Automatic Key-phrase Extraction to Support the Understanding of Infrastructure Disaster Resilience

X. Lv[a] , S.A. Morshed[b] and L. Zhang[c]

[a] Florida International University, 10555 West Flagler Street, EC 2956, Miami, FL 33174
[b] Florida International University, 10555 West Flagler Street, Miami, FL 33174
[c] Florida International University, 10555 West Flagler Street, EC 2935, Miami, FL 33174
E-mail: xulv@fiu.edu, smors005@fiu.edu, luzhang@fiu.edu

Abstract –

Preventing natural disasters from causing substantial social-economic damages relies heavily on the disaster resilience of the nation's critical infrastructure. According to the National Academy of Sciences, research on understanding and analyzing the disaster resilience of our infrastructure systems is a "national imperative". To address this need, this paper proposes an automatic keyphrase extraction methodology to extract relevant phrases on disaster resilience from documents in infrastructure domain. In developing the proposed methodology, a document collection including research papers and public reports are prepared. Noun phrases are first extracted from every sentence in the collection and form the candidates for keyphrases following a filtering procedure. Each candidate phrase is then represented as a global semantic vector and a local semantic vector. To select relevant phrases on disaster resilience, a semantic similarly measure is proposed to incorporate the semantics of candidate phrases in both the general and infrastructure domain. Ten physical resilience concepts from a pre-developed community resilience hierarchy is selected as the target concepts to evaluate the performance of the proposed methodology. When evaluated on the document collection, the proposed methodology achieved 66% of precision at top 20 extracted keyphrases on average.

Keywords –

Infrastructure disaster resilience; Automatic keyphrase extraction; Natural language processing

1 Introduction

Preventing natural disasters from causing substantial social-economic damages relies heavily on the disaster resilience of the nation's critical infrastructure. U.S. National Academies has defined resilience as "the ability to prepare and plan for, absorb, recover from, and more successfully adapt to adverse events" [1]. At present, United States is in dire need of resilient infrastructure systems as its physical infrastructure is aging and deteriorating. On this context, a report published by the National Academy of Sciences has defined the research on understanding and analyzing the disaster resilience of our infrastructure systems as a "national imperative" [1]. To facilitate infrastructure disaster resilience, one major prerequire is to understand disaster resilience in a more explicit and deep manner. Existing research typically provides conceptual or theoretical framework that identifies some key characteristics (e.g., robustness, redundancy) of disaster resilience without directly linking them to more detailed and specific concepts. There is, thus a need to analyze the vast amount of text documents (e.g., reports, papers, news articles) on disaster resilience to facilitate a better understanding on the definition, interpretation and classification of disaster resilience in the infrastructure domain.

To address this need, this paper proposes a methodology that automatically extracts keyphrases about disaster resilience from documents in the infrastructure domain. Extracted keyphrases can provide highly condensed and valuable summary of disaster resilience in the infrastructure domain. The proposed methodology and the extracted keyphrases could facilitate the information and knowledge management of disaster resilience, such as the preparation and development of disaster recovery/resilience guidance manuals, and the retrieval of disaster recovery/resilience best management practices. The remainder of the paper discusses about the background and knowledge gaps, presents our proposed methodology, and analyzes the experimental results.

2 Background and Knowledge Gaps

Automatic keyphrase extraction is the process of selecting important and topical phrases from the body of

a document [2]. Keyphrase extraction starts with identifying a list of candidate phrases using some heuristics rules, which rely on syntactic features like part of speech tags, and/or statistical features like the frequency of n-grams [3]. There are commonly two different approaches to determine which candidates are correct keyphrases: supervised and unsupervised approaches. Supervised approaches utilize supervised machine learning algorithms to learn how to extract keyphrases from pre-labeled training documents and formulate the keyphrase extraction task either as a text classification problem or a sequence labeling problem [3]. For example, John et al. [4] presented a multi-feature supervised automatic keyphrase extraction system which uses a combination of statistical, linguistic, syntactic, semantic, and topic-based features with a Random Forest classifier. Zhang et al. [5] proposed a deep recurrent neural network (RNN) model that jointly extracts and ranks keyphrases from tweets based on keywords and context information.

