DIMPLE: DIsaster Management and Principled Large-scale information Extraction

Workshop Programme

31st May 2014

INTRODUCTION

09:00-09:15 Khurshid Ahmad, *Introduction: Disasters, Ethics, Terminology and Ontology*

ONLINE INFORMATION SYSTEMS & DISASTER MANAGEMENT

09:15– 09:35 Alexander Lörch & Mathias Bank, *Topic Analyst® - A framework for recognizing early warnings*

09:35– 09:55 Enrico Musacchio & Francesco Russo, *An Emergency Management System: Sistema Informativo Gestione Emergenze Protezione Civile*

09:55– 10:10 Questions & Discussion

COMMUNICATIONS DURING AND AFTER DISASTERS AND EMERGENCIES

10:10– 10:30 Henrik Selsøe Sørensen, *Multi-Lingual and Multi-Cultural Aspects of Post-Disaster Emergency Communication - the LinguaNet® Experience*

10:30 – 11:00 Coffee break

11:00 -11:20 Maria Teresa Musacchi, *Social media and disaster management: US FEMA as a benchmark for its European counterparts?*

11:20 – 12:00 Cilian Fennell, *How Communications Companies Can Help Organisations Prepare for Disasters*

1140 – 12:00 Maria Grazia Busa & Sara Brugnerotto, *Italian doctor-patient interactions: a study of verbal and non-verbal behavior leading to miscommunication*

12:00 – 12:15 Questions & Discussion

TRUST IN AND OF THE SOCIAL MEDIA

12:15 – 12:35 Carl Vogel, *Anonymous FreeSpeech*

13:35 – 12:55 Martin Mackin & Sadhbh McCarthy, *Trust & Transparency in Social Media for Disaster Management*

12:55 – 13:10 Questions & Discussion

13:10 – 14:00 Lunch

TERMINOLOGY AND ONTOLOGY OF DISASTERS

14:00 – 14:20 Bodil Nistrup Madsen & Hanne Erdman Thomsen, *Terminological Ontologies for Risk and Vulnerability Analysis*

14:20 – 14:40 Xiubo Zhang & Khurshid Ahmad, *Ontology and terminology of disaster Management*

14:40 – 15:00 Questions & Discussions

INFORMATION EXTRACTION FROM DISASTER DOCUMENT AND MEDIA STREAMS

15:00 – 15:20 Lars Döhling, Jirka Lewandowski & Ulf Leser, *A Study in Domain-Independent Information Extraction for Disaster Management*

15:20 – 15:40 Daniel Isemann, Andreas Niekler, Benedict Preβler, Frank Viereck & Gerhard Heyer, *OCR of Legacy Documents as a Building Block in Industrial Disaster Prevention*

15:40 – 16:00 Stephen Kelly & Khurshid Ahmad, *Determining levels of urgency and anxiety during a natural disaster: Noise, affect, and news in social media*

16:00 – 16:30 Coffee break

16:30 – 17:00 Discussion

17:00 Workshop Closed

Editors

Khurshid Ahmad	Trinity College Dublin, IRELAND.
Carl Vogel	Trinity College, Dublin, IRELAND.

Workshop Organizers/Organizing Committee

Khurshid Ahmad	Trinity College Dublin, IRELAND.
Carl Vogel	Trinity College, Dublin, IRELAND.

Workshop Programme Committee

Khurshid Ahmad	Trinity College Dublin, IRELAND.
Gerhard Heyer	University of Leipzig, GERMANY
Linda Hogan	Trinity College, Dublin, IRELAND.
Bodil Madsen	Copenhagen Business School, DENMARK.
Sadhbh McCarthy	Centre for Irish and European Security, Dublin, IRELAND.
Maria Teresa Musacchio	University of Padova, ITALY.
Henrik Sørenson	Copenhagen Business School, DENMARK.
Carl Vogel	Trinity College, Dublin, IRELAND.

Table of contents

Author Index

Introduction: Disasters, Ethics, Terminology and Ontology

Khurshid Ahmad
School of Computer Science and Statistics, Trinity College Dublin, IRELAND
kahmad@cs.tcd.ie

This workshop addresses the use and implications of the use of technology in the management of the disaster life cycle: starting from warnings about impending disasters to suggestions about recovery. The technology issue has become more poignant with the advent of social media and its continually increasing use in disasters across the world. Consider the major disasters of this century, which has just begun, including hurricanes in the USA[1,2], earthquakes in Haiti[3,4], and tsunamis in Japan[5,6]. In each of these cases there is documented evidence that social media is quite helpful in disseminating life and business critical information quickly and effectively. The citizen is playing or should play an active role in monitoring, averting and rehabilitating before, during and after a disaster[7]. One learns from the disaster archives referenced above that there is a need for an ethically well-grounded and accessible system that can harness the limitless data that streams through the social media and the formal media.

To set the scene two systems are presented that have a potential for improving communications during a disaster: Alexander Lörch & Mathias Bank (CID, Germany) present their social media monitoring and search system Topic Analyst®, and Enrico Musacchio & Francesco Russo (Data Piano, Italy) present their emergency management system that can be interfaced to geographical information systems and to databases comprising information relating to hazards for instance.

The use of social media requires an understanding of how human beings communicate during a disaster where the goal is to communicate maximal amount of information, often across cultural and linguistic barriers, whilst ensuring that the communication does not unnecessarily alarm the recipients of the information. In this context Henrik Sørensen (CBS, Denmark) informs us about languages that have developed specially for communications between stakeholders in life- and business-critical missions especially in a multi-lingual and multi-cultural environment. Teresa Musacchio (Univ. Padova, Italy) continues with the multi-lingual theme and analyses the English and Spanish language documents prepared by the US Federal Emergency Management Agency(FEMA) together with the various blogs on the topics in the documents to describes the strategies used in communicating different kinds of disaster-related information. Cilian Fennell (Stillwater, Ireland) relates his experience in the public information and media management

---

[1]Imran, Muhammad, Shady MamoonElbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. "Extracting information nuggets from disaster-related messages in social media." In *ISCRAM'13: Proceedings of the 10th International ISCRAM Conference,* Vol. 26. 2013.

[2]Grinberg, Nir, MorNaaman, Blake Shaw, and GiladLotan. "Extracting diurnal patterns of real world activity from social media."In *Proc. of the 7th Int. AAAI Conference on Weblogs and Social Media (ICWSM–13)*. 2013.

[3]Chunara, Rumi, Jason R. Andrews, and John S. Brownstein. "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak." *The American Journal of Tropical Medicine and Hygiene* 86, no. 1 (2012): 39-45.

[4]Hughes, Amanda Lee, and LeysiaPalen. "Twitter adoption and use in mass convergence and emergency events." *International Journal of Emergency Management* 6, no. 3 (2009): 248-260.

[5]Ng, Kwan-Hoong, and Mei-Li Lean. "The Fukushima nuclear crisis reemphasizes the need for improved risk communication and better use of social media." *Health physics* 103, no. 3 (2012): 307-310.

[6]Johansson, Fredrik, Joel Brynielsson, and Maribel NarganesQuijano."Estimating citizen alertness in crises using social media monitoring and analysis." In *Intelligence and Security Informatics Conference (EISIC), 2012 European*, pp. 189-196. IEEE, 2012.

[7]Swigger, Nathaniel. "The Online Citizen: Is Social Media Changing Citizens' Beliefs About Democratic Values?." *Political Behavior* 35 (3) (2013): 589-603.

business to disaster management. Maria Grazia Busà and Sara Brugnerotto (Univ. Padova, Italy) describe their experience of verbal and non-verbal doctor-patient communications and the issue of trust between the two stakeholders in this life-critical business.

The issue of *trust* is the theme of the next two papers in this workshop. Carl Vogel (Trinity College Dublin, Ireland) discusses the relevance and importance of *anonymous free speech* and describes a system that can facilitate such communication whilst conforming to various norms of a civilised society. Martin Mackin and Sadhbh McCarthy (CIES, Ireland) continue with the notion of trust and emphasise the how important is to engage the citizenry at large.

We then turn to the organisation of knowledge relating to disaster, referred to as the ontology of disaster, and show how this organisation can be discerned from a systematic examination of domain specific texts. Bodil Nistrup Madsen and Hanne Erdman Thomsen discuss their work based on the documents of the Danish Emergency Management Agency and how this work will facilitate an analysis of the potential risks of disasters and an analysis of the vulnerabilities. Xiubo Zhang and Khurshid Ahmad continue with the theme of text-based ontology- and terminology-extraction and outline algorithms involving frequency analysis of texts for creating candidate ontologies.

The scale of information propagated via social media exceeds the limits of any possible unaided human professional monitoring. The automatic extraction of information from documents and streams related to disasters is the theme of the concluding session of this workshop: Lars Döhling, Jirka Lewandowski and Ulf Leser (Humboldt-Universität Berlin, Germany) argue that disaster information extraction systems are often crafted and trained for one specific type of disaster: for instance, *earthquake* information management system cannot be easily adapted as a flood information management system; they describe methods for domain independent information extraction. Daniel Isemann and colleagues (Universität Leipzig, Germany) look at the extraction of information from restored versions of documents destroyed or damaged due to incidents or aging and describe a strategy that involves optical character reading. Finally, the concluding paper (by Stephen Kelly and Khurshid Ahmad, Trinity College Dublin, Ireland) deals with the issues raised in this workshop related to trust, related to ontology/terminology, and to information extraction: The authors deal with (a) the potentially intrusive nature of information extraction systems, that is their ability to extract and store data about named *entities* –names of people, places and objects, and (b) systems that can extract candidate affect terms – *negative/positive, strong/weak* and *active/passive* affect- in the documents related to disasters. They demonstrate their methods by looking at hurricane and earthquake reports.

The participants in this workshop are from the academia and from small-to-medium size enterprises (SME) in social media and in emergency management, with particular interest in disaster management: The academics have their own discourse style and the SME's have their own.

We gratefully thank EU sponsored Slándáil Project (FP7 Security sponsored project #6076921, 2014-2017).

Topic Analyst® - A framework for recognizing early warnings

Alexander Lörch, Mathias Bank
CID GmbH
Gewerbepark Birkenhain 1, 63579 Freigericht
E-mail: a.loerch@cid.de, m.bank@cid.de

Abstract

The enormous quantity of information available today offers an unbelievable treasure of knowledge, which can provide relevant information to companies. The amount of data is however also a problem as it makes it very difficult to identify the relevant data and thus to generate knowledge. We present a comprehensive analysis tool named Topic Analyst® which enables the user to interactively investigate a huge amount of data. It enables the user to identify available topics, their trends and their tonality to make informed decisions.

Keywords: early warnings, alerting, topic analysis, semantic analyses, trend detection

1. Introduction

The enormous quantity of information available today offers an unbelievable treasure of knowledge, which can provide relevant information to companies – "can" is the crucial word: It is not known which source may possibly publish information with a certain relevance and when – but it could. Analysts constantly face the challenge of continuously monitoring all sources which could potentially provide relevant information in order to actually obtain the desired message at the right time.

