# Summarizing Conceptual Graphs for Automatic Summarization Task

Sabino Miranda-Jiménez, Alexander Gelbukh, and Grigori Sidorov

Natural Language and Text Processing Laboratory,
Center for Computing Research, National Polytechnic Institute
Av. Juan de Dios Bátiz, s/n, esq. Mendizábal,
Col. Nueva Industrial Vallejo, 07738, Mexico City, Mexico
sabino@sagitario.cic.ipn.mx, www.gelbukh.com,
www.cic.ipn.mx/~sidorov

**Abstract.** We propose a conceptual graph-based framework for abstractive text summarization. While syntactic or partial semantic representations of texts have been used in literature, complete semantic representations have not been explored for this purpose. We use a complete semantic representation, namely, conceptual graph structures, composed of concepts and conceptual relations. To summarize a conceptual graph, we remove the nodes that represent less important content, and apply certain operations on the resulting smaller conceptual graphs. We measure the importance of nodes on weighted conceptual graphs by the HITS algorithm, augmented with some heuristics based on VerbNet semantic patterns. Our experimental results are promising.

**Keywords:** Automatic summarization, conceptual graphs, graph-based ranking algorithms, HITS algorithm.

## 1 Introduction

With the overwhelming amount of information available today on the Internet and elsewhere, summarization technologies are essential to improve the access to this information. High-quality automatic text summarization is a challenging task that involves text analysis, text understanding, the use of domain information, and natural language generation.

Summarization approaches can be categorized as extractive and abstractive. The limitations of extractive approach are well known: in the first place, low quality of the generated summaries. On the other hand, abstractive summaries have not been sufficiently explored because of the need in a deeper text analysis required for understanding the texts, and complexity associated with it. Such a deep analysis is indispensible to improve the quality of summaries [1].

We propose a method for single-document abstractive summarization, based on conceptual graphs as the underlying text representation [8]. This kind of representation has not been used for automatic summarization so far. We focus on ranking nodes and applying a kind of pruning operation, namely, selecting the most important

H.D. Pfeiffer et al. (Eds.): ICCS 2013, LNAI 7735, pp. 245–253, 2013.

nodes according to HITS algorithm [5] over weighted conceptual graphs and using other heuristics based on semantic patterns of VerbNet [13]. The summary at semantic level is the resulting structure of selected nodes. Automatic generation of conceptual graphs from text is beyond the scope of this paper.

This paper is organized as follows. Section 3 describes our approach. Section 4 presents the experimental results. Finally, Section 5 gives the conclusions and future work.

## 2 Related Work

In recent years, there has been an increase in the interest to graph-based methods in Natural Language Processing. Graph-based approaches such as LexRank [2] and TextRank [3] have been used for keyword extraction for extractive summarization. In these approaches, graphs are usually considered undirected and unweighted; their nodes are either sentences, words, or other kind of units, and edges are defined by overlaps of the content between units. In these approaches, well-known iterative algorithms are used such as HITS or PageRank to rank the nodes in order to select salient ones. The selected nodes represent the summary; non-salient nodes are removed from the graph.

Other approaches use word order to create the graphs [6]. The graphs are directed. Nodes are words, and the edges represent the precedence of the word in the sentence, that is, the word in the word order is important. The resulting graph is ranked similar to TextRank approach.

In [3] the notion of weighting edges was introduced in HITS algorithms. Overlap of sentences was used as a kind of weight, but because of an unnatural way of using weights, the study was mainly on undirected and unweighted graphs. In contrast, a conceptual graph can be considered as a weighted graph having sense because conceptual relations between concepts provide a semantic flow through the graph, namely, the semantic flow over agent relations, object relations, attribute relations, etc. Another feature in our model is the preference of the node in order to select concepts (nodes) which the users are interested (see Section 3.3 and Section 3.4.).

There have been attempts to use the semantics of the document, such as in Semantic Graphs approach [4, 7]. This approach uses triplets (subject—predicate—object). Each triple is characterized by a rich set of linguistic, statistical, and graph attributes. A Support Vector Machine classifier is used to identify important triples to generate the summary. Nevertheless, a real and complete, fine-grained semantic representation is not used.

