additional libraries such as Mahout and WordNet, to discover semantically related results. For simplicity Google online search engine was used in the implementation.

Step 4. The next step is to obtain relevant information from the top ranked results, this includes the human readable text and metadata such as creation time and source information. This information can later be used to refine the results in the causal network.

Step 5. The readable text is split into sentences, this facilitates processing the results and allows a summary results to be presented in the causal graph.

Step 6. Validating phrases to be NTR is resource intensive, therefore only selected phrases that contain the seed NTR or an alias, and a causal marker are further processed.

[5] The system for *Heuristics 1* is available upon request, please contact the author.
[6] `http://www.google.com/trends/correlate`

Step 7. The causal marker and pattern approach is used to define the causally related items, if one of the items is the seed NTR or an alias the counterpart is marked as a NTR candidate, to be evaluated in the next step.

Step 8. The NTR candidates are evaluated, all the validated NTRs are collected and alias matching is performed in **Step 9**, to reduce the result set. The causally related news topics can then be used as new seed NTR to continue expanding the network.

Step 10. At the end of the process we have a seed NTR and causally related NTRs, and the user is able to review the sources of the information at a sentence level. Additional information such as temporal sorting and ranking of causal relations may also be presented.

6 Results

From the process flow in 1 there is a series of intermediate results, such as generating the query in **Step 2** and extracting causal relations from sentences in **Step 7**, then evaluating the cause and effect as valid references to news topics in **Step 8**. The final result to the initial query from the user is a collection of phrases that are references to a news topics, clustered for simplification. The phrases are provided with additional information such as the document source and sentence from which the phrase were extracted. This information can be presented in a graphical interface to make the information more intuitive, similar to a Bayesian network.

7 Evaluation

Several components of the system have been evaluated, for the evaluation of the proposed hypotheses the complete system is required and under development. The evaluated components demonstrate the viability of the approach, these are: event classification from Wikipedia articles, NTR validation, NTR alias discovery and causal relation extraction between NTRs.

We conducted an empirical test to evaluate the efficiency of our event classification algorithm. For this, we randomly selected a Wikipedia article from the database and let the system try to identify it as an event. This process was repeated till one hundred events were found. The algorithmically identified events were then manually rechecked, to see if they were correctly identified or if they are false positives. We also rechecked the Wikipedia articles which did *not* pass the algorithm to find false negatives. The Table 1 shows the results.

Table 1. Wikipedia event article classification

	Positive (predicted)	Negative (predicted)	**Recall**
Positive (actual)	81	138	**37%**
Negative (actual)	19	3686	
Precision	**81%**		**Accuracy 82%**

An analysis of the results shows that the performance is highly affected by the quantity of information in articles, for example articles that contain many sections or articles that do not contain an infobox were in many cases misclassified. In future work we consider information quantity as a feature for classification.

The NTR validation was evaluated on a collection of 60 phrases extracted by the prototype system. To extract the phrases the system was initialized with 20 distinct seed NTR. The seed NTRs were not expanded with alias NTR to facilitate the evaluation. From the resulting 215 phrases the system classified as causally related news topic, a random selection of 60 were selected for evaluation. The evaluation by human annotators was done via an online survey system[7]. Annotators were given the NTR that is causally related to the seed NTR and the phrase from which the NTR was extracted, then the users were asked if the NTR refers to an event or activities in the given phrase. The possible annotations were true, false or unknown, the latter option is to consider phrases that are unclear or incorrectly parsed from the original source. Results show that from the phrases that could be classified by the annotators close to 64% were correctly classified by the system.

In future work a collection of phrases composed of known NTRs and additional phrases will be used to test if the system can sort them correctly. The performance of this component is comparable to the keyword extraction accuracy of 24% to 60% in [2].

To evaluate causal relation extraction between NTRs, human annotators were provided information that was extracted by the prototype. The information consisted of a seed NTR, a phrase that the system classified as causally relating the seed NTR to an another NTR, and the additional NTR extracted from the causal phrase. Because the system is intended to causally relate NTR it follows that the performance of the causal relation extraction is dependent on the performance of NTR validation. The triplets of seed NTR, causal phrase and additional NTR were generated based on 20 seed NTR, 40 generated triplets were randomly selected for evaluation. Annotators were asked to confirm if the seed NTR and the additional NTR were causally related in the given phrase. The system showed an accuracy of around 24%, an analysis of the results show that many of the extracted NTR were incorrect and that question phrases were wrongly classified as causal. The performance of causal relation extraction between news topic references is similar to causal relation extraction systems trained in one domain and evaluated in a different domain [11]. In future work causal relation extraction will be evaluated separately from NTR validation.

To evaluate NTR aliases annotators were provided 40 pairs of phrases the system classified as aliases, these aliases were generated based on eight seed NTRs. Annotators were given the phrase pairs and asked if they refer to the same event or activity. Annotators could answer yes, no or unknown, the latter option was given to consider terminology that may be obscure or unknown to the user. The overall accuracy was 65%. The results are an improvement over 36% accuracy in [2] for linking similar topic event pairs.

