Automatic Extraction of Abbreviation for Emergency Management Websites

Min Song
New Jersey Institute of Technology
min.song@njit.edu

Peishih Chang
New Jersey Institute of Technology
peishih.chang@njit.edu

ABSTRACT

In this paper we present a novel approach to reduce information proliferation and aid better information structure by automatically generating extraction of abbreviation for emergency management websites. 5.7 Giga Byte web data from 624 emergency management related web sites is collected and a list of acronyms is automatically generated by proposed system (AbbrevExtractor). Being the first attempt of applying abbreviation extraction to the field, this work is expected to provide comprehensive and timely information for emergency management communities in emergency preparedness, training and education. Future work is likely to involve more data collection and intelligent text analysis for dynamically maintaining and updating the list of acronyms and abbreviations.

Keywords

Abbreviation extraction, web mining, intelligent text analysis, emergency management

INTRODUCTION

After the Asian tsunami, US hurricane Katrina and the Kashmir earthquake, people around the world recognize Internet as a powerful facilitator in relief efforts. There is incredible amount of reports, forums, messages and sites in emergency management fields that grows and multiplies at incredible speed. With a single Google search, 17 millions ‘emergency management’ sites and documents have been listed. Various phases of emergency preparedness, planning, training or education, response, recovery and assessment (Van de Walle and Turoff, 2007) have been extensively explored by practitioners, scholars and even normal citizens (Currion, DeSilva and Van de Walle, 2007; Palen, Hiltz and Liu, 2007). While people are able to share information and coordinate citizen-led efforts in addition to any official government and non-governmental websites (Palen et al., 2007), information proliferation and overload emerges as a new problem for the society. Due to broad aspects of knowledge and the involvement of multiple disciplines (i.e., medical, fire, police and public safety, call coordinators, and etc.) in emergency management (Kristensen, Kyng and Palen, 2006), much effort has been devoted in manually creating guide to emergency management and identifying related terms, definitions and acronyms. Many emergency websites tend to each provide a glossary of emergency terms, yet the list is neither definitive nor exhaustive.

While it is not difficult for a person to identify an abbreviation, although sometimes time consuming, it is a far more difficult question to determine which ‘meaning’ was intended. Many questions after hurricane Katrina have been raised by general public. What does ‘FEMA’ stand for? What does ‘AEC’ mean? While the former creates no confusion and could be easily searched using Google, the latter has multiple meanings in different contexts (e.g. *army environmental command, agency emergency coordinators, atomic energy commission*, etc.).

This research is motivated by a belief that ensuring accurate, comprehensive and timely information is likely to lead to better knowledge about emergency management. Abbreviation extraction, one of advanced text mining techniques (Cohen and Hersh, 2005), is used in this study to derive high-quality information from the text. Being the first attempt of applying abbreviation extraction to the emergency field, the proposed system allows the users to concentrate on higher level interpretation and actual action by removing some troublesome and time consuming tasks. A robust system like this may well benefit emergency preparedness and provide timely and structured materials for emergency training and education purpose.

A brief introduction of proposed automatic abbreviation extraction system is followed by a series of illustrations concerning system architecture and techniques used to identify terms and acronyms. The procedure of data collection is then presented, followed by comparisons and analyses of automatically generated acronym list with manually created one. A discussion of directions for future work and fruitful extensions is then concluded in this paper.

ABBREVEXTRACTOR: AN ABBREVIATION EXTRACTION SYSTEM

In this section, we describe the proposed hybrid abbreviation extraction system, called AbbrevExtractor, combining the Support Vector Machine-based (SVM) noun chunking technique with pattern matching techniques. In the following section, we present the system architecture of AbbrevExtractor. In addition, we discuss the noun chunking technique. Finally, Section 3.3 explains our abbreviation extraction algorithm.

The System Architecture

The system architecture of our hybrid abbreviation extraction system, AbbrevExtractor, is illustrated in Fig. 1. AbbrevExtractor consists of seven major components: 1) web crawler, 2) data converter, 3) sentence parser, 4) pos tagger, 5) noun chunker, 6) abbreviation matcher, 7) best-match selector component.

The Noun Chunker component applies a SVM-based text chunking technique. A typical text chunking algorithm seeks a complete partitioning of a sentence into chunks of different types (Kudo and Matsumoto, 2000). Since our chunking technique requires identifying POS (Part-Of-Speech) tags for individual words, we incorporate Brill's POS Tagger into AbbrevExtractor. Brill's technique is one of the high quality POS tagging techniques.