## 3 Approach Using Conceptual Graph

### 3.1 Conceptual Graphs Formalism

Conceptual Graphs (CGs) [8] are structures for knowledge representation based on first-order logic. They are natural, simple, and fine-grained semantic representations

to depict texts. A conceptual graph is a finite, connected and bipartite graph. It has two kinds of nodes: *conceptual relations* (ovals) and *concepts* (rectangles) (Fig. 1) [8]. A concept is connected to a related concept by conceptual relation. Each conceptual relation must be linked to some other concept.

In our approach, by concepts, we consider content words (that is, all except for stop words); by conceptual relations we consider semantic roles [11]: agent, causer, instrument, experiencer, patient, location, time, object, source, and goal, as well as some other relations, such as attribute, quantity, measure, etc.— approximately 30 relations used in [8].

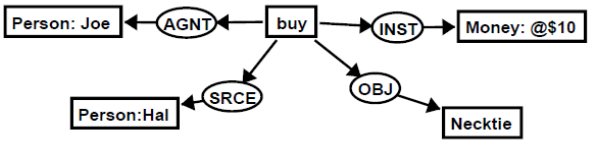


**Fig. 1.** Conceptual graph for sentence: *Joe buys a necktie from Hal for $10*

Other element of CGs is concept types. Concept types represent classes of entities (*Person*, *Money,* Fig. 1), attribute, state and event. It is also called concept type hierarchy that represents an AKO (is-a-kind-of) hierarchy, and it is used to map concepts into the hierarchy for inference purposes [8, 15]. For example, in Fig. 1, *Person:Joe* denotes the concept type *Person*, and its referent *Joe* is an instance of *Person*.

CG Framework allows graph-based operations for reasoning. A number of operations such as: restriction, simplification, unification (join), graph matching (projection), and indexing can be performed to create, manipulate and retrieve large sets of conceptual graphs [8, 15].

### 3.2 Construction of Conceptual Graphs

The construction of a conceptual graph from a text is not direct. It requires an additional process to discover relationships among text units. Approaches have been proposed for automatically generating conceptual graphs such as in [10], but tools are not available. Thus, we manually created the collection of conceptual graphs based on news of DUC-2003 competition in order to prove our ideas.

We use simple conceptual graphs (without negations, situations, or contexts) to simplified our task. For instance, the conceptual graphs for the following news are shown in Figure 2: "*Typhoon Babs weakened into a severe tropical storm Sunday night after it triggered massive flooding and landslides in Taiwan and slammed Hong Kong with strong winds. The storm earlier killed at least 156 people in the Philippines and left hundreds of thousands homeless.*"

In Figure 2, we use a notation '(number)' for a concept that would be referred, and '#' for a co-reference to the concept marked with the specific number; for instance, #3 refers to the concept *Typhoon-Babs* (3). In addition, we use the hierarchy of WordNet [12] to map a referent to its concept type. For instance, *Hong-Kong*, *Taiwan* is mapped to *City*.