[7] `https://crowdflower.com/jobs/65074/`

8 Discussion

Much research in causal relation extraction and topic modeling has already been completed, but there are still gaps to be filled, in order to complete the system as proposed here. Though the experimental prototype performs better than state of the art only in very narrow aspects of the preliminary evaluation, it does function as a proof-of-concept system and provides an intuition of the overall architecture of the system. Because the proposed method for causal relation extraction is based on a novel topic model and vice versa, these two must be developed in parallel. These technical contributions will allow a more general contribution of a useful tool to aid in understanding current news, a novel approach to navigating between news topics and a method to potentially present past news topics as newly relevant.

9 Future Work

In [11] the probability of verb pairs being in a causal relation was estimated based on temporally ordered text, it follows that news topic references, that are also temporally ordered, can also be used to calculate causal relation probability. News topic references for past event can be extracted from knowledge bases such as Wikipedia, a greater challenge is discovering news topic references for current events that are not yet in knowledge bases, by using known news topic references and explicit causal relations it is possible to extract novel news topic references, this process will be developed in future work. Based on the aforementioned future works a system that automatically generates a network of causally related news topics to facilitate used understanding will be presented, this system will be used to evaluate the hypotheses presented above.

10 Conclusion

In this paper we propose a system to reduce news information overflow by automatically generating a network of causally related news topics. To generate the network, information is extracted from distributed news sources and presented in a simplified format. To causally related news topics, a topic model called "News Topic Reference" is presented. This topic model is suitable for extracting causal relations from natural language, it is semantically comprehensive and it can be used to cluster results from the causal relation extraction, this model is one of the contributions of this work. Using features found in news topics such as occurrence count and temporal order, it is possible to aggregate information to causal relations in order to provide more comprehensive results.

The proposed system is based on several assumptions that will be fully evaluated once the system is past the prototype stage. These assumptions focus on the usability of the system and the availability of information for the system to function. A preliminary evaluation provides a proof-of-concept for the viability of the proposed system.

References

1. Feng, A., Allan, J.: Incident threading for news passages. In: Proc. of the 18th ACM Conference on Information and Knowledge Management, pp. 1307–1316 (2009)
2. Ishii, H., et al.: Causal Network Construction to Support Understanding of News. In: Proceedings of HICSS 2010, pp. 1–10 (2010)
3. Shahaf, D., Guestrin, C.: Connecting the dots between news articles. In: Proc. of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 623–632 (2010)
4. Nallapati, R., et al.: Event threading within news topics. In: ACM International Conference on Information and Knowledge Management, pp. 446–453 (2012)
5. Kim, D., Oh, A.: Topic chains for understanding a news corpus. In: Computational Linguistics and Intelligent Text Processing, pp. 163–176 (2011)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
7. Ahmed, A., et al.: Unified Analysis of Streaming News. In: Proc. 20th International Conference on World Wide Web, pp. 267–276 (2011)
8. Harabagiu, S., Lacatusu, F.: Topic themes for multi-document summarization. In: Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2005)
9. Huang, A.: Similarity Measures for Text Document Clustering. In: Proc. 6th New Zealand Computer Science Research Student Conference, pp. 49–56 (2008)
10. Girju, R., Moldovan, D.: Text mining for causal relations. In: Proceedings FLAIRS Conference, pp. 360–364 (2002)
11. Rink, B., et al.: Learning Textual Graph Patterns to Detect Causal Event Relations. In: Proc. 23rd Florida Artificial Intelligence Research Society International Conference, Applied Natural Language Processing Track, pp. 265–270 (2010)
12. Beamer, B., Girju, R.: Using a Bigram Event Model to Predict Causal Potential. In: Gelbukh, A. (ed.) CICLing 2009. LNCS, vol. 5449, pp. 430–441. Springer, Heidelberg (2009)
13. Radinsky, K., Davidovich, S.: Learning Causality from Textual Data. In: NLU 21 (2011)
14. Wolff, P., et al.: Models of causation and causal verbs. In: The Meeting of the Chicago Linguistics Society, pp. 607–622 (2002)
15. Joshi, S., et al.: Lexico-syntactic causal pattern text mining. In: Proc. 14th WSEAS International Conference on Computers, pp. 446–451 (2010)
16. Chang, D.-S., Choi, K.-S.: Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. Information Processing & Management 42(3), 662–678 (2006)
17. Ittoo, A., Bouma, G.: Extracting Explicit and Implicit Causal Relations from Sparse, Domain-Specific Texts. In: Muñoz, R., Montoyo, A., Métais, E. (eds.) NLDB 2011. LNCS, vol. 6716, pp. 52–63. Springer, Heidelberg (2011)
18. Butnariu, C., et al.: SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. In: Proc. 5th International Workshop on Semantic Evaluation, pp. 39–44 (2010)
19. Chan, Y.S.: Minimally Supervised Event Causality Identification. In: Proc. Conference on Empirical Methods in Natural Language Processing, pp. 294–303 (2011)
20. Riaz, M., Girju, R.: Another Look at Causality: Discovering Scenario-Specific Contingency Relationships with No Supervision. In: IEEE 4th International Conference on Semantic Computing, pp. 361–368 (2010)
21. Allan, J., et al.: Topic detection and tracking pilot study final report (1998)