# NEWS: News Event Walker and Summarizer

Radityo Eko Prasojo
RPrasojo@unibz.it
Free University of Bozen-Bolzano
Bozen-Bolzano, Italy

Mouna Kacimi
Mouna.Kacimi@unibz.it
Free University of Bozen-Bolzano
Bozen-Bolzano, Italy

Werner Nutt
Werner.Nutt@unibz.it
Free University of Bozen-Bolzano
Bozen-Bolzano, Italy

## ABSTRACT

Most news summarization techniques are static, and thus do not satisfy user needs in having summaries with specific structures or details. Meanwhile, existing dynamic techniques such as query-based summarization fail to handle content-independent queries that target the type of summary information such as time, location, reasons, and consequences of reported events. The NEWS system supports multi-granular summarization along two dimensions: the level of detail and type of information. The system employs fine-grained information extraction to extract facts and their facets with type tagging. The extracted information is then modeled as a graph used to create summaries. The system incrementally expands summaries based on the nodes visited by users, folding related events into the search space.

## CCS CONCEPTS

• **Information systems → Document structure**; **Summarization**; • **Computing methodologies → Natural language processing**.

## KEYWORDS

Summarization; Information Extraction; Knowledge Graph

**ACM Reference Format:**
Radityo Eko Prasojo, Mouna Kacimi, and Werner Nutt. 2019. NEWS: News Event Walker and Summarizer. In *2019 International Conference on Management of Data (SIGMOD '19), June 30-July 5, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3299869.3320243

## 1 INTRODUCTION

Abstractive summarization approaches [2, 5, 6, 8] have recently gained traction due to their functionality to paraphrase texts from multiple sources as input. While these approaches can provide reasonably good results, they output a single summary of each input text. Thus, by design, they cannot be customized to satisfy user needs in having summaries with specific type of detail or focus. To address this issue, query-based techniques were proposed [1, 3, 11]. The idea is to introduce a pre-processing phase where query-relevant text should be extracted before performing summarization. These approaches are able to customize the input text to user needs, but they suffer from two main limitations: (1) users should have a minimum knowledge about the content of the input text to be able to pose queries, which is not always possible; (2) content-independent queries that target the structure of the summary are not handled, such as reporting the reasons for all mentioned facts.

In this paper, we present NEWS, a multi-granular summarizer of news events. The system performs semantic summarization [5, 7] by extracting facts, represented as triples, from the input text and modeling them as a graph. The graph model naturally connects facts along the paths with their complementary information, which we call *facets.* Thus, it allows for an incremental gathering of information along the paths leading to summaries with different granularity. Moreover, by exploiting graph properties such as node degrees, NEWS can effectively find important facts. Few techniques have used graph models to create summaries [5, 9]. They all rely on open information extraction tools that are not fine-grained and do not provide facet tagging such as *time*, *location*, *reason*, and *consequence*, which are crucial for type-based expansions of summaries. By contrast, NEWS uses StuffIE [10], a fine-grained information extraction tool with facet tagging. NEWS leverages node degrees and fact saliency to rank paths. This results in boosting paths that have multiple authoritative nodes and therefore identifying important facts to be included in the summary. The final product of the system offers the possibility for users to browse through the graph of events and decide which type of information to be included in the summary and at which level of detail.

## 2 NEWS EVENT GRAPH

We use StuffIE [10] to perform a fine-grained extraction of events in news articles. Events are extracted as a set of facts and facets. Facts are represented as triples of the form ⟨*arg1; predicate; arg2*⟩, which typically occur with complementary information called *facets.* We consider the following example:

*SIGMOD '19, June 30-July 5, 2019, Amsterdam, Netherlands*

ACM ISBN 978-1-4503-5643-5/19/06…$15.00
https://doi.org/10.1145/3299869.3320243

> *S1: "U.S. President Donald Trump fired Secretary of State Rex Tillerson on Tuesday. Trump has nominated CIA Director Mike Pompeo to replace Tillerson as America's new top diplomat."*