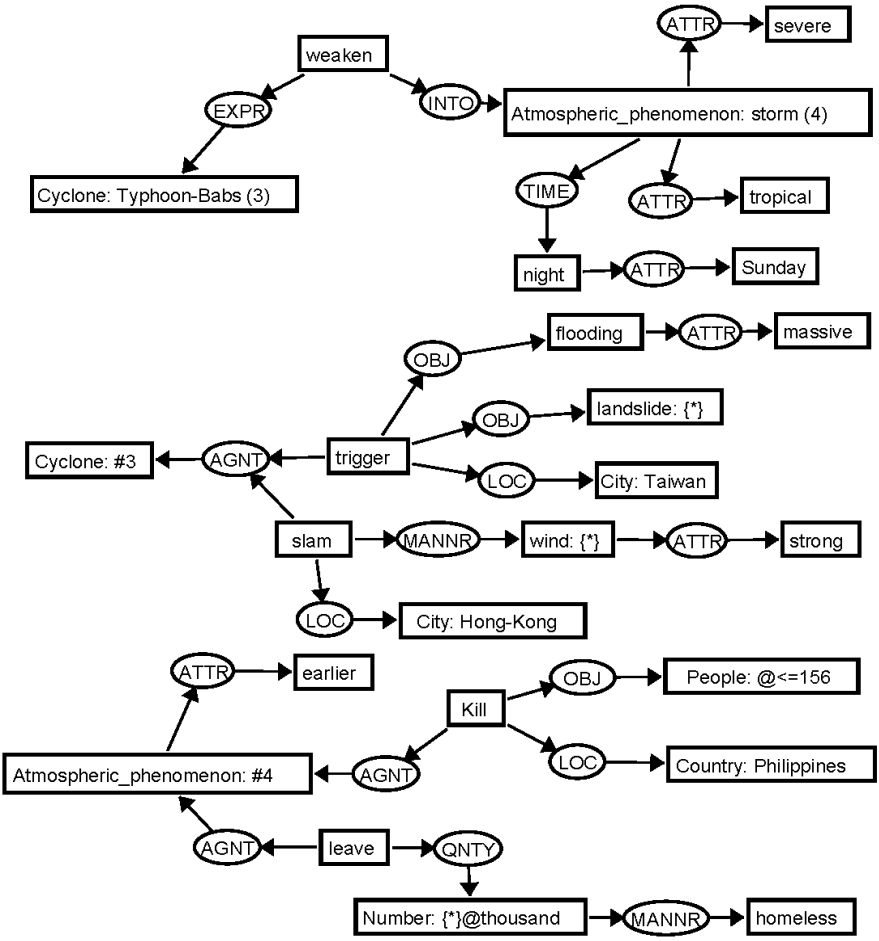


**Fig. 2.** Example of news as conceptual graph

### 3.3 Weighted Conceptual Graphs

We introduce a weighted conceptual graph (Fig. 3). The idea behind these kinds of conceptual graphs is the interest in the semantic flow of graphs. In our approach, edges and nodes have weights. The edge weights are assigned according to the semantic flow in the graph—flow through conceptual relations—, and node weight measures the degree of interest of the topics to the user.

Thus, if the interest is on some semantic flows such as agents, locations, attributes, or other thematic roles, the edge weight that pass through them should be increased in order to reward the flow that pass through them such as in Fig 3. Similar to node preference, a value greater than 1 rewards the topic preference; a value less than 1 penalizes the preference; a value equal to 1 for no reward.

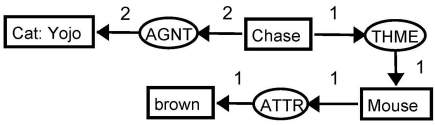


**Fig. 3.** A weighted conceptual graph for sentence: *The cat Yojo is chasing a brown mouse*

For example, if we are interested in the flows that pass through agent relations (AGNT) the incoming and outgoing edges for these conceptual relations are set to value of 2 (see Fig 3).

### 3.4 Ranking Algorithm

HITS [5] is an iterative algorithm that takes into account both in-degree and out-degree of nodes for ranking. The algorithm makes a distinction between authorities (nodes with a large number of incoming links) and hubs (nodes with a large number of outgoing links). For each node, HITS produces two sets of scores: **AUTH**ority and **HUB**. We use the authority score (means that the node is good as information source) in order to choose the nodes that will take part in the summary. We used a modified version of HITS algorithm similar to the proposed in [3].

The equations (1) and (2) are used to compute authorities and hubs scores. Where $I$ is the set of incoming links for node $V_i$; O is the set of outgoing links for node $V_i$; $W_{ki}$ is the weight of semantic flow of edge; and PREF is the node preference.

$$AUTH(V_i) = \sum_{V_k \in I(V_i)} W_{ki} \cdot HUB(V_k) \cdot PREF(V_k) \tag{1}$$

$$HUB(V_i) = \sum_{V_k \in O(V_i)} W_{ik} \cdot AUTH(V_k) \cdot PREF(V_k) \tag{2}$$

### 3.5 Ranking Algorithm of Conceptual Graphs

In order to select the important nodes in CGs, we carry out the following steps:

1. Set hub and authority scores associated to each node a value of 1.
2. Apply the operation Authority, equation (1).
3. Apply the operation Hub, equation (2).
4. Normalize the Authority and Hub values by Euclidian norm.
5. Repeat from 2–4 up to convergence or *N* iterations.
6. Sort nodes by authority values in descending order.
7. Expand the connected concepts for each selected conceptual relation.
8. Expand the associated nodes for each selected concept (verb concept) according to its semantic pattern.
9. Select the top concepts according to a threshold in order to prune the graph.

Mihalcea and Tarau [3] used 20–30 iterations to converge the HITS algorithm; others use one iteration [6]. We identified that more than 15 iterations are enough in our collections of graphs.