Therefore, the fundamental task is to more precisely define your own information needs in order to support them as comprehensively and broadly as possible by automated information gathering.

What's going on with...?

- Competitors, customers and suppliers
- Legislation and court rulings
- Research and development
- Patent and brand registrations
- Other sectors

but also:

- disaster management and emergency relief

The sources required for information gathering and the methods for continuously building up the knowledge base can be defined on the basis of such an information concept.

In this process, we highlight the following aspects:

- Web Crawling & Search
- Why can Google, Google Alerts & Co. only supply impulses?
- How can crawlers recognize and "gather" information on the web?
- Which challenges exist, e.g. technically and legally?
- How can information from so-called "Deep Web", hence from closed information sources and social media, for example, be added?
- Internal Interfaces
- How can intranet systems such as SharePoint, Confluence and file stores be included?

2. Information Gathering & Analysis

Crawlers ensure the continuous collection of information from as many potentially relevant sources as possible and build up a transparent, quality-secured information basis. Software for automated language processing analyses this primarily textual information, structures its contents, semantically links documents with persons, companies, products, etc., thus allowing for the most efficient search possible.

Analysis functionality based on semantic and statistical algorithms enables a comprehensive investigation of a large amount of textual data and the identification of the individual relevant aspects, topics and trends including their tonality and correlation with additional topics and aspects. In addition, semantic structures make it possible to combine various data types in order to enable combined consideration of news reports and revenue changes.

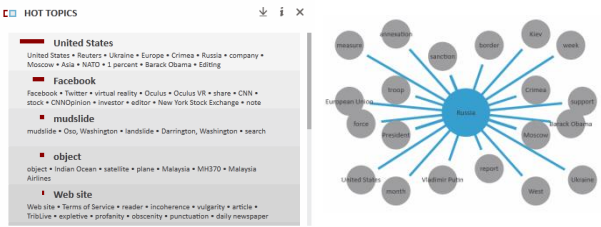


Figure 1: Sematic and statistical algorithms enable different insights to the underlying textual data. The example shows dynamically calculated clusters and co-occurrence graphs based on a data set which is interactively filtered by the user.

3. The analysis cockpit “Topic Analyst®”

The information explorer and analysis cockpit Topic Analyst® is a generic system that continuously imports data from a multitude of sources – such as the Internet – and uses this data to automatically create real-time analyses and reports. Some of its functions include the tracking of positive or negative reporting about topics important to you in the media, the analysis of opinions in social media networks or the recognition of trends in further business-relevant sources.

Topic Analyst® offers comprehensive visualization options, so-called "Micro Analysts", for topics, trends, current entities and in particular for the display of relationship networks, such as those between people (cf. figure 1). These networks enable an efficient navigation between individual aspects and quickly introduce new knowledge for further research. Topic Analyst® also links single documents with other thematically related documents in order to link individual pieces of information to each other.
The Micro Analysts include display forms such as graphics, network displays, charts and graphs in order to make information quickly accessible. The user can combine several displays onto personal dashboards in order to compare information between data quantities or to constantly keep an eye on current information (cf. figure 2). Since Topic Analyst® constantly records new information, dashboards and the Micro Analysts contained on them are continuously updated. In this way new trends are made pro-actively visible.

Figure 2: Topic Analyst® supports several different analyses based on semantic and statistical algorithms, visualized in individual “Micro Analysts”. These can be arranged flexibly on different dashboards.

Topic Analyst® additionally offers a semantic search with dynamic filter options ("faceted browsing") which can be used to target specific types of documents. Along with the automated offering of information without the need of a manual search, a semantically supported search is also available to the user.

Semantic Language Processing & Cross-Linking of Information

- Semantic Language Processing & Cross-Linking of Information
- Assurance of a high-quality information basis
- Inclusion of existing specialized knowledge
- High-precision recognition and monitoring of competitors, products, technologies and more
- Correct linking of different types of data via semantics
- Linking of structured data such as industry sales data or advertising expenditures with unstructured data for a combined analysis Interactive and configurable dashboard

Semantic Search

- Recognition of semantic concepts for significantly increasing the quality of searches during research ("medicine" - medication or hygienics?)

Analysis

- Usage of indicators to identify topics, trends & tonality in real time
- Uncovering "hidden" information which would otherwise get lost in the background noise of the vast sea of information
- Combination of "classic" BI key performance indicators with (text) information-based indicators
- Alerting
- Automated notification of new information and market changes

Using Topic Analyst®, you actively explore huge amounts of information using interactive analysis views that offer overview and insights just at a glance. During a search, suitable matches are not going to be output according to intransparent criteria of relevance anymore, issues in texts and also numerical values can rather be linked to each other and prioritized according to their significance. Then they are ready for a graphical evaluation in many ways. Within big data, you can thus deduce previously unknown correlations from a word cloud or bar charts, for instance, without difficulty. With the support of these intelligent visual analysis tools you can variously influence the displayed search results to make them usable for your informational purposes. And if you should need to collect all the necessary information about a particular subject - Topic Analyst® makes transparent what potentially could be found, and therefore, it is also ideally suited to a systematic and focused ad-hoc search.

An emergency management system: Sistema informativo gestione emergenze protezione civile

Enrico Musacchio and Francesco Russo
Datapiano S.r.l., San Don à di Piave
Galleria Progresso 5, 30027 San Don à di Piave (Italy)
enrico.musacchio@protecoeng.com;francescorusso@datapiano.it

abstract

The software available today for managing civil protection actions during disasters emergencies is based on the collection of data about staff , work vehicles and rescue equipment , implemented and ordered in a structured database , used to raise the various resources , directing and coordinating them with human intervention. In many cases , the software is also able to display by means of interactions with a GIS, locating resources and means on the territory and possibly pre-established risk scenarios . In the light of current knowledge and awareness of the availability of modern facilities and web-based services that provide real-time interaction with the outside world , the described methodology of approach to the problem on which are based on the available software should be considered outdated and insufficient to ensure timely reaction and management of real-time communications. According to the developments of our research team , the approach to the problem of civil defense emergency management must be fully modernized , opening it to new technologies and services available, so that the reaction to adverse events can take place in real time and emergency operators can actually interact with the outside world and with social media, managing external communications and rescue operations with the necessary authority.

Introduction

We have developed a system for the management of civil protection emergencies, enabling real-time performance of all the functions necessary for the operation of a control centre. *Sistema informativo gestione emergenze protezione civile* (S.I.G.E) comprises fully-relational real-time information system based on web-interfaced GIS of a given locality. Once S.I.G.E is populated with data , it allows an operator to interact with it in real time (see Figure 1).

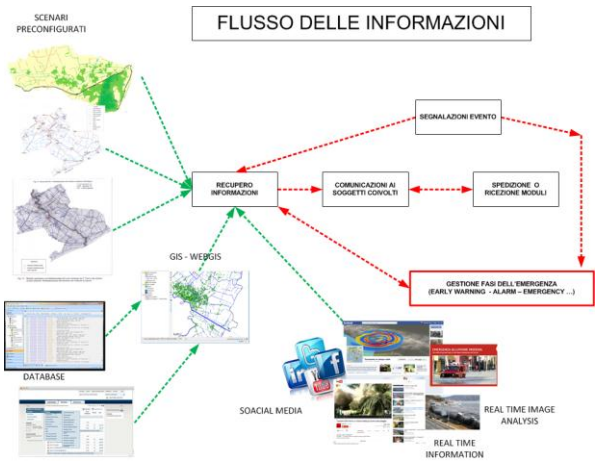


Figure 1: The architecture of S.I.G.E. system showing access to a GIS, strategic databases, and sources of real-time sensor information linked to visualization system.

The operator can not only view the default risk scenarios, but with S.I.G.E can obtain (a) available data relating to persons involved in a disaster, and (b) operational resources or emergency staff. Information can be retrieved not only from pre-configured databases or pre-defined scenarios, but also from social media, real time external modules performing particular tasks and focussing constantly on emergency management.

As described in the figure 1 (above), software can collect data from many sources, and then use all the data to generate reports that may contain images as well. Given the legal constraints of emergency management and concomitant disaster management methods and techniques, S.I.G.E. helps an operator to manage communications protocols for compiling and sending relevant information automatically, by different media, to all the stakeholders including the first responders and the victims. The decisions made by the operators can be tracked and verified in real-time. The system records all information received and sent and maintains an audit trail.

The interaction with the web and the use of a WEB-GIS system, allowing the real-time display of all information, and interact with social media to manage communications. The result is a real time verifying of any effects of disaster managers actions.

S.I.G.E Knowledge base

S.I.G.E contains a knowledge base of rules and associated data sets that encode all the (reasonable) actions that can help in performing key tasks in disaster management.

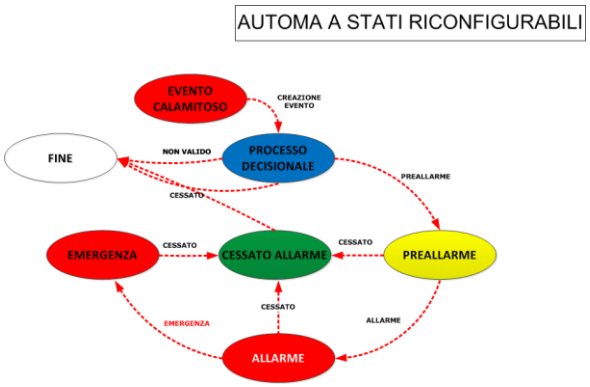


Figure 2: a schematic representation of automaton.

Both actions (the text in the ovals in the figure 2 above) and rules (the arrows) are fully editable, so that software can manage any possible event, applying any possible rule.

In the following figure 3 is represented an evolution of the automaton, as an example of software capabilities. Even when actions and rules become complex, software requires only a different data table editing.

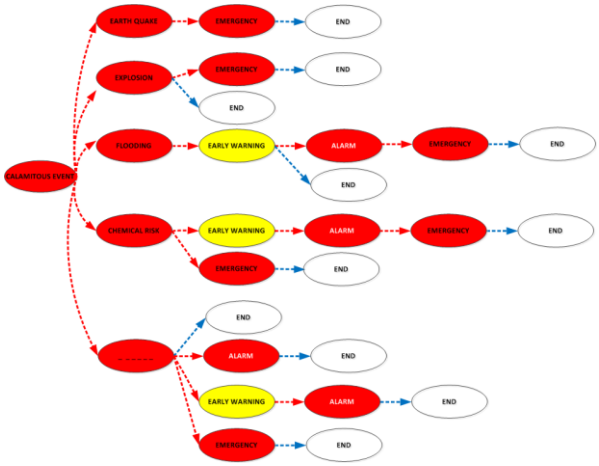


Figure 3: a schematic representation of a more complex automaton.