As labeled training documents can be hard to obtain, unsupervised approaches have also been widely adopted, which commonly identify keyphrases using graph-based ranking method or topic-based clustering method [3]. For example, Liu et al. [6] proposed the topical PageRank (TPR) model, which decomposes traditional PageRank into multiple topic-specific PageRanks and extracts keyphrases based on their importance scores to different topics. Mahata et al. [7] proposed the Key2Vec model which trains multi-word phrase embeddings as thematic representation of scientific articles and ranks extracted keyphrases using theme-weighted PageRank algorithm.

Despite the above-mentioned research efforts, two major knowledge gaps for automatic keyphrase extraction have been identified: (1) most of the research works focus on extracting keyphrases from scientific papers or social media posts and have not yet been evaluated on public reports. Public reports are documents developed by public agencies such as department of transportation, metropolitan planning organization, and emergency management department. Public reports provide us an opportunity to understand the view of disaster resilience from public agencies perspective; (2) most of the existing works rank the importance of candidate phrases based on statistical and document structure features, which has limited capability to incorporate the semantics of candidate phrases into the evaluation process.

3 Proposed Automatic Keyphrase Extraction Methodology

To address the above-mentioned knowledge gaps, the paper proposes an automatic keyphrase extraction methodology that extracts relevant phrases on disaster resilience concepts from documents in infrastructure domain. In preparing the document collection, public reports developed by important agency stakeholders during the infrastructure planning and design process were collected in addition to scientific papers. When identifying keyphrases for disaster resilience concepts, a semantic similarity measure is proposed to capture the semantics of candidate phrases in both the general and infrastructure domain. The proposed methodology includes six primary tasks: data collection, data preparation, reference hierarchy selection, candidate phrase extraction, and keyphrase ranking.

3.1 Data Collection

To create a document collection, the keyphrase "infrastructure disaster resilience" was used to search for scientific papers and public reports from search engines including Google and Google Scholar. The titles of the public reports and scientific papers collected are shown in Table 1 and 2. In total, the document collection contains 11 public reports developed by 8 agencies, and 8 scientific papers. For each paper or report, the original pdf file was first converted to a txt file. Only the textual contents in the main body were kept excluding the figures, table of contents, references, and appendix.

3.2 Data Preparation

To prepare for the implementation of keyphrase extraction, each document in the collection is divided into individual sentences based on sentence boundaries, such as period, question mark, and exclamation mark. Each sentence is further divided into individual tokens (e.g., words and punctuations), and every word is converted into its lowercase form. Lemmatization is then conducted to remove inflectional endings of a word, and return its base or dictionary form, which is known as the lemma. For example, after the lemmatization, the words "rebuilds", "rebuilt", and "rebuilding" would all be transformed into their lemma "rebuild".

3.3 Reference Hierarchy Selection

In order to identify keyphrases relevant to disaster resilience, a reference concept hierarchy is used to provide additional semantic information for the keyphrase ranking process. As there is no existing concept hierarchy available for the infrastructure disaster resilience, the community disaster resilience hierarchy (taxonomy) developed by Taeby and Zhang [8] was selected due to its similarity to the domain. As shown in Figure 1, the selected reference hierarchy includes 29 practices (R1 – R29) to enhance the disaster

resilience in residential communities.