The fact ⟨*Trump*; *fired*; *Secretary of State Rex Tillerson*⟩ has one facet *"on Tuesday"* about time. By contrast, the fact ⟨*Trump*; *has nominated*; *CIA Director Mike Pompeo*⟩ has two facets. The first one is *"to replace Tillerson"* which represents a *consequence*. The second facet is *"America's new top diplomat,"* which is about a *role*. Facet tagging is provided by StuffIE using a distant-learning approach that exploits Oxford English dictionary. We model the extracted information as a graph where a node can be either a fact or a facet, whereas an edge represents a relation between two facts or between a fact and a facet. Figure 1 shows an example of a semantic graph generated from a set of news articles about *Trump firing Tillerson.* For the sake of simplicity we just show a subgraph that contains the facts and facets extracted from the following sentences:

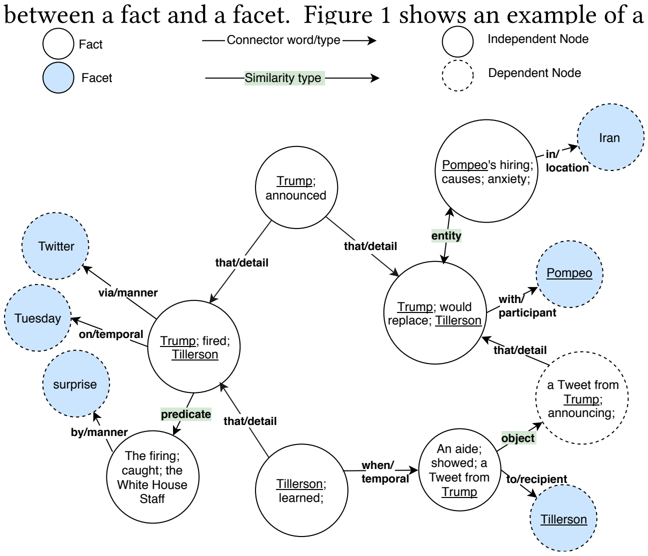


**Figure 1: Example of a news event graph.**

> *S2: "Mr. Tillerson learned he had been fired on Tuesday morning when a top aide showed him a tweet from Mr. Trump announcing the change." [New York Times] S3: "Mr. Trump announced he would replace Mr. Tillerson with Mike Pompeo." [New York Times]*
> *S4: "President Donald Trump announced Tuesday morning that he had fired Secretary of State Rex Tillerson." [NBC News]*
> *S5: "US President Donald Trump has fired Secretary of State Rex Tillerson via Twitter." [BBC News]*
> *S6: "Pompeo's hiring causes anxiety in Iran." [★ Adv.]*

Nodes in the graph are either independent or dependent. A node is dependent if information contained in it cannot stand alone; it depends on its predecessor node. In Figure 1, independent and dependent nodes are shown with solid and dotted lines, respectively. We can see for example that the fact node ⟨*a Tweet from Trump*; *announcing*;⟩ is a dependent node. It depends on the fact node ⟨*Trump*; *would replace*; *Tillerson*⟩. Typically, facets are always represented as dependent nodes, while facts can be of any type.

Edges in the graph carry labels that depend on the type of relation between nodes. Edges between facts and their facets are labeled with connectors such as "*that*", "*on*", "*with*", "*by*", and "*via*". By contrast, edges between facts can have two different types of labels. The first type consists of connectors similarly to the case of facets. The second type reflects similarities between facts. Facts can have similar subjects, predicates, objects, or share entity occurrences also through their facets. In Figure 1, we can see that the fact ⟨*Pompeo's hiring*; *causes*; *anxiety*⟩ is connected with an edge of type *entity* with the fact ⟨*Trump*; *would replace*; *Tillerson*⟩ because it has the facet "*with Mike Pompeo*."

## 3 SYSTEM ARCHITECTURE

Figure 2 shows an overview of the architecture of NEWS.