The Best-Match Selector component identifies the correct long form from a set of candidate long forms within the sentence by computing the proximity of candidate long forms to a short form.

The outline of the approach described in Figure 1 is as follows:

1) A list of web sites is provided to collect data.
2) Web data is crawled and downloaded into AbbrevExtractor.
3) Various different forms of data such PDF, WORD, PS and PPT are converted into plain text.
4) Sentences are identified and parsed by the Sentence Parser component.
5) Each sentence is split into phrase groups by the SVM-based noun chunking component.
6) Short forms and candidate long forms are identified by the pattern matching-based Abbreviation Matcher component.
7) Correct long forms are determined and selected from candidate long forms by the best match selector component.
8) A pair of a short form and a long form is inserted into the database of Ontologies for abbreviations.

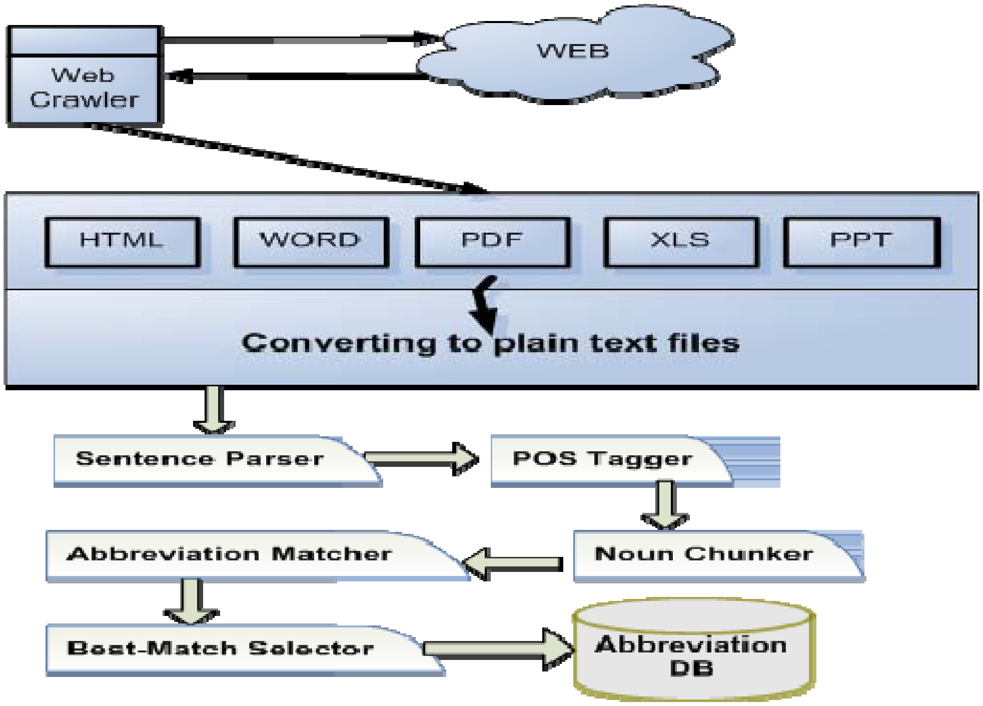


Figure 1. System Architecture

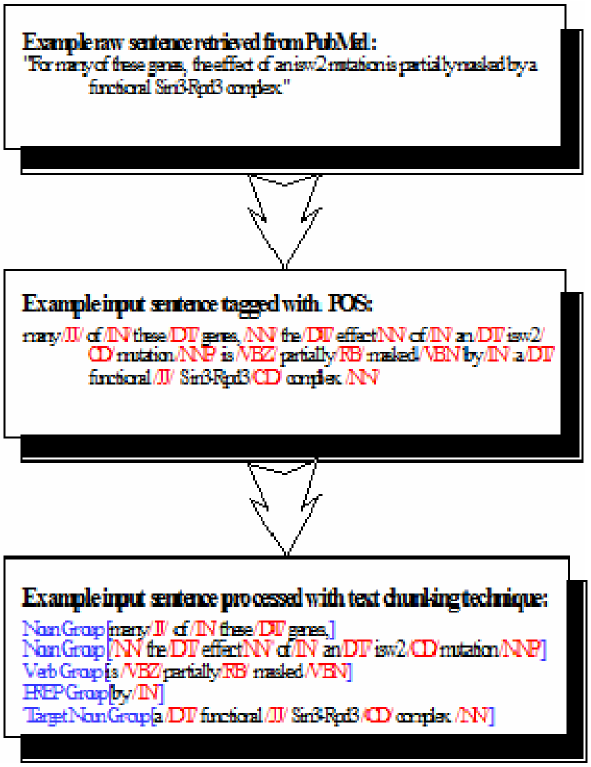
Sentence Chunking by the Support Vector Machine Technique