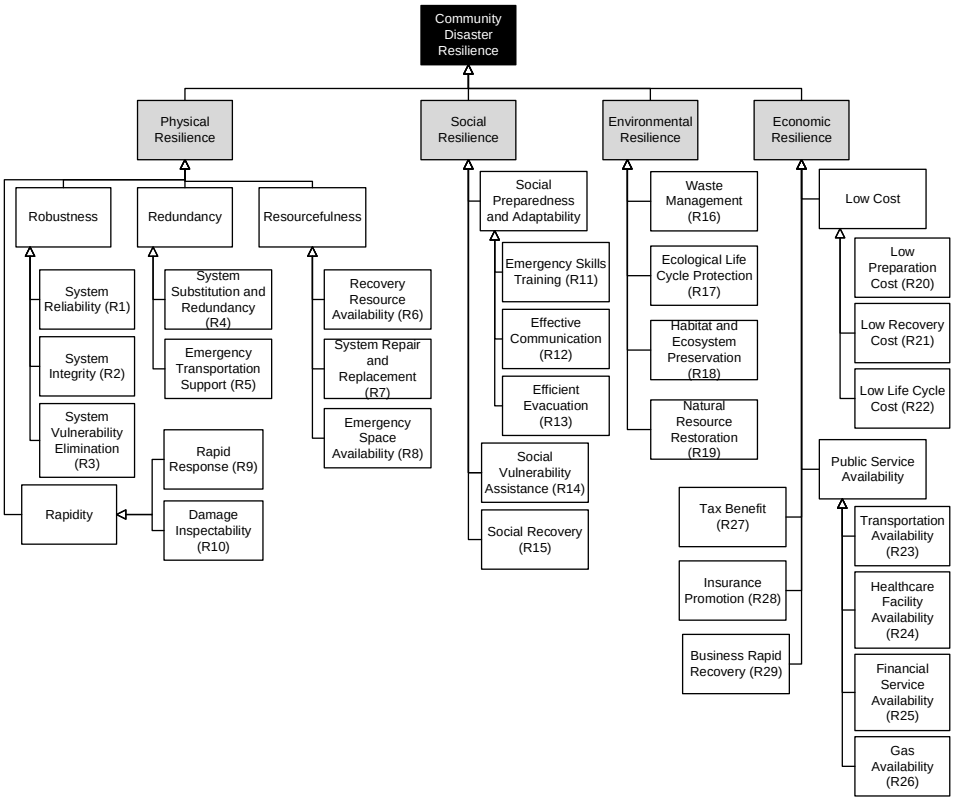

Figure 1. Community disaster resilience hierarchy [8]

Table 1. Public reports in the document collection

Public report	Reporting agency
• A Charrette on Florida's Future New Corridors	• Florida Department of Transportation (FDOT)
• Assessing Criticality in Transportation Adaptation Planning	• Federal Highway Administration (FHWA)
• Assessment of Key Gaps in the Integration of Climate Change Considerations into Transportation Engineering	• FHWA
• Community Impact Assessment: A Quick Reference for Transportation	• FHWA
• FHWA Climate Resilience Pilot Program - Alaska DOT	• Alaska DOT and FHWA
• FHWA Climate Resilience Pilot Program - Michigan DOT	• Michigan DOT and FHWA
• FHWA Climate Resilience Pilot Program - Oregon DOT	• Oregon DOT and FHWA
• MnDOT Flash Flood Vulnerability and Adaptation Assessment Pilot Project	• Main DOT
• Performance Report FDOT 2017	• FDOT
• Post disaster redevelopment plan	• Florida Department of Community Affairs • Florida Division of Emergency Management
• Post Hurricane Sandy Transportation resilience Study in	• FHWA

NY NJ Con

Table 2. Scientific papers in the document collection

Scientific paper
• A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities
• A review of definitions and measures of system resilience
• Critical Infrastructure, Interdependencies, and Resilience
• Fostering resilience to extreme events within infrastructure systems Characterizing decision contexts for mitigation and adaptation
• Resilience and Sustainability of Civil Infrastructure Toward a Unified Approach
• Resilience of Critical Infrastructure Elements and Its Main Factors
• Review on resilience in literature and standards for critical built-infrastructure
• Robustness, Adaptivity, and Resiliency Analysis

3.4 Candidate Phrase Extraction

The Stanford CoreNLP toolkit [9] is used to extract all the noun phrases in each sentence from the collection. The following actions are then adopted to filter the extracted noun phrases:

- Remove the starting token of a noun phrase, if it (1) is a determiner (e.g., "a", "the", "this", and "that"); (2) is a number; or (3) belongs to a standard list of English stop word
- Remove all single-word noun phrases
- Keep only the unique phrases

After the filtering procedure, the resultant noun phrases form the list of candidate phrases. For example, for the sentence "these critical facilities include water and power lifeline, acute-care hospital, and organization that have the responsibility for emergency management at the local community level", the candidate phrases extracted are "critical facility", "water and power lifeline", "acute-care hospital", "emergency management", and "community level".