Steps 1–6 calculate the HITS scores. Step 7 applies rules to expand the concepts that a conceptual relation connects; for instance, the relation *OBJ(trigger,flooding)* (see Table 1) is expanded into two concepts *flooding* and *trigger*. Step 8 applies the verb pattern rules in order to keep coherent structures.

The semantic patterns of verb concepts were extracted from VerbNet [13]. For example, the pattern for the *chase* concept (Fig. 3) is identified in the VerbNet class ID *chase-51.6*. The pattern is *NP V NP* (Noun Phrase, Verb, Noun Phrase), and the verb is *Basic Transitive.* The role for the first NP is *agent,* and the second NP is *Theme.* Both of them are required for the concept *chase* because it is defined as transitive verb. Thus, the *agent* and the *theme* must be included in a summary.

After applying steps 1–8, Step 9 applies the pruning operation by means of a threshold set by user. It selects nodes without duplicates according to the threshold. The selected nodes represent the summary at the semantic level (see Table 2).

## 4 Experimental Results

We carried out our experiments on the collection of news articles provided by the DUC 2003 [9]. We selected news with length from 40 to 60 words. For each article, there are 3 summaries on average made by humans.

We created three groups of documents from DUC: 2-senteces, 3-sentences, and 4-sentences length such as news in Fig 2. Each group consists of 4 documents represented as conceptual graphs. We set the threshold for pruning operation to 20% of concepts of the original document. As a baseline, we selected the first concepts beginning at the first paragraphs up to the established threshold (except stop words). We set the semantic flow value for agent relations to value of 2. Standard metrics (precision and recall) are used to evaluate our method. **Recall** is the fraction of concepts chosen by the human that were also correctly identified by the method. **Precision** is the fraction of concepts chosen by the method that were correct. **F-measure** is the harmonic mean of precision and recall.

Table 1 shows the selected nodes by ranking method including conceptual relations. Also, expansions of conceptual relations are shown such as object relations (OBJ). Table 2 shows the selected concepts by the method that are part of the summary considering their interrelationships between them; **(req)** indicates that the concept was added because the verb pattern requires it, i.e., *kill* pattern requires its Object (*People:@lt=156*).Table 3 shows the average of the evaluation of the approach for the three collections of graphs.

Our method slightly outperforms the baseline. It is because text documents are very short and the baseline covers the concepts in a good way. Although other approaches have demonstrated that the first and last sentences in the paragraphs are good indicators to find relevant information [14], our method uses all the net and outperforms the baseline. It demonstrates that the method in huge graphs could

operate equally as in small graphs. Finally, the selected concepts in Table 2 represents the summay; according to the CG representation in Fig 2. It could be read: "*Typhoon-Babs triggered flooding and landslides in Taiwan. The storm killed at least 156 people. Typhoon-Babs slammed in Hong Kong.*"

**Table 1.** Selected concepts and conceptual relations by ranking method with expansion of conceptual relations

| NODE | RELATION EXPANSION | AUTH | HUB |
|---|---|---|---|
| Cyclone:Typhoon-Babs | - | *0.729* | *0.3E-16* |
| Atmospheric_phenomenon:storm | - | *0.680* | *0.70E-03* |
| AGNT(trigger-Cyclone:Typhoon-Babs) | trigger/:Typhoon-Babs | *0.054* | *0.147* |
| OBJ(trigger-flooding) | trigger/ flooding | *0.027* | *0.10E-04* |
| OBJ(trigger-landslide) | trigger/landslide | *0.027* | *0.67E-05* |
| LOC(trigger-City:Taiwan) | trigger/:Taiwan | *0.027* | *0.137* |
| AGNT(kill- Atmospheric_ phenomenon:storm) | kill/:storm/ People:@lt=156 **(req)** | *0.022* | *0.147* |
| AGNT(slam-Cyclone:Typhoon-Babs) | slam/:Typhoon-Babs/ City:Hong Kong **(req)** | *0.022* | *0.67E-05* |
| LOC(kill-Country:Philippines) | kill/:Philippines | *0.011* | *0.38E-16* |