Disaster Risk Management in S.I.G.E

The software prototype is also equipped with special modules , which can generate real-time risk areas on the basis of specific events (explosions, spills of toxic substances , earthquakes, accidents during dangerous transports etc.). All information, both pre-defined or real time set, may be presented on GIS so you can view the affected areas and immediately identify people or property at risk. After that you can automatically generate appropriate communications and send them to all involved subjects, according to the instructions of the operating room, both internal management of civil protection) and external (direct to the public). This information can also contain instructions for behavior to better face the emergency in the “warm” phase of it.

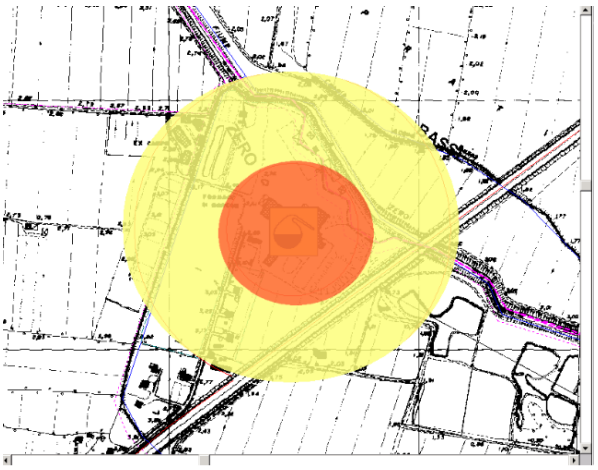


Figure 4 – A simple external module perform real time calculation of toxic gas dispersion in absence of wind and immediately represent the result on GIS screen, over a territory map.

Given an actual or potentially disaster impacted area, our system can provide an estimate of the size of the population in the area by linking directly to databases that comprise population details like birth and death registrars data bases and electoral rolls. This estimate can be transmitted instantly to the relevant disaster managers for immediate action.

Future Developments

Modules are being developed that will allow the software to receive information in real time from social media, through analysis of texts and images (conforming to ethical and privacy statutes). Using this information , the software can generate useful and authoritative communications addressed to the population outside , making sure to correct, almost in real time, the effect of any inaccuracies or errors.

From the structural point of view , the system works according to a finite state automaton , completely reconfigured according to mode of reaction to the event set by the user for each type of risk to be addressed. This automaton, although bypassable by the user, suggests real-time proper actions for the management of emergencies, avoiding blunders due to emotional operator. The complex relational database set is also equipped with tables for storing information relating to the means , the staff, the characteristics of the territory and so on. All tables can be completely reconfigured by the user, connected to each other and with the main tables of the software to set relationships one to one or one to many, also configurable. For this reason the system is able to "learn" any procedures set for emergency management (as a set of tables and their relationship) by storing in tables and reports behavioral

elements necessary. Operating in this way, software can react to emergency event immediately after it, acting as required and stated by human operator before. Of course, system can be bypassed by authorized personnel at any time, so the operator can act differently if circumstances require. .

The prototype generates already a real-time journal of the event, containing the details of all operations , which guarantees the storage of actions, reactions , and staff who made them , creating a historical archive. This special feature allows you to reconstruct past events in order to learn from their mistakes and reconfigure the management of events, or even to simulate the effects of different behaviors to respond to emergencies . Ongoing research will equip the tactical software module , which will simulate the events , to create a virtual environment with which to assess the effects of actions and emergency communications on the basis of the data acquired through the web and social media during real events. On the basis of criteria to be specified in the research, respectful of ethics and privacy, you can also ensure the modification of the behavior of the software by means of self-learning.

Multi-lingual and Multi-Cultural Aspects of Post-Disaster Emergency Communication – the LinguaNet® Experience

Henrik Selsøe Sørensen

Dept. for International Business Communication (IBC), Copenhagen Business Scool (CBS),
Dalgas Have 15, DK-2000 Frederiksberg
E-mail: hss.ibc@cbs.dk

Abstract

LinguaNet® is a system for fast multi-lingual communications between police forces co-operating across frontiers. It has been in operation for more than two decades and proved its worth. From 1995 to 1998, CBS developed a number of add-ons to the system in the framework of an EU project in order to improve the multi-lingual and multi-cultural efficiency of the system. Three of these, namely multi-lingual casualty registration, cross-cultural ontology work, and a method for handling ontological discrepancies when country-specific data elements clash, are reported in this paper as examples. Given recent technological advances and the upcoming of social media as well as intelligent and instant big-data analyses since the reported research was carried out, it is suggested that the original ideas be revisited and re-engineered in view of improving efficiency in cross-frontier post-disaster emergency situations, where speed, robustness and reliability with respect to very divergent user profiles is a sine qua non.

Keywords: post-disaster emergency communication, multi-lingual and multi-cultural ontologies, LinguaNet®

1. Background

This contribution outlines a small set of the functionalities developed by CBS within the framework of LinguaNet®. LinguaNet is a multinational, multimedium communications system specially designed for cross border, mission critical operational communication by police, fire, ambulance, medical, coastguard, disaster response coordinators. It is tailored to work for specific professional applications over independent networks, LinguaNet employs user-specified message templates and other linguistic controls, graphics and sounds to achieve fast and accurate messaging and information provision

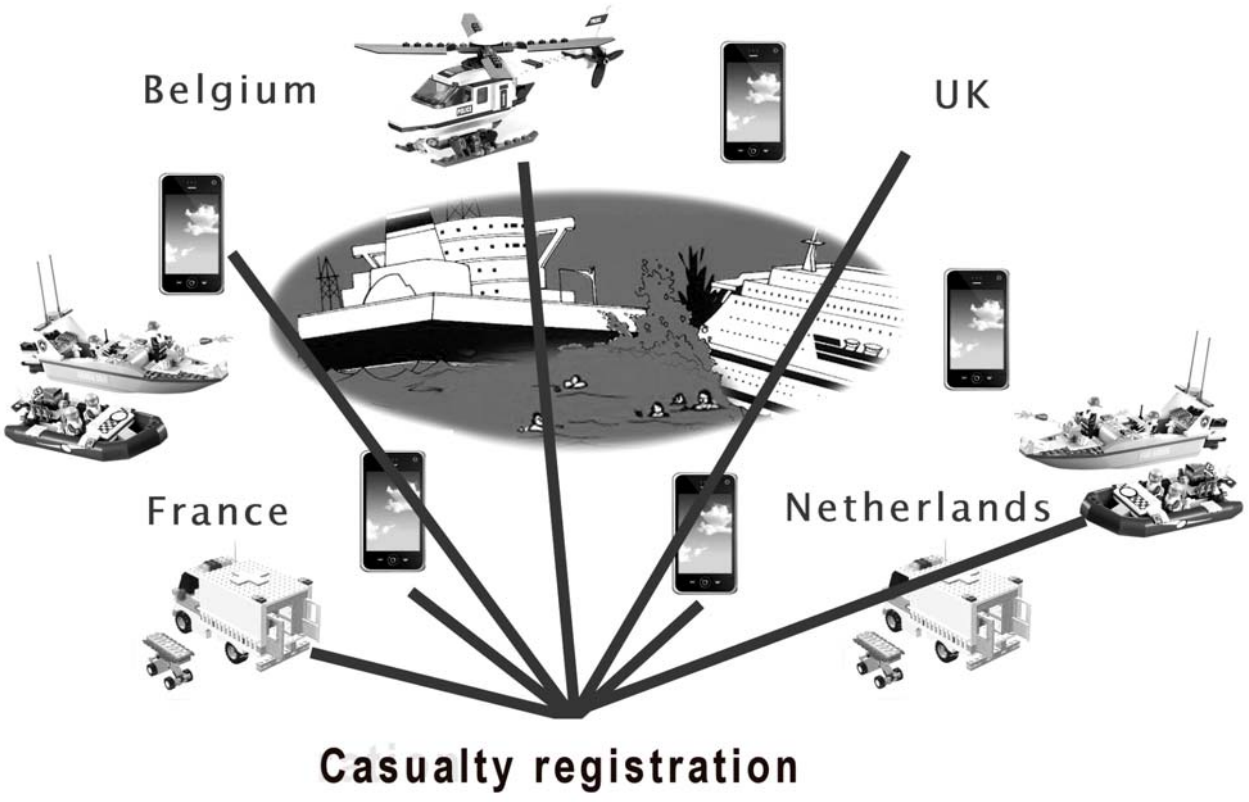


Figure 1: Disaster scenario involving at least 4 languages and countries

across language and administrative barriers. LinguaNet has grown from twenty years of successful research and design of operational language and protocols for sea (Weeks et al., 1984), air (Robertson et al., 1988), police (Johnson et al., 1993c), business and Channel Tunnel emergency service operations (ACPO, 1987, Briggs, 2000, Gallagher, 2000, Johnson et al. 1993a, 1995a). It has applications in many networks in both the private and public sectors and is highly relevant to emergency scenarios. Fifteen forces in five countries worked with the developers to enhance and extend the LinguaNet prototype developed mainly by Edward Johnson and Prolingua Ltd. (Johnson 1993, 1996, Johnson et al 1993a-c, 1995a-c). Pilot implementations of multi-lingual and multi-cultural add-ons developed by the CBS team contain several ideas that are especially practicable in the context of current system developments.

My focus will be the add-ons to LinguaNet developed by myself and my CBS colleagues Inge Gorm Hansen and Bianca Hede in the framework of "Test-bed LinguaNet"; partly funded (1995-98) under the European Commission's 4th Framework Programme - Telematics Applications of Common Interest (Language Engineering). The experimental add-ons described below (Gorm Hansen et al. 1998) are not part of the operational LinguaNet system.

2. Challenges

The challenges that were addressed more specifically at CBS were the multi-lingual and multi-cultural aspects of a cross-border major disaster scenario, see figure 1, where information must be shared instantaneously by different authorities such as rescue centres, police, hospitals, ambulances etc. Using just one system to add and retrieve information safeguards against misunderstandings and communication breakdowns.

Reports are exchanged in several languages and possibly stored in different national systems. Culture-bound discrepancies combined with language barriers constitute serious challenges. Differences in procedures and legal requirements may slow down operations even though the police and rescue forces from different countries have worked together and are trained to avoid operational inadequacies. Instant access to data repositories containing up-to-date contact information and addresses in a neighbouring country as well as to a description of procedures to follow in a foreign country (with translations in several languages) may prove to be crucial.