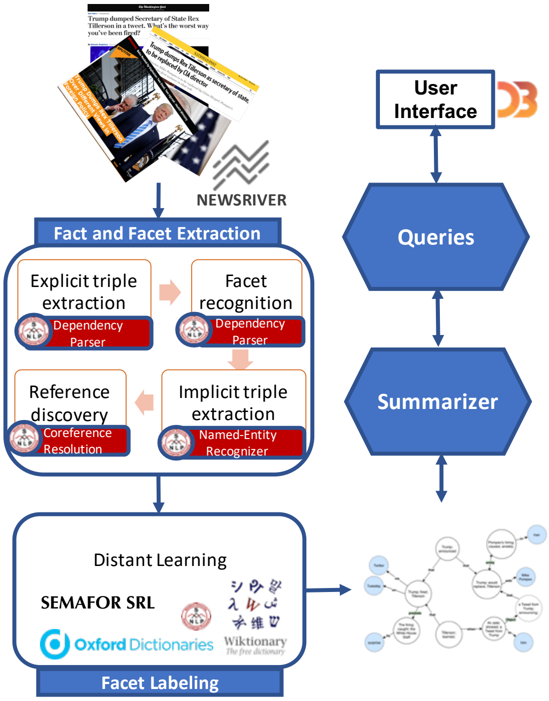


**Figure 2: NEWS System Architecture**

Information from news articles is extracted using StuffIE where events are represented by facts and facets. Then distant learning is employed to label the extracted facets giving more semantic insights to the data. The outcome of this process is an event graph from which summaries are built incrementally, based on a set of selected nodes. The nodes can be selected automatically using graph properties such as node degrees. Alternatively, they can be selected based on user queries that specify the type of information and the level of detail that should be included in the summary.

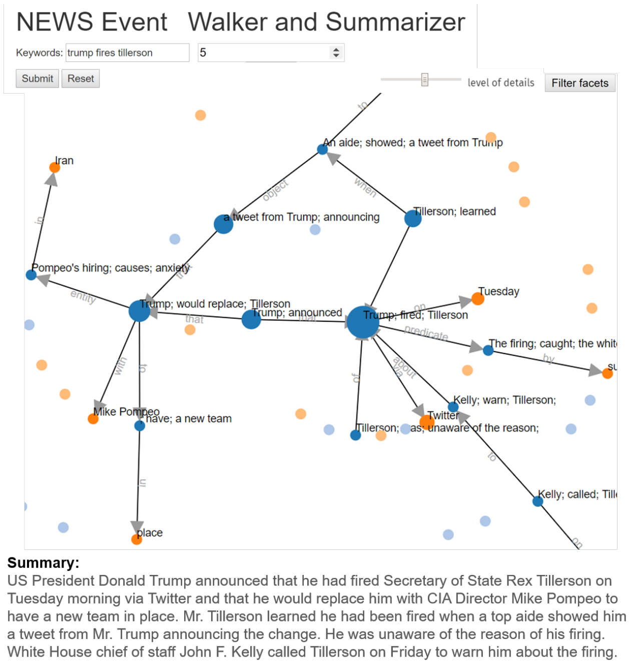
**Fact and Facet Extraction.** Each fact is represented as a triple relation. To extract triples from unstructured text, StuffIE proceeds as follows. The Stanford dependency parser generates a grammatical tree from each sentence. The nodes of the grammatical tree are the words of the sentence, while the edges represent their syntactical relationships. The extraction starts by turning each verb node into the predicate of a triple relation. Then, using handcrafted rules, we process all the paths issuing from that verb node to find the corresponding subject, object, and facets, based on the types of the dependencies in the grammatical tree. After extracting facts by finding the subject and object of each predicate in a sentence, we proceed with the identification of the facets related to each fact. To this end, we use verb complements as indicators for the presence of facets in a sentence. We note that in StuffIE we also extract nested relations that correspond to the case where subjects or objects of triples are not simple noun phrases but clauses containing other facts. As a further enhancement of the extraction process, we extract implicit relations and perform co-reference resolution as described in our previous work [10].

**Facet Labeling.** After extracting facts and facets, we proceed with labeling the facets to indicate their role with respect to the corresponding facts. For example, a facet can be the *"reason"*, the *"purpose"*, or the *"consequence"* of its related fact. To achieve that, we employ a distant learning approach, which is divided into two main steps. First, we construct the set of facet labels by observing the meanings (i.e., glosses) of English prepositions, which are the main connectors of facets. As a result of this analysis, we came up with 22 handcrafted labels. Then, we consolidated these labels by comparing them with the labels provided by Illinois SRL [4] using the same descriptions, reaching 35 labels. The second step is to build a classification model that maps each facet to a label given the connector to the corresponding fact and the context sentence. We used a multi-nominal logistic regression model to perform facet labeling. The training data were constructed automatically based on the preposition glosses provided by Oxford Dictionary and Wiktionary [10].