3.5 Keyphrase Ranking

To select phrases relevant to disaster resilience, a candidate phrase p_i is first represented as a global semantic vector $\widehat{V_i}$ and a local semantic vector V_i. A semantic vector is a real-valued vector of features that characterizes the meaning of a candidate phrase. A global semantic vector represents the contexts in which the candidate phrase appears in the general domain corpus. The global semantic vector is the weighted aggregation of the global word embeddings of its terms.

$$\widehat{V_i} = \{\cup_{j=1}^{m} w_j \, \hat{E}(t_j)\} \quad (1)$$

As per Equation (1), $\hat{E}(t_j)$ is the global word embedding for term t_j, w_j is the weight term of t_j, and m is the total number of terms in phrase p_i. For each term, the global word embedding is obtained from the pre-trained Fasttext embeddings on Wikipedia [10]. An entropy-based weight is adopted to accommodate terms with different contribution to the semantics of a candidate phrase. As per Equation (2) [11], $P(t_j|D_B)$ is the probability of the term t_j appears in the background sentences, and $P(t_j|D_T)$ is the probability of the term t_j appears in the thematic sentences. To differentiate thematic sentences from the background sentences, a set of keywords represent the concepts from the reference hierarchy was created. If the sentence contains any of the pre-defined keywords, it would be considered as a thematic sentence, otherwise a background sentence. When forming the global semantic vectors, common terms (words have similar probabilities to appear in both background and thematic sentences) would have lower weights and distinctive terms (words have very different probabilities to appear in background and thematic sentences) would have higher weights.

$$w_j = -P(t_j|D_B)\log\left(P(t_j|\,D_B)\right) - P(t_j|D_T)\log\left(P(t_j|\,D_T)\right) \quad (2)$$

A local semantic vector represents the contexts in which the candidate phrase appears in the domain-specific corpus. For a candidate phrase p_i, its local semantic vector V_i is defined in Equation (3), where $E(p_i)$ is the multi-word phrase embedding directly obtained using Fasttext embedding model [10] trained on the document collection. During the training process, a candidate phrase is treated as a single token and added to the vocabulary of the embedding model, a multi-word phrase embedding can thus be obtained directly without aggregating embeddings from term level.

$$V_i = E(p_i) \quad (3)$$

To select the keyphrases relevant to disaster resilience, a semantic similarity measure is proposed to incorporate the semantics of candidate phrases in both the general and infrastructure domain. For a candidate phrase pi and a concept c_k in the reference hierarchy, their semantic similarity $SS(p_i, c_k)$ is defined in Equation (4), where $\hat{S}(p_i, c_k)$ is the global similarity and $S(p_i, c_k)$ is the local similarity.

$$SS(p_i, c_k) = \hat{S}(p_i, c_k) + S(p_i, c_k) \quad (4)$$

The global similarity measures the semantic similarity between pi and c_k in general domain. As per

Equation (5), the global similarity $\hat{S}(p_i, c_k)$ is defined as the cosine similarity between $\hat{V}_i$ and $\hat{V}_k$, which are the global sematic vectors of p_i and c_k respectively.

$$\hat{S}(p_i, c_k) = \frac{\hat{V}_i * \hat{V}_k}{||\hat{V}_i|| * ||\hat{V}_k||} \quad (4)$$

The local similarity between pi and c_k measures their semantic similarity in the infrastructure domain. As per Equation (6), the local similarity $\hat{S}(p_i, c_k)$ is defined as the cosine similarity between V_i and V_k, which are the local sematic vectors of p_i and c_k, respectively.

$$S(p_i, c_k) = \frac{V_i * V_k}{||V_i|| * ||V_k||} \quad (5)$$

3.6 Evaluation

All the bottom-level sub-concepts (R1 - R10) under physical resilience were selected as the target concepts for evaluating the proposed keyphrase extraction methodology. For each target concept, the top 20 ranked keyphrases were extracted based on their semantic similarities. The authors then evaluated each extracted keyphrase and determined its relevance to the target concept based on the majority votes. The performance of the proposed methodology was then evaluated using precision at top 20, which is defined as the ratio of the number of relevant keyphrases over the top 20 ranked keyphrases.