When it comes to casualty registration, besides the language barriers, a very important function of the casualty forms is to avoid information being recorded more than once, and survivors being interviewed more

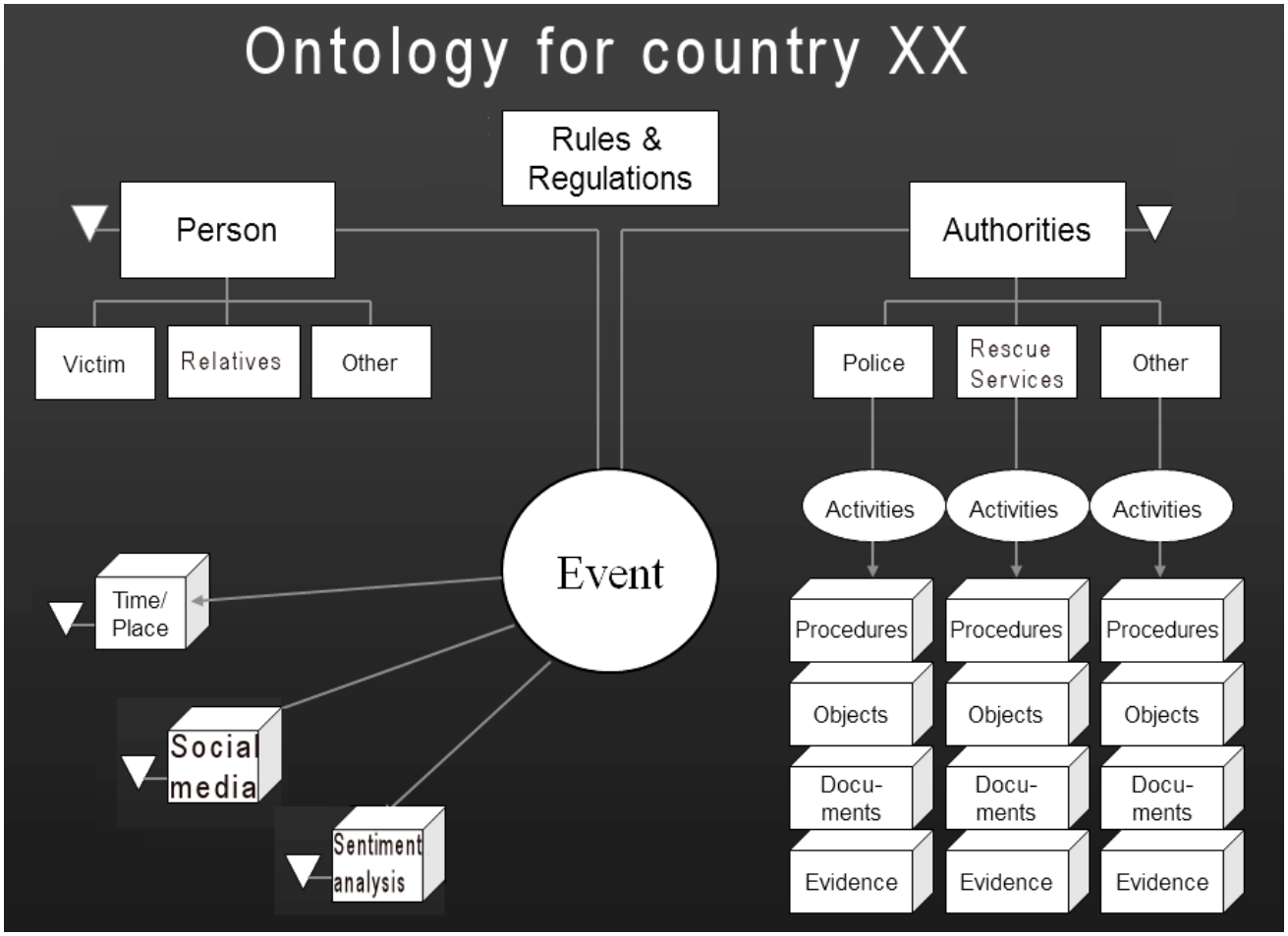


Figure 2: A model for using country- and event-specific ontologies for overcoming administrative discrepancies

than once, by different services. Once the name of a casualty or the telephone number of a relative to be informed has been given, it should be available to all involved parties. The operational response side would benefit clearly from single sourcing across language barriers when information needs to be conveyed to e.g. relatives in search for details about potentially missing persons under circumstances, where errors can be fatal. For a rescue operation to be successful, the communication side must be as efficient as the actual operations.

Audio files, videos and pictures from whatever source, mobile devices on location or social media, should be collected and analysed jointly as they seldom stand alone. For this, annotation is an essential tool in improving communication and allowing fast retrieval.

When it comes to person description and picture annotation, descriptive data elements organised in ontologies and accounting for language- or culture-bound discrepancies may come into play in order to assure fast and reliable retrieval of data in the aftermath of a major incident.

Below, some examples will be given of how challenges of the types mentioned were tackled in the form of experimental add-ons to the LinguaNet system.

3. CBS experimental add-ons

The immediate physical rescue operations require important efforts of communication and coordination and so do the subsequent disaster follow-up and business continuation effort. The CBS LinguaNet team developed various components of which three will be given as examples below. Being research-oriented developments, none of these reached a stage of sufficient robustness during the EU funded project, to be actually integrated into the LinguaNet system, however, it seems to be worthwhile revisiting – and revising - them in the light of possibilities opened by new technologies and social media.

3.1 Alignment of culture-bound ontologies

Stakeholders and procedures, be they administrative or purely technical, differ from country to country and most importantly they are subject to change all the time. An updated map in the form of an ontology of relevant procedures, authorities, communication channels, relevant data repositories (text, audio, video, still pictures) for each country must be at hand, cf. the overview given in Figure 2. Each ontology would hold not only maps of typical procedures and related rules and country-specific caveats but also contact information for involved authorities and services as well as a reasonable level of multi-lingual support in the form of explanatory texts translated into relevant languages.

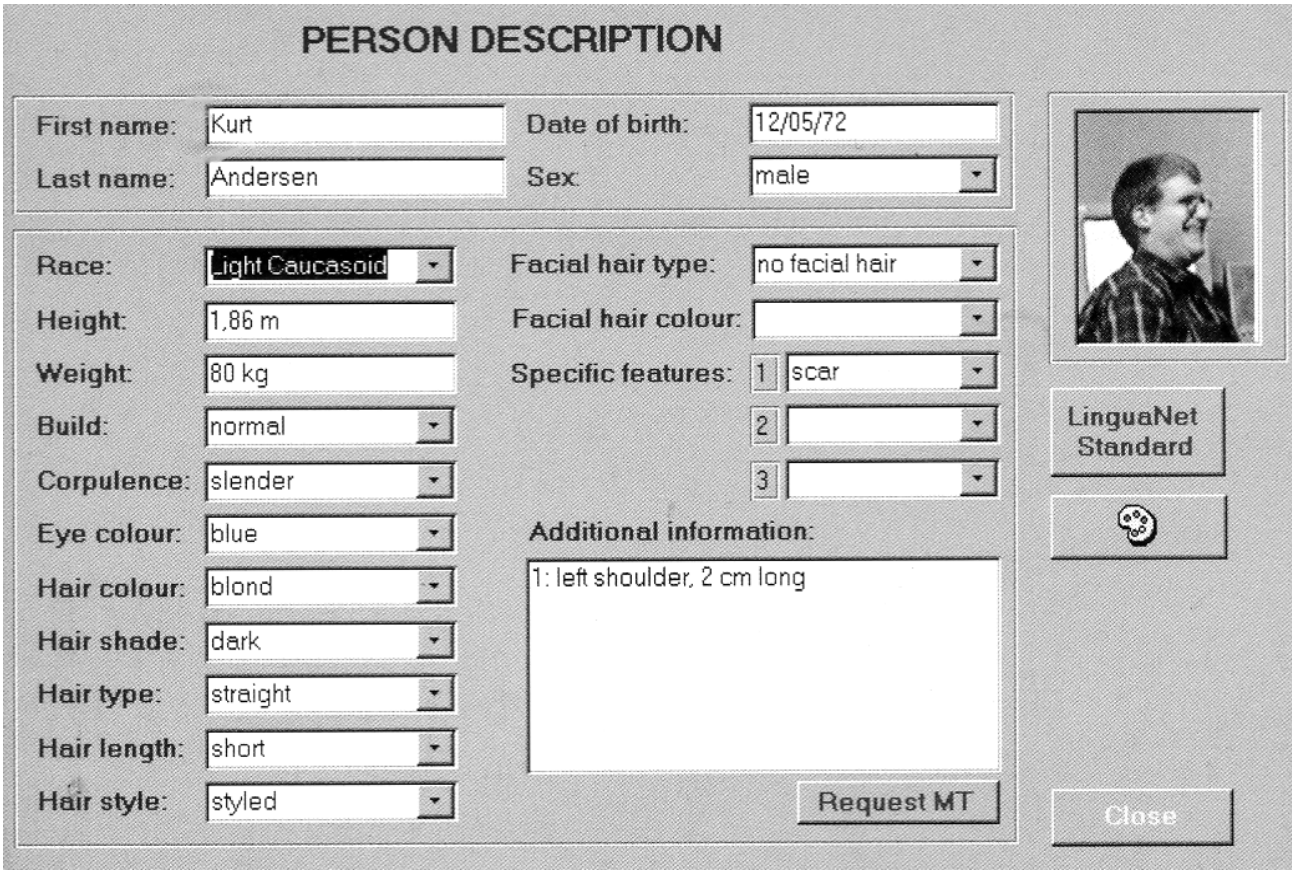


Figure 3: LinguaNet multi-lingual template for person description

Different ontologies for different events must be envisaged. Furthermore, links between the country-specific ontologies and tools to clarify and overcome discrepancies must be made available. Procedures for getting access to sensitive data may differ from country to country. When a closed circuit communication channel at the national level is used for cross-border messages, does it remain closed, and, if not, is there an alternative – this is one of many questions that could be answered by the proposed knowledge base. It would even be useful in cases where the answer is that you are simply not allowed to get access to the wanted data from a foreign country. Even this negative message would be useful and save time.

The mentioning of social media and sentiment analysis in figure 2 signals that the availability of new media and tools should definitely be incorporated in the set-up of a new version of LinguaNet tuned for managing disasters and emergencies. The implementation remains to be done, but it is clear that surveillance based on information extraction from social media in the form of sentiment analysis or other ways of extracting information from "big data" will constitute a valuable resource in disaster management, especially if it not hampered by language barriers.