**Summary Generation.** We select the summary content by taking the top-$K$ fact nodes in the graph. We call these nodes the *seed nodes*. The seed nodes are ranked according to three main criteria: (1) the document frequency with which they appear, (2) the total number of their occurrences across all articles, and (3) their node degree, which is the total number of their incoming and outgoing edges. Technically, from each of the top-$K$ seed nodes selected earlier, we iteratively add neighboring nodes (facets or other facts) to generate a summary based on the selected nodes. We give as input a window size of how far we want to traverse the graph. The window size corresponds to the number of selected facets and the depth of the traversal. We note that facets are selected with respect to a certain ranking that we learned from the most common types of facets used in the training data of TAC and DUC[1]. Alternatively their types can be decided by the user as input. Once the set of facts and facets to be included in the summary is selected, we proceed with the generation of summary sentences. This is done by transforming the selected graph nodes into a readable text. We start from our previous approach [9], which naively unfolds facts and their facets based on their canonicalized form. Then we enhance the process to solve the problem of fluency and redundancy by (1) grouping related facts, (2) avoiding facet duplicates, and (3) reducing repetitions based on co-references.

## 4 DEMO DESCRIPTION

We have implemented a web-based tool that demonstrates NEWS. Figure 3 shows a screenshot of the demo.



NEWS Event Walker and Summarizer

Keywords: trump fires tillerson

5

**Summary:**
US President Donald Trump announced that he had fired Secretary of State Rex Tillerson on Tuesday morning via Twitter and that he would replace him with CIA Director Mike Pompeo to have a new team in place. Mr. Tillerson learned he had been fired when a top aide showed him a tweet from Mr. Trump announcing the change. He was unaware of the reason of his firing. White House chief of staff John F. Kelly called Tillerson on Friday to warn him about the firing.

**Figure 3: NEWS demo - graph and summarization**

A user starts by inputting keywords of a news event of their interest and the number of articles that NEWS will collect to be summarized. Our tool will fetch the articles using the Newsriver API[2] and then run them through the pipeline that has been described in Section 3. The graph, which follows the description shown in Figure 1, is visualized in the browser

[1] https://tac.nist.gov/ and http://duc.nist.gov/
[2] https://newsriver.io/

using a variant of force-directed graph of D3.js.[3] Blue nodes represent facts, whereas orange nodes represent facets. Bigger nodes represent the more important facts and facets. The nodes are not rigid, that is, the user may drag them around to read the text better and to have a better view of the whole structure. Beside the graph, our tool displays the summary generated from the selected nodes in the graph. The node selection is based on (1) the level of detail, that is, the window size of the number of facets and the traversal-depth starting from the seed nodes and (2) the facet roles. Nodes that are not selected are faded-out, and their labels and edges are completely hidden. By default, our tool sets the window size in the middle and sets all facet roles to be considered in the summary. The user may change them by interacting with the slider and by clicking the "Filter facets" button, respectively. By doing so, the tool will re-visualize the graph following the new settings and also generate a new summary accordingly.

The user may double click a fact node to view some additional detail of the fact, including the involved entities, the facets and their roles. Figure 4 shows a screenshot of this.

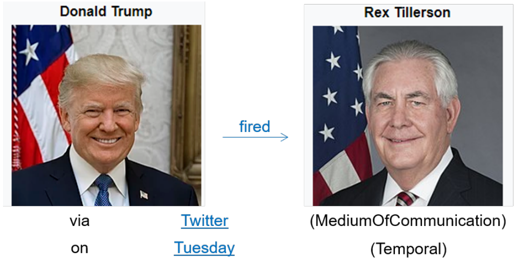


**Figure 4: NEWS demo - fact view**

The user may click the predicate of the fact to see a list of links, each containing an article that reports the fact. The user may also click a row containing the facet information, which will also show a list of links to the articles that contain the facet. The subject, the object, or some facets of the fact may refer to some other facts (in the case of nested relations) or some entites. In both cases they are represented as links. Clicking on a link referring to another fact will show the user a fact view page that represents the fact. On the other hand, clicking on one of the entities will bring the user to the entity detail page, which contains the Wikipedia infobox of the selected entity alongside fact-view links that shows where the entity appears in the graph, either as a subject, object, or one of the facets. For example, clicking on the photo of Rex Tillerson on Figure 4 will show Figure 5, while clicking on the link of "Trump; fired" on Figure 5 will bring the user back to Figure 4.