Table 3. Performance of the proposed methodology

Disaster resilience concepts	Precision at top 20	Example keyphrases
• System reliability (R1)	85%	• system robustness, utility system
• System integrity (R2)	55%	• building integrity, structural integrity
• System vulnerability elimination (R3)	65%	• direct vulnerability reduction, vulnerability assessment
• System substitution and redundancy (R4)	50%	• power supply redundancy, structural redundancy
• Emergency transportation support (R5)	15%	• emergency transportation lifeline safety plan, public transportation emergency relief program
• Recovery resource availability (R6)	75%	• recover service, traditional disaster recovery funding
• System repair and replacement (R7)	100%	• rapid repair technology, housing repair
• Shelter availability (R8)	30%	• emergency shelter, special need shelter
• Rapid response (R9)	100%	• emergency response time, emergency response personnel
• Damage inspectability (R10)	85%	• damage evaluation, damage detection technology

4 Experimental Results and Analysis

For each target concept, the performance of the keyphrase extraction and two example keyphrases extracted are shown in Table 3. The average precision at top 20 for robustness (R1, R2, R3), redundancy (R4, R5), resourcefulness (R6, R7, R8) and rapidity (R9, R10) concepts are 68%, 33%, 68%, and 93% respectively. The proposed methodology extracted much fewer relevant phrases on redundancy concepts compared with on other concepts. This indicates that the redundancy concepts are less domain-specific and have a high level of ambiguity. For example, the concept "emergency transportation support" has the lowest precision at top 20 (15%) among the 10 target concepts evaluated. Many phrases related to transportation system are falsely extracted as keyphrases, such as "mass transportation system", and "key transportation infrastructure". On the other hand, concepts "system repair and replacement", and "rapid response" achieved 100% of precision at top 20 because of the less ambiguity they have in the infrastructure domain. Overall the proposed methodology achieved 66% of precision at top 20 extracted keyphrases on average.

5 CONCLUSION AND FUTURE WORK

This paper presents an automatic keyphrase extraction methodology to identify relevant phrases on disaster resilience from documents in infrastructure domain. In developing the proposed methodology, each candidate phrase is represented as a global semantic vector and a local semantic vector. A semantic similarity measure is proposed to incorporate the semantics of candidate phrases in both general and infrastructure domain. When evaluated on the document collection, the proposed methodology achieves 66% on average in terms of precision at top 20 extracted keyphrases.

In the future work, the authors will continue to improve the current work in four directions: (1) conduct surveys or focus-group meetings with experts on infrastructure disaster resilience to further validate the keyphrases extracted; (2) evaluate the proposed methodology on more disaster resilience concepts with larger document collection; (3) investigate how the extracted keyphrases are similar or different between scientific papers and public reports; (4) explore how the extracted keyphrases could facilitate the automatic construction of disaster resilience ontology or taxonomy.

References

[1] NRC (National Research Council). Disaster resilience: A national imperative. Washington, DC: National Academies Press, 2012.

[2] Peter Turney. 2000. Learning algorithms for keyphrase extraction. Information Retrieval, 2:303–336.

[3] Hasan KS, Ng V. Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, (1):1262-1273, 2014.

[4] John AK, Di Caro L, Boella G. A supervised keyphrase extraction system. In Proceedings of the 12th International Conference on Semantic Systems, 57-62, 2016.

[5] Zhang Q, Wang Y, Gong Y, Huang X. Keyphrase extraction using deep recurrent neural networks on Twitter. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing: 836-845, 2016.

[6] Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. Automatic keyphrase extraction via topic decomposition. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 366–376, 2010.

[7] Mahata D, Kuriakose J, Shah RR, Zimmermann R. Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (2): 634-639, 2018.

[8] Taeby, M. and Zhang, L. Exploring Stakeholder Views on Disaster Resilience Practices of Residential Communities in South Florida. Natural Hazards Review, 20(1): 04018028, 2018.

[9] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of Association for Computing Linguistic: System Demonstrations, 55-60, 2014.

[10] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135-46, 2017.

[11] Aggarwal, C.C. and Zhai, C. Mining text data. Springer Science & Business Media, 2012.