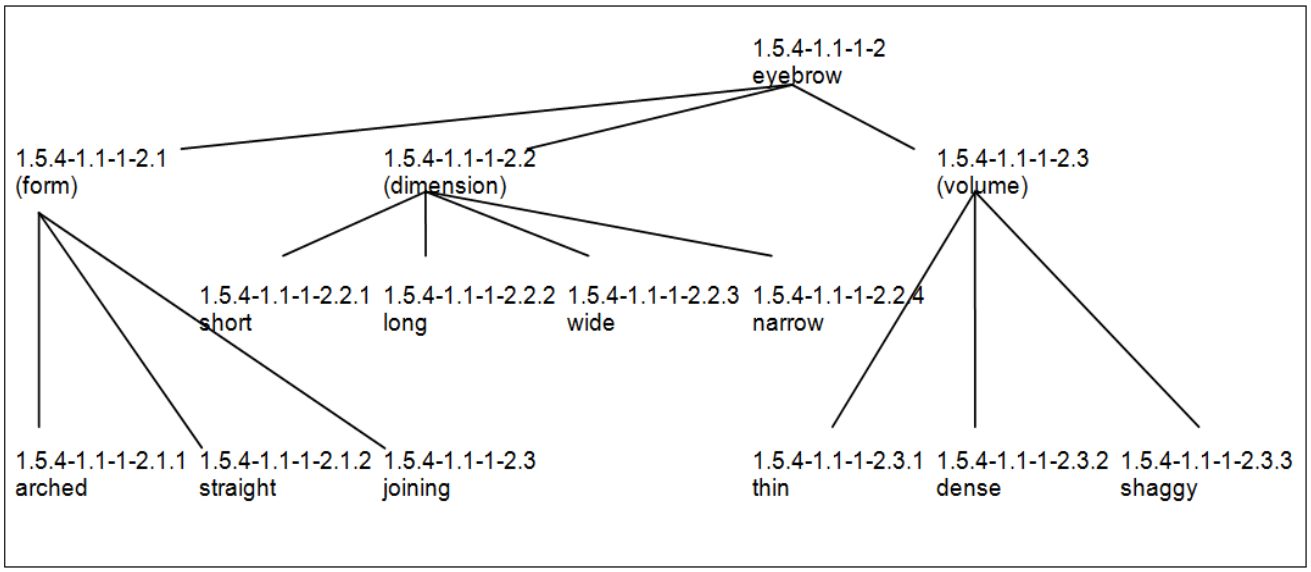
3.2 Overcoming language barriers

For obvious reasons, language barriers are a major hindrance to cross-border operations, because it is crucial that delays are reduced to a minimum both when rescue operations are underway and when victims need to be identified and relatives contacted in the transition phase to post emergency logistics (Stuart-Black, 2008).

As an example of a method to overcome the language barrier, the CBS team of LinguaNet prepared a prototype for a multi-lingual casualty registration system. The prototype was implemented in MicroSoft Access as a flexible modular system for registering and retrieving incident and person data in connection with a major incident. It combines information needed for different purposes, operational, administrative, etc., and comprises a number of individual forms prepared for recording information about a rescued or a missing person. The method could be reused for more advanced registration work today, which would require re-engineering of the database and integration of e.g. videos.

The person description form shown as Figure 3 was the result of extensive terminological work on assembling data elements and building ontologies for person description. The information categories represented had been selected by comparing and contrasting existing country-specific forms in an attempt to satisfy both the need for quick identification where only a few information categories had been filled in. The objective was to create a standard for a detailed person description taking into account multi-lingual and multi-cultural data elements, see example in 3.3 below. The contents of the pick lists reflects the work done on person description terminology and relies on elements drawn from sources made available by authorities in the respective countries. When an international standard was found, e.g. 'race: light caucasoid' etc., this standard was obviously used. In case of discrepancies due to lack of an international standard, a LinguaNet standard was proposed, which was the optimal compromise e.g. for hair styles, cf. discussion below in section 3.3 of what the compromise looked like for the description of eyebrows (based on English, French, Danish standards at the time).

Thanks to the picklists, i.e. templates with controlled content data elements referred to as the "LinguaNet® standard", it is possible to access data in any of the languages involved (English, French, Danish as it were) and still respect the principle of single-sourcing in order to minimise the risk of duplicate recordings.



Figur 4: LinguaNet standard for the description of eyebrows

For the field named “Additional information” (see Figure 3), free text could be entered on an experimental basis, and a machine translation would be proposed as an option with a clear statement of the fact that the translation was indicative, and that a human should be contacted, given that the users are not expected to be linguists (Gorm Hansen et al., 2002).

3.3 Dealing with ontological discrepancies

The methodology used must allow for structured and user-friendly inclusion of new languages, subject fields and data in general. To achieve this, relevant data elements in the involved languages were analyzed in order to identify data elements and organize them in ontologies, particularly in the numerous cases where an international standard had not yet been established. The sources used were made available by the National Commissioner in Copenhagen, the Kent County Constabulary, Interpol (their Disaster Victim Identification Form in Danish, French and English), forms from the Casualty Bureau in England and other European Police and Rescue organizations.

The person description template for the human head and face illustrates the proposed LinguaNet standard which was developed. At one stage, all data elements related to the specific concept ‘eyebrow’ had been identified and organised in systems of concepts, one for each language which were then compared in search for discrepancies, and they were then merged into a common denominator which was subsequently accepted by the involved police forces and rescue services. If one language, e.g. Danish, had only two options for the volume of an eyebrow, namely ’thin’ and ‘dense’ (not ‘shaggy’), then Danish users were offered this third option. The underlying database, furthermore, held a drawing of ‘shaggy’ eyebrows, so that no mistake would be possible.

It was clear that reuse of existing ontologies were definitely purpose-oriented, and that existing ontologies used in other contexts would not be suitable for police users. It is obvious that data elements for ‘eyebrow’ or ‘nose’ for police purposes would differ from what is used in a medical context, e.g. within plastic surgery.

4. Conclusion

Disasters and major incidents tend to have implications beyond the borders of a single country and this fact calls for constant improvements and incorporation of new communication tools. The methods developed as LinguaNet add-ons reported here need to be remodeled and enhanced using up-to-date technologies for sharing and annotating videos etc. and instantaneously extract relevant information from ‘big data’. At the same time, robust multi-lingual aids tailor-made for communication between rescue crews who are not linguists must be improved or created. This report on some selected multi-lingually and multi-culturally oriented methods for how to enhance and smoothen post-disaster emergency and recovery efforts across borders could hopefully serve as inspiration for further developments.

One main conclusion to be passed on from the experience gained from the work with LinguaNet® (with over 50 points of contact involving networks of police and rescue forces in Belgium, France, the Netherlands, Spain, United Kingdom, Denmark Germany) is that robustness is key, because a lot of different services and user profiles who have their full attention fixed on very practical matters are going to create, process, and disseminate information which must be absolutely reliable.

5. Acknowledgements

LinguaNet™ developed by Edward Johnson and Prolingua Ltd.

“Test-bed LinguaNet” Project partly funded under the European Commission’s 4th Framework Programme - Telematics Applications of Common Interest (Language Engineering), 1996-98.

6. References

ACPO (1987). *Emergency Procedures Manual of the ACPO General Purposes Committee* (Standing Sub-Committee on Emergency Procedures). London HMSO.

Briggs, R., (2000). *The European Maritime Disaster Project Final Report.* Essex County Council, Chelmsford, ECC.

Gallagher, D. (1998). *European Police Co-operation: Its development and Impact between 1967-1997 in an Anglo/French trans-frontier setting.* Southampton University PhD thesis.

Gorm Hansen, I., Selsøe Sørensen, H., Johnson, E. (1998).: LinguaNet? We Need it Now:Delivering Multilingual Messaging and Language Resources to the Police. In *Journal of the European Language Resources Association ELRA*. Vol. 3 n. 4, pp. 7--9.

Gorm Hansen, I.; Selsøe Sørensen, H. (2002). LinguaNet : Embedded MT in a Cross-Border Messaging System for European Law Enforcement. In *Machine Translation*, Vol. 17, No. 2, pp. 139--163.

Johnson, E., (1989/1990). *Les langues et la concurrence économique* (1989). Proceedings of the International Language Symposium Paris Volume 4. Also published in English (1990) as: *Language and Economic Competition*, Fachsprache, International Journal of LSP, Vienna 1-2 1990, pp. 2--17.

Johnson, E., (1993). Språkproblem bland poliser under Engelska kanalen. Swedish Centre for Technical Terminology *TNC-Aktuellt* Nr.2., Stockholm, TNC, pp 6--8.

Johnson, E. (1996). Setting a Linguist to Catch a Thief. Published Papers of the Association for Information Management: *Translating and the Computer 18 Conference Proceedings*, London, ASLIB, http://mt-archive.info/Aslib-1996-Johnson.pdf, pp. 1--8.

Johnson, E., Garner, M., Hick, S., Matthews, D., (1993a). *PoliceSpeak – Police Communications and Language and the Channel Tunnel - Research Report* .Cambridge, PoliceSpeak Publications.

Johnson, E., Garner, M., Hick, S., Matthews, D., (1993b). *PoliceSpeak – Police Communications and Language, English-French Police Lexicon.* Cambridge, PoliceSpeak Publications.

Johnson, E., Garner, M., Hick, S., Matthews, D. (1993c). *PoliceSpeak – Police Communications and Language - Speech and Text Recommendations, Cambridge,* PoliceSpeak Publications.

Johnson, E., Garner, M., Hick, S., Matthews, D. (1995a). *INTACOM - Inter-Agency Communications - Report and Recommendations (Volume 1).* Cambridge, PoliceSpeak Publications.

Johnson, E., Garner, M., Hick, S., Matthews, D. (1995b). *INTACOM - Inter-Agency Communications – Emergency Services of Britain and France (Volume 2).* Cambridge, PoliceSpeak Publications.

Johnson, E., Garner, M., Hick, S., Matthews, D. (1995c). *INTACOM - Inter-Agency Communications - English-French and French-English.*

Robertson F.A., Johnson, E. (1988), *AirSpeak: Radiotelephony Communication for Pilots.* Oxford, Prentice Hall.

Stuart-Black S., Stuart-Black, J. and Coles, E. (2008) *Health Emergency Planning: A Handbook for Practitioners*. The Stationery Office, London.

Weeks, Capt. F.,Glover, Lt. A., Johnson, E., Strevens, P. (1984). *Seaspeak Reference Manual.* Oxford, Pergamon Press.

Social media and disaster management: US FEMA as a benchmark for its European counterparts?

Maria Teresa Musacchio
Dipartimento di Studi Linguistici e Letterari, Università di Padova
Via Beato Pellegrino 26, 35137 Padova (Italy)
mt.musacchio@unipd.it

Abstract

Over the last two decades disaster management and its public communication have been substantially transformed by the development of digital media such as blogs, wikis, social media and Youtube. Taken together, these shifts raise a series of issues for work in disaster management. Managing emergencies is a complex undertaking that relies extensively on knowledge management systems. Unlike European counterparts such as the German Bundesdienst für Bevolkerungs-schutz und Katastrophenhilfe (BBK, Federal Office of Civil Protection and Disaster Assistance) and the Italian Protezione Civile (Civil Protection), the US Federal Emergency Management Agency (FEMA) employs social media technologies such as blogs, Facebook and Twitter as relevant disaster management and knowledge sharing mechanisms. This paper investigates how FEMA uses blogs, Facebook and Twitter, what knowledge is shared through these social media, and how knowledge sharing is facilitated and expedited through the use of these systems. Linguistic analysis is conducted to investigate the use of terminology, appraisal, syntactical structures, markers of shared and unshared information, thematic structure and text complexity with a view to identifying what knowledge is transmitted in emergencies, how alerts are provided that do not spread panic or descriptions supplied that form accountable reports of disaster management.