**Table 2.** Final selected concept by the ranking method

| NODE | AUTH | HUB |
|---|---|---|
| Cyclone:Typhoon-Babs | *0.729* | *0.3E-16* |
| Atmospheric_phenomenon:storm | *0.680* | *0.70E-03* |
| trigger | *0.054* | *0.147* |
| flooding | *0.027* | *0.10E-04* |
| landslide | *0.027* | *0.67E-05* |
| City:Taiwan | *0.027* | *0.137* |
| kill | *0.022* | *0.147* |
| slam | *0.022* | *0.67E-05* |
| City:Hong Kong | *0.022* | *0.147* |
| People:@lt=156 | *0.022* | *0.70E-03* |
| Country:Philippines | *0.011* | *0.38E-16* |

**Table 3.** Evaluation of the system

| | ***Precision*** | | ***Recall*** | | ***F-Measure*** | |
|---|---|---|---|---|---|---|
| | Baseline | System | Baseline | System | Baseline | System |
| Group I (2-sentences) | 0.38 | 0.50 | 0.38 | 0.50 | 0.38 | 0.50 |
| Group II (3-sentences) | 0.11 | 0.25 | 0.13 | 0.22 | 0.12 | 0.23 |
| Group III (4-sentences) | 0.43 | 0.50 | 0.45 | 0.50 | 0.43 | 0.50 |

## 5 Conclusions

We have proposed a novel graph-based approach for single-document summarization. Our approach is based on the Hub-Authority framework and conceptual graphs as underlying semantic text representation. It combines the text content with semantic roles into graph-based ranking algorithms. The method uses semantic patterns from VerbNet to keep coherent structures when a threshold is applied in order to prune the nodes. Furthermore we introduced a weighted conceptual graph to provide a flexible schema to focus on certain semantic flows or topics by means of weights and preferences. We evaluate our method on DUC-2003 data. The results show that our approach is promising.

In future work, we plan to apply operations such as generalization and join on resulting conceptual graphs in order to improve the quality of the generated summaries. Also, we expect to improve the results on larger conceptual graphs, 500–1000 words per document.

**Acknowledgments.** This work was done under partial support of the Mexican Government (SNI, COFAA-IPN, PIFI-IPN, SIP-IPN 20121823 and 20120418, CONACYT 50206-H and 83270), CONACYT-DST India (122030, "Answer Validation through Textual Entailment"), Mexico City Government (ICYT PICCO10-120), and European project WIQ-EI 269180.

## References

1. Spärck Jones, K.: Automatic summarising: The state of the art. Information Processing & Management 43(6), 1449–1481 (2007)
2. Erkan, G., Radev, D.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22(1), 457–479 (2004)
3. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, pp. 404–411 (2004)
4. Leskovec, J., Grobelnik, M., Milic-Frayling, N.: Learning Semantic Graph Mapping for Document Summarization. In: Proceedings of ECML/PKDD 2004, Workshop on Knowledge Discovery and Ontologies, Pisa, Italy, pp. 1–6 (2004)
5. Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5), 604–632 (1999)
6. Litvak, M., Last, M.: Graph-based keyword extraction for single-document summarization. In: Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization, Manchester, United Kingdom, pp. 17–24 (2008)
7. Tsatsaronis, G., Varlamis, I., Nørvåg, K.: SemanticRank: ranking keywords and sentences using semantic graphs. In: Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, pp. 1074–1082 (2010)
8. Sowa, J.F.: Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, Reading (1984)
9. DUC. Document Understanding Conference (2003), `http://duc.nist.gov/pubs.html#2003`

10. Hensman, S., Dunnion, J.: Automatically Building Conceptual Graphs Using VerbNet and WordNet. In: Proceedings of the 3rd International Symposium on Information and Communication Technologies, Las Vegas, USA, pp. 115–120 (2004)
11. Jackendoff, R.: Semantic Interpretation in Generative Grammar. MIT Press, Cambridge (1972)
12. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
13. Kipper, K., Trang Dang, H., Palmer, M.: Class-Based Construction of a Verb Lexicon. In: Proceedings of Seventeenth National Conference on Artificial Intelligence (AAAI 2000), Austin, TX, pp. 691–696 (2000)
14. Hovy, E., Chin-Yew, L.: Automating Text Summarization in SUMMARIST. In: Mani, I., Maybury, M.T. (eds.) Advances in Automatic Text Summarization, pp. 81–94. MIT Press, Cambridge (1999)
15. Chein, M., Mugnier, M.-L.: Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs. Springer, London (2009)