Text chunking is defined as dividing a text into syntactically correlated parts of words (Kudo and Matsumoto, 2000). Chunking is recognized as a series of processes – first, identifying proper chunks from a sequence of tokens, and second, classifying these chunks into some grammatical classes. Major advantages of using text chunking over full parsing techniques are that partial parsing such as text chunking is much faster, more robust, yet sufficient for abbreviation extraction.

Support Vector Machine (SVM) based text chunking was reported to produce the highest accuracy in the text chunking task (Kudo and Matsumoto, 2000). The SVMs-based approach such as other inductive-learning approaches takes as input a set of training example and finds a classification function that maps them to a class.

SVMs are known to robustly handle large features (Cortes and Vapnik, 1995). This makes them an ideal model for abbreviation extraction. SVMs are particularly useful for real world data sets that often contain inseparable data points. Although training is generally slow, the resulting model is usually small and runs quickly as only the patterns that help define the function that separates positive from negative examples. In addition, SVMs are binary classifiers, and thus we need to combine several SVM models to obtain a multiclass classifier. Due to the nature of the SVM as a binary classifier it is necessary in a multi-class task to consider the strategy for combining several classifiers. In this paper, we use Tiny SVM (Kudo and Matsumoto, 2000) in that Tiny SVM performs well in handling a multi-class task.

Figure 2 illustrates the procedure of converting a raw sentence from PubMed to the phrase-based units grouped by the SVM text-chunking technique. The top box shows a sentence that is part of abstracts retrieved from PubMed. The middle box illustrates the parsed sentence by POS taggers. The bottom box shows the final conversion made to the POS tagged sentence by the SVM based text chunking technique. Table 1 lists a set of tagging types for phrases and its description.



Example raw sentence retrieved from PubMed:
"For many of these genes, the effect of an isw2 mutation is partially masked by a functional Sin3-Rpd3 complex"

Example input sentence tagged with POS:
many/JJ/ of/IN/ these/DT/ genes,/NN/ the/DT/ effect/NN/ of/IN/ an/DT/ isw2/CD/ mutation/NNP/ is/VBZ/ partially/RB/ masked/VBN/ by/IN/ a/DT/ functional/JJ/ Sin3-Rpd3/CD/ complex/NN/

Example input sentence processed with text chunking technique:
NounGroup[many/JJ/ of/IN/ these/DT/ genes,]
NounGroup[/NN/ the/DT/ effect/NN/ of/IN/ an/DT/ isw2/CD/ mutation/NNP/]
VerbGroup[is/VBZ/ partially/RB/ masked/VBN/]
PREPGroup[by/IN/]
TargetNounGroup[a/DT/ functional/JJ/ Sin3-Rpd3/CD/ complex/NN/]

Figure 2. A Procedure of Sentence Parsing (JJ denotes adjective, IN denotes preposition, DT denotes determiner, CD cardinal number, NN denotes singular noun, NNP denotes proper noun, VBZ and VBN denote verb, RB denotes adverb.)

Tagging Type	**Description**
NP_GROUP	Noun phrase group
VP_GROUP	Verb phrase group
PP_GROUP	Preposition phrase group
CONJ_GROUP	Conjunction phrase group
ART_GROUP	Article phrase group
CM_GROUP	Conjunction phrase group
ADV_GROUP	Adverb phrase group
ADJ_GROUP	Adjective phrase group
PART_GROUP	Miscellaneous phrase group

Table 1. Tagging Types

DETERMINATION OF CORRECT SHORT FORMS AND LONG FORMS

AbbrevExtractor selects a short and candidate long forms within noun phrase groups. Given this prerequisite, AbbrevExtractor applies pattern matching-based rules to identify short forms and long forms, similar to ExtractAbbrev and ALICE, which stands for Abbreviation LIfter using Corpus-based Extraction.