1. Introduction

Communication plays a crucial role in disaster management. Spreading timely and accurate information to all stakeholders – the general public, national and local government authorities and the media – is essential in disaster response and recovery activities and in educating the public in anticipation of severe weather events with a view to reducing the risk of present and future disasters. Over the last two decades communication in disaster management has been transformed by the rise of social media, which enable response organisations to quickly provide information to, interact and maintain contact with the public also before and after emergencies to ensure that people are aware of what to do to prepare and confront severe weather events. Social media are extremely useful in disaster management because people are familiar with them as they use them regularly for other purposes and because more people can be reached through them. However, there is always a risk that information goes viral in a negative sense and spreads panic rather than mitigating the effects of emergencies (Haddow & Haddow 2014: 3). To avoid problems deriving from the dissemination of unsuitable information, the US Federal Emergency Management Agency (FEMA) has developed the following schedule in five points: 1. identify key information that needs to be communicated to the public; 2. craft messages conveying key information that are clear and easily understood by all, including those with special needs; 3. prioritize messages to ensure timely delivery of information without overwhelming the audience; 4. verify accuracy of information through appropriate channels; and 5. disseminate messages using the most effective means available (FEMA 2008 in Haddow and Haddow 2014: 10-11). At issue here is clearly to what extent FEMA successfully applies these principles in practice.

Over the last decade, patterns of communication via social media in emergencies and disaster management have been investigated from a number of perspectives. In public relations, for example, research has focused on emergency knowledge management (Yates and Paquette 2011), on information framing via social media in health communication (Fischer Liu and Kim 2011), on differences in routine as opposed to critical communication (Kavanaugh et al. 2012), on how medium, type of crisis and emotions affect people's reactions (Utz et al. 2013) and how value modelling can be used to assess social media messages in a severe weather event (Freberg et al. 2013). In linguistics, attention has been directly pragmatically at such topics as the production of descriptions that are 'thick' enough – in other words, that convey enough meaning – in initial emergency reports (Cromdal et al. 2008); the formality or informality of safety critical communication (André et al. 2010); or more generally how language is used to create affiliation via social media (Zappavigna 2012). Given the fairly recent nature of these studies, there is scope for further investigation into the language of social media in the highly constrained environment of emergencies, disaster management and preparedness. Constraints are both a consequence of the conciseness which is typical of Facebook posts and especially tweets and interaction which aims to be maximally effective while ambiguity is reduced to a minimum to avoid spreading panic.

The shift to social media or Web 2.0 has turned the web into a resource which can be used not only to read, watch and listen to content created by web authors, but also to write and contribute content in a dynamic rather than static way through the use of computers, tablets, smartphones and the like. Whether they are conveyed via blogs, Facebook posts or tweets, social media messages are presented in reverse chronological order, can incorporate links to other sites like Youtube, and include metadata such as the time when they were created and the ID of the user. Facebook and Twitter also provide information about followers. Research into the use of

social networking services suggests that posts do not only contain information, but they also provide recommendations and opinions. Therefore, investigations into what knowledge is shared and how, and also how contact is established and maintained in social media is highly relevant.

This paper compares FEMA blogs, Facebook posts and tweets in English, and contrasts tweets in English and Spanish to identify what knowledge is transmitted, how alerts are provided that do not spread panic during emergencies or how descriptions are supplied that form accountable reports of disaster management and/or prepare and educate people to face future disasters. FEMA blogging is analysed with a view to establishing whether it can be regarded by its European counterparts as a benchmark against which to develop similar web content.

2. Aims and Methodology

Social media in emergencies and disaster management are investigated to identify patterns of meaning which can throw light on what knowledge is shared and how knowledge is facilitated and expedited through the use of these systems. The method chosen to study meaning combines a qualitative analysis of corpora with a qualitative approach to discourse analysis which aims to trace the underlying social semiotics of sharing information and maintaining contact with the public.

Interactions through social media are usually considered a kind of conversation where the relationship construed between participants is typically based on 'information snacking', that is, restricted to making initial contact and then resuming it occasionally at important dates (Zappavigna 2012: 6) which – in the case of disaster management – would be emergencies and the necessary training and updating to be provided between one emergency and the next. Negotiating and maintaining relationships in disaster management thus implies using social media to find people and bond with them around shared values such as mitigation, preparedness, response and recovery. Mitigation consists in implementing strategies, technologies and actions that reduce the number of casualties and loss of property during disasters. Preparedness is achieved through messages that educate and train people in anticipation of disasters. Response means warning, informing people of a disaster and of evacuation operations. Recovery gives individuals and communities affected by a disaster information on how to apply for disaster relief (Haddow & Haddow 2014: 4). Research indicates that microblogs such as Twitter are searched more for social content and events, while web searches yield more information on facts and navigation (Teevan et al. 2011). As users access social media to find information that interests them and drop in and out of the discussion over time, organizations such as FEMA need to leverage this kind of sporadic access by establishing a set of practices in which their online contacts become 'followers', i.e. they are constructed as an audience and regular access is ensured through audience management and carefully presented content. A question then arises as to how language is used to create and maintain these social bonds as it amounts to construing ideational meaning as describing human experience and coupling it with interpersonal meaning as enacting personal and social relationships between people (Halliday 2004: 29). This coupling can then be investigated to define online communities through their discursive practice (what they actually do with language), their metadiscursive practice (what they write about the language they use) and their implicit metapragmatics or what their language covertly signals (Johnson & Ensslin 2007: 10). This is based on a study of the essential features of blogging and microblogging such as the use of terminology, appraisal, syntactical structures, markers of shared and unshared information, thematic structure and text complexity (Zappavigna 2012).

As corpora to be investigated, social media include a range of components of different kinds – typically text, image and video – which cannot be studied just by using corpus analysis tools and pose problems of representativeness, balance and comparability, especially as they are time-sensitive texts. To what extent do the samples reflect the patterns to be investigated? Does the corpus contain equal ranges of the text types in question? What is the corpus potential in terms of the possibility to compare it with similar corpora?

First, corpus analysis here focuses on (written) text and is thus eminently linguistic. Second, in order to identify what information is transmitted about disasters, wordlists of Facebook posts, tweets and blogs were used to obtain a set of keywords, to compare with each other and contrast with those of Spanish tweets. For both English and Spanish, keywords were extracted using news articles as a larger, reference sub-corpus of general language texts on the same topic(s). Third, once keywords had been extracted, KWIC concordances were run in the three English subcorpora and in Spanish to investigate use in greater detail. Fourth, attitudinal and affective words as expressions of appraisal were searched amongst keywords to establish any differences in use in the four text types present in the corpus. Here the reference framework is Martin and White's study of appraisal (2005), though for the purposes of this preliminary study only the three broad categories of appraisal will be considered, namely affect which expresses emotion, judgement which assesses behaviour and appreciation which estimates value. Finally, quantitative measures of text complexity such as type/token ratios and sentence length were supplemented with qualitative analyses of syntactic structures, markers of shared and unshared information and thematic structures to outline trends in the language of Facebook posts, tweets and blogs for communication in emergencies. Corpus analysis was conducted using version 3.2.4 of Laurence Anthony's AntConc software and version 6 of Mike Scott's WordSmith Tools software.

3. Data Collection

Data were collected from the FEMA website and through Lexis Nexis as a computer-assisted research service to retrieve legal documents and news articles from the web. The corpus includes three English sub-corpora of texts from social media: a) a collection of Facebook posts; b) all FEMA tweets till mid-March 2014; c) the ten most

recent pages of FEMA blogs. For the purposes of comparison, the corpus includes all FEMA tweets available in Spanish and news articles in English, Spanish and Italian retrieved from Lexis Nexis using three keywords – weather, emergency, disaster – and their equivalents in Spanish and Italian.

Text type	Source	Language	Tokens
Facebook	FEMA	English	38,448
Twitter	FEMA	English	60,531
Twitter	FEMA	Spanish	19,913
Blogs	FEMA	English	409,514
News articles	Lexis Nexis	English	477,748
News articles	Lexis Nexis	Spanish	398,183
News articles	Lexis Nexis	Italian	177,054
TOTAL			**1,581,391**

Table 1: Components of the disaster management corpus including three sub-corpora of social media and reference corpora of news articles in three languages.

As can be seen from Table 1, FEMA blogs, Facebook posts and tweets involve issues of representativeness as they constitute samples which may or may not reflect the typical patterns of the corresponding text types. They also pose problems as to balance, because the size of FEMA blogs, posts and tweets is different. Finally, comparability is somewhat limited as the English and Spanish tweets are different in size. The corpus, which currently has 1,581,391 tokens distributed amongst three languages, is somewhat representative, but is not balanced and only allows comparison to a given extent. This is not due to corpus design, but rather to the fact that social media in disaster management is still something of a novelty and that FEMA mainly publishes texts in English and has no Facebook posts or blogs in Spanish; even its tweets seemed to be translated into Spanish at the beginning. Moreover, the number of Italian news articles that can be retrieved through Lexis Nexis on the topic of emergencies and disaster management is much lower than that of Spanish and cannot compare with English. In terms of size, the English subcorpus – with a total of 986,241 tokens – is twice as large as the Spanish one (418,096 tokens), which in turn is two and a half times bigger than the Italian one. Yet, considering the preliminary nature of this work, the data available from FEMA anyway, and the possibility to compare Spanish and Italian as close Romance languages, it was felt that investigating the corpus from a qualitative and – to a lesser extent – from a quantitative point of view could still provide some insights into the use of social media in disaster management to be subsequently explored, confirmed and validated through analysis of larger, more balanced corpora. To mention but one of the problems encountered in compiling the corpus, Facebook posts in English from the beginning in mid-2009 to 2014 had to be manually downloaded to ensure that comments from the public in nested form could be identified and retrieved, though this was only possible for the most recent posts as lots of comments had been withdrawn or deleted as no longer relevant. No doubt, a routine can be developed for 'scraping' social media texts to ensure that all Facebook posts and tweets are automatically retrieved once a research project is clearly defined. Still, at the moment the corpus is representative in parts, as it does include all FEMA tweets in English and Spanish.

4. Analysis

FEMA's mission is "to support our [US] citizens and first responders to ensure that as a nation we work together to build, sustain, and improve our capability to prepare for, protect against, respond to, recover from, and mitigate all hazards" (www.fema.gov). To accomplish this mission FEMA has a "Blog, newsroom, videos & photos" section to provide information and communicate with all stakeholders. Within this section a "Social media" subsection can be found which includes links to FEMA on Twitter, FEMA on Facebook, and FEMA on YouTube. A comparison of FEMA on Twitter and FEMA on Facebook with FEMA blogs suggests that tweets and Facebook posts or microblogs are increasingly concise versions of FEMA blogs and are indeed conceived as short notes that refer back to blogs for more extended information.