## 5 CONCLUSION AND FUTURE WORK

We have presented NEWS, a multi-granular summarizer of news events that provides an interactive experience that

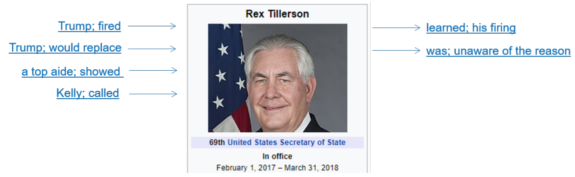


**Figure 5: NEWS demo - entity view**

allows the user to browse through graph of related events and to modify the content of the graph by setting the level of detail and the desired facet roles. Then, our tool will generate a summary according to the settings.

NEWS leverages NLP pipelines and we plan to improve some of them. For example, co-reference resolution is an important part of NEWS, in particular in building the event graph which requires the identification of similar entities and events. However, existing techniques and our enhancement still have some limitations, including resolving plural pronouns (e.g. we) and pronouns taken from quoted speech (e.g. I). Both are common occurrences in news articles.

## REFERENCES

[1] T. Baumel, M. Eyal, and M. Elhadad. 2018. Query Focused Abstractive Summarization: Incorporating Query Relevance, Multi-Document Coverage, and Summary Length Constraints into seq2seq Models. *CoRR* abs/1801.07704 (2018). http://arxiv.org/abs/1801.07704

[2] L. Bing, P. Li, Y. Liao, W. Lam, W. Guo, and R. Passonneau. 2015. Abstractive Multi-Document Summarization via Phrase Selection and Merging. In *Proc. ACL'15*, Vol. 1. 1587–1597.

[3] E. Canhasi and I. Kononenko. 2014. Weighted Archetypal Analysis of the Multi-element Graph for Query-focused Multi-document Summarization. *Expert Syst. Appl.* 41, 2 (2014), 535–543. https://doi.org/10.1016/j.eswa.2013.07.079

[4] J. Clarke, V. Srikumar, M. Sammons, and D. Roth. 2012. An NLP Curator (or: How I Learned to Stop Worrying and Love NLP Pipelines). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey, x–y.

[5] P. Li, W. Cai, and H. Huang. 2015. Weakly Supervised Natural Language Processing Framework for Abstractive Multi-Document Summarization. In *Proc. CIKM'15*. ACM, 1401–1410.

[6] W. Li. 2015. Abstractive Multi-document Summarization with Semantic Information Extraction.. In *EMNLP'15*. 1908–1913.

[7] T. Oya, Y. Mehdad, G. Carenini, and R. Ng. 2014. A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships. In *Proc. INLG'14*. ACL, 45–53. http://www.aclweb.org/anthology/W14-4407

[8] D. Pighin, M. Cornolti, E. Alfonseca, and K. Filippova. 2014. Modelling Events through Memory-based, Open-IE Patterns for Abstractive Summarization.. In *ACL (1)*. 892–901.

[9] R. E. Prasojo, M. Kacimi, and W. Nutt. 2018. Modeling and Summarizing News Events Using Semantic Triples. In *Proc. ESWC'18*. 512–527. https://doi.org/10.1007/978-3-319-93417-4_33

[10] R. E. Prasojo, M. Kacimi, and W. Nutt. 2018. StuffIE: Semantic Tagging of Unlabeled Facets Using Fine-Grained Information Extraction. In *Proc. CIKM'18*. 467–476. https://doi.org/10.1145/3269206.3271812

[11] S. Xiong and D. Ji. 2016. Query-focused Multi-document Summarization Using Hypergraph-based Ranking. *Inf. Process. Manage.* 52, 4 (July 2016), 670–681. https://doi.org/10.1016/j.ipm.2015.12.012

[3] https://beta.observablehq.com/@mbostock/d3-force-directed-graph