In ExtractAbbrev, short forms are selected if the following conditions are satisfied: 1) it consists of at most two terms, and its length is between two to ten characters, at least one of the characters is a letter, and the first character is alphanumeric. Finding correct long forms is based on starting from the end of both the short form and the long form, moving right to left, trying to find the shortest long form that matches the short form. In ALICE, short forms are determined by the rules of nine discard conditions and four acceptance conditions. Long forms are selected by five discard conditions and 16 templates.

Compared to these ExtractAbbrev and ALICE, AbbrevXtractor identifies correct short and long forms within noun phrase groups. The selection rules for short forms are adapted from ExtractAbbrev (Schwartz and Hearst, 2003). The selection rules for candidate long forms are as follows: 1) The first word of the candidate long form is not in the first word list of candidate long forms. 2) The candidate long forms do not consist of only one word in the long form list. 3) The number of words in a noun group less than 10. 4) The characters in short forms are matched in capitalized characters in candidate long forms. 5) Candidate long forms are one word and the name of its POS is CD.

Once all the candidate long forms are identified, we compute distance between short form and candidate long forms, based on order it is presented in the sentence. Figure 3 shows how candidate long forms and a short form are located in a sentence. A candidate long form in the shortest distance with a short form is selected as the best matched long form.

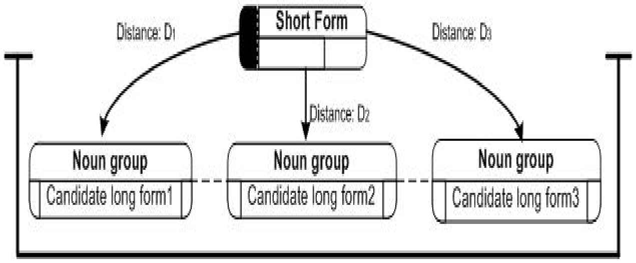


Figure 3. Topology of Short and Candidate Long forms

WEB DATA COLLECTION

We collected about 5.7 Giga Byte web data from 624 emergency management related web sites. The list of web sites was provided by experts of emergency management. For this study, we spidered these sites with depth 1. The depth 1 refers to the first level of the given target web site. Table 2 shows the sample of the list.

WEB SITES
Federal Radiological Emergency Response Plan
Center for Biodefense, George Mason University
Center for State Homeland Security
Chemical and Biological Defense Information Analysis Center (CBIAC)
MEDLINE Plus, Disasters and Emergency Preparedness Web Reference, National Library of Medicine.
Survival Ring, Private Civil Defense Site

Table 2: Web sites that were crawled

Comparison with manually created acronym list for emergency management

We extracted 70250 abbreviations from the 5.7 GB web data. In 70250 abbreviations, multiple appearance of the same abbreviation was permitted. To evaluate the performance of our system, we compared the list of extracted abbreviation with the one manually created for emergency management. Acronyms are complied from various resources. The main resource is Emergency Management Glossary and Acronyms, found at http://lepc.co.lake.in.us/empglos.pdf. The number of acronyms in the manually created list is 856. Out of 70250, the total matched count is 18057 and the total unique matched count 743, which results in 87% accuracy rate.

Table 3 shows the results of automatically extracted abbreviation in XML form and table 4 is the sample of Gold Standard used for comparison. Our initial experiments show the promising evidence of applying our automatic abbreviation extraction technique to emergency management.

```
<RESPONSE>
  <ENTRY ABBREV="ATSDR" LONG_FORM="Agency for Toxic Substances and Disease
Registry"/>
  <ENTRY ABBREV="CABS" LONG_FORM="Chemical Agent Briefing Sheets"/>
  <ENTRY ABBREV="ATSDR" LONG_FORM="Agency for Toxic Substances and Disease
Registry"/>
  <ENTRY ABBREV="CABS" LONG_FORM="Chemical Agent Briefing Sheets"/>
  <ENTRY ABBREV="CEP" LONG_FORM="Completed Exposure Pathway"/>
  <ENTRY ABBREV="FAQs" LONG_FORM="Frequently Asked Questions"/>
  <ENTRY ABBREV="GIS" LONG_FORM="Geographic Information Systems"/>
  <ENTRY ABBREV="HSEES" LONG_FORM="Hazardous Substances"/>
  <ENTRY ABBREV="MRLs" LONG_FORM="Minimum Risk Levels"/>
  <ENTRY ABBREV="PAHs" LONG_FORM="Polycyclic Aromatic Hydrocarbons"/>
  <ENTRY ABBREV="CABS" LONG_FORM="Chemical Agent Briefing Sheets"/>
 <ENTRY ABBREV="CAP" LONG_FORM="Community Assistance Panel"/>
  </RESPONSE>
```