Calculation of keyness yielded keywords for the three FEMA subcorpora in English. Keyword lists were compared to find common items and study those in greater detail. To decide how typical of each individual component of the corpus these items were, the "weirdness" coefficient was calculated (Ahmad 2007). This coefficient can be used to identify the items that occur comparatively more frequently in a given specialist corpus than in a reference general-language corpus (Lexis Nexis, in our case). The coefficient is calculated as a ratio: R_s/R_g, where R_s is the relative frequency of an item in a corpus of N_s words and R_g is the relative frequency of the same item in a general language corpus of N_g words. The higher the weirdness value of an item, the more typical it is of the corpus it appears in. Results in Table 2 below suggest prominence given to two items in all three subcorpora, *tip** and *safe**.

Key-word or root	Twitter	Facebook	Blogs
ready	**11.2**	5.8	0.5
tip*	**34.3**	**45.4**	**2.6**
safe*	**8.2**	**6.1**	**1.6**
recover*	3.1	2.9	0.3
watch*	7.9	3.8	**1.4**
surviv*	6.1	**7.3**	1.1
prepar*	2.1	3.5	0.9
severe	1.7	1.7	0.5

Table 2: Weirdness of key-words or roots common to the three FEMA subcorpora in English.

The keywords or key roots selected to investigate weirdness in Table 2 are interrelated and highly productive in the FEMA sucorpora in English both with reference to domain-specific terms and text composition. *Ready* is a US government website (www.ready.gov) providing information on how to prepare for an

emergency and as such becomes a topic – hashtag – in Twitter (*#ready*), but also provides advice such as *Be ready!* or *Get ready!* where the use of exclamation marks assists in intensifying the force of the message. *Tip** mainly appears in *safety tip(s)* or *preparedeness tip(s)* and is also a component of hashtags in Twitter such as *#winter safety tip*. *Safe** is used both as an adjective in recommendations like *Stay safe!* and as a term entering compounds or phrases like *winter safety* or *disaster-specific safety tips*. *Recover** is generally found in its noun form, *recovery*. It is one of the four phases of emergency management identified by FEMA, namely mitigation, preparedness, response and recovery (Haddow & Haddow 2014: 3-4). As such, it frequently appears in the binomial *response and recovery* which also relies on consonance to boost its effect on Facebook:

15 February 2013
New Jersey Recovery: 100 Days After Sandy
It's been over 100 days since Hurricane Sandy struck the East Coast, but response and recovery efforts are still going at full speed. Here's a look at some of the key points to recovery in New Jersey, where over 58,000 applicants have been approved for federal disaster assistance, resulting in over $358 million going to impacted individuals and families thus far. There is still much work to do. For more on the recovery efforts in New Jersey, visit www.fema.gov/SandyNJ

*Watch** is found as a verb in the name of a blog, *What we are watching*, and as a term as in the following example from Twitter:

"Winter Storm Watch" means severe #wx such as heavy snow or ice is possible, follow local news reports & be alert to changing wx conditions

where *wx* is the Twitter abbreviation for weather. *Watch* as a term also appears in the phrase *watches or warnings* and FEMA makes a distinction between the two, as shown in the Facebook post below:

Jeff Hillendahl The two tier weather alerts provide citizens the ability to prepare. The "watch" is telling you that you need to get your stuff together, but don't freak out just yet. The "warning" is telling you to take the immediate appropriate measures for the severe weather that the "watch", hopefully, alerted you to before.

*Surviv** is mainly found in the form *survivor(s)* in collocations such as *disaster/flood/tornado survivors* or in the hashtag *#Sandy survivors*. Similarly to *ready*, *prepar** appears in recommendations such as *Get prepared!* Like *watch** it also describes a phase in emergency management and is found in compounds such as *national/domestic/community/family/emergency/disaster preparedness*. As a measure of the keyness of *preparedness* in this domain, one should consider that in the 450m-word Corpus of Contemporary American English it has a frequency of 0.000003%. Finally, *severe* is most frequently used in the *severe weather* collocation, though in emergencies it is also found in *severe damage* and more specific types of bad weather – *severe drought*, *severe cold*. Though with reference to weather it is unmistakably negative in connotation, it occupies an intermediate position in the range exhibited in news articles, which goes from *poor*, *adverse*, *nasty* through *inclement*, *rough* and *severe* to *extreme* and *wild*. The following tweet shows use of some of these keywords in combination:

1/29 @usNWSgov forecasts call for chance of severe storms in central US. Safety tips & watch/warning definitions http://www.fema.gov/blog/2013-01-29/severe-weather-throughout-south-midwest...

Apart from the metalanguage – *expand*, *reply*, *retweet(ed)*, *favorite*, *watch* etc. – for the exchange of messages, which in FEMA tweets is English even in the Spanish version, keywords in Spanish tweets largely rely on the common origin of words in the two languages for FEMA 'household' terms such as *recuperación* (recovery) and *preparación* (preparedness) and general concepts like *emergencia(s)*, *desastre* and *rumor*. Both in English and in Spanish, FEMA expresses a preoccupation to keep rumours under control for fear that inaccurate information may spread panic during emergencies. Compare two similar tweets in English and Spanish:

?@fema 9 Nov 2012
Share this updated #Sandy Rumor Control page to help us provide accurate information: http://www.fema.gov/hurricane-sandy-rumor-control … Tell a friend.
?@FEMAespanol 15 Nov 2012
Los rumores se riegan rápido. Dile a un amigo, comparte esta página y ayúdanos a proveer información correcta: http://www.fema.gov/es/huracan-sandy-controlando-rumores … #Sandy

Interestingly, the order of components in the two messages is different: in English sharing information is thematised and telling friends comes second, whereas Spanish foregrounds sharing information with a friend. This suggests that cultural differences in presenting slogan-like messages need to be investigated extensively. Finally, FEMA's choice to describe adverse weather conditions as 'severe weather' affects Spanish too, where the loan translation *clima severo* is used. This collocation is not present in the Lexis Nexis Spanish reference corpus and is only found in internet resources such as Webcorp with reference to the US, Mexico and central American countries. Indeed, Spanish *clima* is a closer equivalent to English *climate*, and bad weather is usually described as *mal tiempo*. A comparison with Italian news articles suggests that in-depth analysis of Spanish tweets can help to develop best practices in the use of social media for disaster management by drawing on the similarities of Italian and Spanish as Romance languages.

With reference to appraisal in the context of disaster management, the example of severe weather above, indicates that FEMA chooses its words very carefully, as a governmental agency is supposed to do in order to keep the middle ground and avoid expressions of emotion. The case of rumour, suggests however that, in the case of potentially dangerous information, FEMA is called upon to play its role as guardian of smooth, safe operations in emergencies. Forced to take a stance on rumours by expressing judgement in the form of polarity – either positive or negative – FEMA chooses the former and lays

the stress on accurate information and involvement (*share…help us/comparte…y ayúdanos*) to build trust and favour affiliation. Yet, social media imply dialogue and exchanges in Facebook and Twitter can lead the public to express criticism as in the following post:

5 March 2014
Nicholas Limon Shame on you FEMA for denying aid to Washington, IL!! If we can send $1 Million to Ukraine, we obliviously have the money to help those, who are still recovering. You people are a disgrace to this country!! HELP US OUT!!

However, once contact has been made, it needs to be kept. When there are no emergencies, it is a good time for FEMA to train people and thus create a bond with blog, Facebook or Twitter followers. In this context, the keyword is *learn*. As can be seen from Table 3 below, KWIC concordances indicate that FEMA messages revolve round the slogan-like collocations *learn about, learn from, learn how, learn more* and *learn what*:

Example	Source
Our best forecasts and warnings mean nothing if YOU don't do something with this information. So, please join us. Take this week to **learn about** the threats.	Blog
FEMA Sandy ?@FEMASandy 29 Nov 2012 It's been said "**Learn from** the past, prepare for the future," which is exactly what you can do when rebuilding http://www.fema.gov/protecting-homes … #Sandy	Twitter
19 August 2009 FEMAinFocus Tip: **Learn how** to help your neighbors during a disaster. Join a local Community Emergency Response Team in your state http://tinyurl.com/pa8tbg	Facebook
11 October 2011 If an earthquake happens, drop, take cover under a sturdy object and hold on. To **learn more** about what to do before, during and after a quake, visit http://Ready.gov/earthquakes	Facebook
PrepareAthon ?@PrepareAthon Oct 17 Doing #ShakeOut is a cheap, simple thing to **learn what** to do during a quake. Participate like this at 10:17 am http://ow.ly/i/3rZr4	Twitter

Table 3: KWIC concordance of *learn* in FEMA blogs, Facebook posts and tweets.

Examples suggest that FEMA aims to establish an ongoing dialogue with blog, Facebook and Twitter followers where the use of exclusive *we*, which means 'we the experts' and creates some distance between writer and reader, is counterbalanced by the invitation to 'join' and 'share' and the stress on 'you'. Phatic appeal is also evident in alternative formulations such as *learn how to/learn how you…*:

USA.gov ?@USAgov Apr 22
Learning how to respond to natural disasters is a great way to give back to your community. **Learn how you** can help at http://www.ready.gov/volunteer

Syntactic features such as sentence structure and complexity can be investigated indirectly through statistical measures of the subcorpora. As some corpus components are rather small, statistical reliability can be questioned. For this reason, they are introduced here more for their relevance in developing a method to extract information than for the actual data they yield. Type/token ratio (TTR), standardized type/token ratio (STTR), mean sentence length and standard deviation of sentence length were obtained as part of the statistics in Wordsmith Tools wordlist. Statistics are shown in Table 4.

Statistics	Twitter		Facebook	Blogs
	English	Spanish	English	English
Type/token ratio (TTR)	7.46	11.66	14.17	8.65
Standardised TTR	29.29	25.85	42.12	42.22
Mean sentence length	14.68	12.29	21.18	20.63
Standard deviation of sentence length	10.25	11.05	19.94	13.98

Table 4: Statistics of lexical density and text complexity for FEMA blogs, Facebook posts and tweets.

Though TTR and STTR are quite crude as measures, it is assumed that they provide some information about lexical density. High TTR or STTR indicate a high variety of lexis. In this respect, Twitter and Facebook messages in English are not as varied as blogs, though Facebook posts are only minimally different. The size of the Facebook subcorpus may have an impact on the reliability of results in the case of sentence length (SL) and standard deviation of sentence length (SdSL) too, as figures are higher for Facebook posts than for blogs, which are traditionally regarded as longer texts. By contrast, all measures relating to news articles seem to confirm results from other corpora of the same text type. Spanish and Italian are more inflected, so variation is smaller and sentences are shorter in English, as outlined in Table 5.