Table 3: Sample of Extracted Abbreviations

One issue that needs to be addressed is multiple definitions of an abbreviation, and the term that has a set of different meanings is referred to as polysemy. Our system tackles the polysemy problem by pairing abbreviation and its long forms and presenting them to the users. Given a set of multiple definitions, the user will pick one from the choices. To assist the user to understand the context of where the abbreviation and its long form appear, we plan to present the user the surrounding phrases of the abbreviation.

AAR/BOE Association of American Railroads/Bureau of Explosives
ABHP American Board of Health Physics
ACGIH American Conference of Governmental Industrial Hygienists
ACP Access control point
ACP Access control post
ACRS Advisory Committee on Reactor Safeguards
ADC Authorized Derivative Classifier
ADP Automatic data processing
AEC Agency Emergency Coordinators
AED Aerodynamic equivalent diameter
AEGL Acute Exposure Guideline Level, under development by EPA and the National Academy of Science
AFMIC Armed Forces Medical Intelligence Center
AFOS Automated field operation and services
AHU Air handling unit

Table 4: Sample of Gold Standard

CONCLUSION

Being the first to apply abbreviation extraction to emergency management field, the initial experiments produce promising results. This research proves to be a huge step forward to aid time-consuming tasks and thus provides opportunities to focus our energies to higher level of tasks. Moreover, it contributes to further advance web mining techniques with the proposed hybrid abbreviation extraction system, AbbrevExtractor. It represents a simple, easy to implement and accurate tool set for building emergency abbreviation glossary. The system also provides a solution to address the concerns of handling different meanings in different contexts. We believe this research would benefit emergency communities, and provide timely and structured materials for emergency preparedness, training and education purpose.

Principal extensions of the present study would further advance the contributions of this research program. First, crawling more data and diving deeper in the emergency sites may increase the accuracy rate of the acronym list. The current system is limited to spider only the first level of the sites. The more websites are being crawled, the more exhaustive of the acronym list may become. Second, further evaluation of the system performance using standard precision and recall measures (Cohen and Hersh, 2005) is planed. Finally, the AbbrevExtractor system would be made publicly accessible for emergency management communities. Future study is likely to involve intelligent text analysis for continuous assimilation of new abbreviations and acronyms into domain knowledge bases. In addition, we plan to compare our technique with other techniques proposed in biomedical abbreviation extraction for system evaluation.

ACKNOWLEDGMENTS

Partial support for this research was provided by the National Science Foundation under grant DUE-0434581, by the Institute for Museum and Library Services under grant LG-02-04-0002-04, and by the New Jersey Institute of Technology.

REFERENCES

1. Cohen, A. and Hersh, W. (2005) A Survey of Current Work in Biomedical Text Mining, *Briefing in Bioinformatics,* 6, 57-71.
2. Cortes, C. and Vapnik, V. (1995) Support-vector Networks, *Machine Learning*, 20, 273-297.
3. Currion, P., DeSilva, C. and Van de Walle, B. (2007) Open sources software for disaster management, *Communications of the ACM*, 50, 3, 61-65.
4. Kristensen, M., Kyng, M. and Palen, L. (2006) Participatory design in emergency medical service: designing for future practice, *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 161-170.

5. Kudo, T. and Matsumoto, Y. (2000) Use of Support Vector Learning for Chunk Identification, *Proceedings of the CoNLL-2000 and LLL-2000*, 142-144.
6. Palen, L., Hiltz, S. R. and Liu, S. B. (2007) Online forums supporting grassroots participation in emergency preparedness and response, *Communications of the ACM*, 50, 3, 54-58.
7. Schwartz, A. S. and Hearst, M. A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text, *Proceedings of the Pacific Symposium on Biocomputing*, Kauai, Hawaii, 8, 451-462.
8. Van de Walle, B. and Turoff, M. (2007) Emergency response information systems: emerging trends and technologies, *Communications of the ACM*, 50, 3, 29-31.