Statistics	English	Spanish	Italian
Type/token ratio (TTR)	4.42	6.07	10.49
Standardised TTR	44.89	42.20	50.02
Mean sentence length	21.06	27.43	23.51
Standard deviation of sentence length	18.48	30.15	20.43

Table 5: Table 4: Statistics of lexical density and text complexity for news articles in English, Spanish and Italian retrieved using Lexis Nexis.

5. Discussion

Preliminary analysis of data from the FEMA corpus of social media texts suggests that FEMA adapts its use of language according to text type. Different measures of keyness and weirdness of terms point to FEMA's need to choose different modes of communication depending on the type and length of message. Blogs are longer and differ minimally from newspaper articles while Facebook posts are more concise and rely on addressing the public directly by providing recommendations, advice or instruction but also support through reference to success

stories in disaster management. Contact is frequently established by commands expressed through imperatives – *Get ready!, Be prepared!* – and more detailed information usually takes the form of links to blog pages or other websites. Twitter posts are sparer still and often limited to bare imperatives and topics in the short, easily recognizable form of hashtags. Similar strategies are found in FEMA's Spanish tweets.

Corpus analysis shows that FEMA has a fully developed and consolidated terminology which is adapted to suit the required length of the message, so that 'severe weather' is shortened to *#severewx* in Twitter. Once contact with the public has been established, FEMA tries to create bonding by stressing the need to keep updated but also learn about emergencies and train for disaster management in advance. This educational function is fulfilled by devising slogan-like messages which refer to campaigns tailored to the needs of special groups such as children or people with disabilities. In this context the language of appraisal is common as the importance of accurate information is stressed and positive, reassuring messages are conveyed to avoid causing stress or spreading panic. Use of social media always entails a degree of risk, as communication is a form of dialogue and public response can only be controlled up to a point.

Syntactic structure and text complexity have been quantitatively investigated through the information that can be gleaned from corpus statistics and supplemented with qualitative studies. Though preliminary results confirmed expectations, the size of some subcorpora challenges statistical reliability. Larger corpora including feedback by the public are needed for more robust statistics. Great help could be provided by comparisons with texts by similar government agencies in other countries, such as Canada and New Zealand, where Twitter messages are used in disaster management.

6. Conclusion

In this paper an attempt has been made to develop a framework for the study of the features of language use in social media for disaster management. The US Federal Emergency Management Agency was taken as a test case to investigate what social media technologies are used and how, what knowledge is shared and what strategies make social media messages more effective. FEMA does not only establish contact with the public by providing accountable reports, information and support at the time of emergencies, but also tries to ensure bonding in order to get people involved in education and training in a kind of lifelong learning before, during and after disasters.

The main purpose of analysis was to establish if FEMA practices as to social media can be regarded as a benchmark against which European government agencies can develop similar web content. Linguistic investigation in this paper suggests that FEMA has worked out a number of strategies to ensure that its social media messages are effective and do not spread panic. However, as social media content is multimodal, linguistic analysis needs to rely on larger corpora and to be supplemented with studies of images, videos and audio messages, and of their interaction, to confirm or disprove results from this preliminary study. At the moment, FEMA is the agency that makes the more extensive use of social media for disaster management in the English speaking world and has been doing so for longer. There are clearly lessons to be learned from their experience.

7. References

Ahmad, K. (2007). Artificial ontologies and real thoughts: Populating the semantic web? (pp. 3-23). In Basili R. & Pazienza M.T. (eds) *AI*IA 2007, LNAI 4733*. Berlin/ Heidelberg: Springer Verlag.

André, M. Sanne, J.M. & Linell, P. (2010). Striking the balance between formality and informality in safety-critical communication: Train traffic control calls (pp. 220-241). In: *Journal of Pragmatics*, 42.

Bednarek M. & Caple H. (2012). 'Value added': Language, image and news values (pp. 103-113). In: *Discourse, Context and Media*, 1.

COCA (1990-2012). Corpus of Contemporary American English, Brigham Young University, http://corpus.byu.edu/coca/ (last accessed in mid-March 2014).

Cromdal J., Osvaldsson, K. & Persson-Thunqvist D. (2008). Context that matters: Producing "thick-enough descriptions" in initial emergency reports (pp. 927-959). In: *Journal of Pragmatics*, 40.

Fischer Liu, B. & Kim S. (2011). How organizations framed the 2009 H1N1 pandemic via social and traditional media: Implications for US health communicators (pp. 233-244). In: *Public Relations Review*, 37 (3) September.

Freberg, K., Saling, K., Vidoloff, K.G. & Eosco, G. (2013). Using value modelling to evaluate social media messages: The case of Hurricane Irene (pp. 185-192). In: *Public Relations Review*, 39 (3), September.

Haddow, G.D. & Haddow, K.S. (2014). *Disaster Communications in a Changing Media World*. Amsterdam/Boston: Butterworth-Heinemann.

Halliday, M.A.K. (2004). *An Introduction to Functional Grammar*. London: Arnold.

Johnson S., Ensslin, A. (2007) Language in the media: theory and practice (pp.3-22). In: Johnson S. & Ensslin A. (eds) *Language in the Media*. London: Continuum.

Kavanaugh, A.L., Fox, E.A., Sheetz, S.D., Yang, S., Li, L.T., Shoemaker, D.J., Natsev, A. & Xie, L. (2012). Social media use by government: From routine to the critical (pp. 480-491). In: *Government Information Quaterly*, 29 (4), October.

Martin, J.R. & White, P.R.R. (2005). *The Language of Evaluation: Appraisal in English*. New York: Palgrave Macmillan.

Teevan, J., Rmage, D. & Ringel Morris, M. (2011). \#TwitterSearch: A comparison of microblog search and Web search. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. Hong Kong: ACM.

Utz S., Schultz F. & Glocka S. (2013). Crisis communication online: How medium, crisis type and emotions affected public reactions in the Fukushima Daiichi nuclear disaster (pp. 40-46). In: *Public Relations Review*, 39 (1) March.

Webcorp (2014) *Webcorp. The Web as Corpus*. Birmingham City University, http://www.webcorp.org.uk/live/ (last accessed in mid-March 2014).

Yates, D. & Paquette S. (2011). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake (pp. 6-13). In: *Journal of Information Management*, 31 (1) February.

Zappavigna, M. (2012). *Discourse of Twitter and Social Media. How We Use Language to Create Affiliation on the Web*. London/New York: Continuum.

How Communications Companies Can Help Organisations Prepare for Disasters

Cilian Fennell
Stillwater Communications,
64 Dame Street,
Dublin 2,
Ireland.
E-mail: cilian@stillwater.ie

Abstract

Effective communication can play a crucial role in predicting, preventing and managing a disaster. A good communications company will help its clients prepare for crisis situations by building relationships with stakeholders, creating effective communications structures and developing a crisis communications strategy. For the purpose of this presentation, we will be concentrating on the case study of a flood management crisis in Ireland, a client project which we worked on recently.

Keywords: communication, media relations

Effective communication can play a crucial role in predicting, preventing and managing a disaster. It also helps to speed the return to normality after a disaster has occurred.

Increasingly, communication companies are managing crisis communications strategies of both public and private sector enterprises. Some of these enterprises deal with life, business and safety-critical operations and infrastructure. Communications companies are expected to ensure clear and effective lines of communications across different sections of a society and across different cultures.

The overall role of a communications company is to help its clients communicate with the external world in a truthful, authentic and engaging way. It also helps organisations refine their messages to ensure clarity, consistency and effectiveness of communication.

There are multiple platforms available for this including face-to-face, online, through the media or through the creation of events that convey the messages and ethos of the company to its various audiences. Communicating successfully requires a deep understanding of all aspects of the client's work, in particular any possible threats to its stakeholders, be they customers, clients, employees or the general public.

Disasters can come in many shapes and sizes, from an oil spill in the Gulf, to political upheaval, to a toxic release. Other disasters for a company can include the release of faulty products, a reputational scandal or the misbehavior of employees. Each one of these has the ability to harm its customer base, particularly when the product or service is in the area of health or safety. Examples of this could be in pharmaceutical, medical, childcare or security sectors. Ensuring clear communication during this phase can ameliorate the effect of the crisis on customers and the organisation alike. By providing early warning, clear instructions during the crisis and open and transparent reviews afterwards, communications companies can lessen the actual and reputational harm to the company and its clients.

The communications element of disaster management must begin before there is any threat of disaster. It should be an ongoing priority of any organisation to have good relationships with those entities that inform the public about their operations and their industries. Media relations are particularly important as this is the most powerful way to disseminate information and influence mass audiences. Creating an understanding of your organisation and its operations with those journalists reporting on your sector is vital so that when disasters occur, they are reported in context and by someone who understands your business.

A good communications company will help its clients prepare for crisis situations by building relationships with stakeholders, creating effective communications structures and developing a crisis communications strategy.

For the purpose of this presentation, we will be concentrating on the case study of a flood management crisis in Ireland. This was a client project which we worked on recently.

This flood management plan had the following five elements:

1. Creating good relationships with the media and other important influencers. This serves to improve communication during any crisis or disaster. We give journalists, bloggers and other online influencers a complete understanding of the client's operation, threats, risks and crisis avoidance strategy, thereby easing the flow of communication when a crisis occurs.

2. Training spokespeople and key internal staff in delivering clear and concise communication through the creation of crisis scenarios and role playing. This can include a range of exercises from a full blown disaster management simulation, to on camera media training, to the creation of sound bites and updating messages.

3. Analysing any previous disasters for weaknesses and suggesting improvements to the communications process or the retraining of personnel. Reviewing all media engagements before, during and after the crisis. Tracking reputational damage and customer satisfaction reports.

4. Recruiting allies both internally and externally to warn of any potential danger signs. Identifying key influencers and creating communication champions across all levels of the organization so that traditional and non-traditional channels can be used to disseminate information and influence behavior as effectively as possible.

5. Drafting a crisis communications plan to set out a clear response structure.

These five key steps ensure that the client is prepared when disaster occurs.

Italian doctor-patient interactions: A study of verbal and non-verbal behavior leading to miscommunication

M. Grazia Busà, Sara Brugnerotto
University of Padova
Italy
E-mail: mariagrazia.busa@unipd.it, sara.brugnerotto@studenti.unipd.it

Abstract

This study discusses aspects of doctor-patient communication and presents a preliminary analysis of doctor-patient interactions in Italy. The aim is to gain information on how (mis)communication between doctors and patients may affect the doctor-patient relationship and may lead patients to lack trust in their doctors. The authors use a corpus of existing audio-video materials on Italian doctor-patient interactions, and analyse doctors' use of verbal and nonverbal expressions in their exchanges with their patients. The analysis is aimed to identify which features may engender communication problems –leading to misunderstandings and the perception of doctors as distant or unreliable. The preliminary findings reveal that patients' lack of trust in doctors may also be the result of doctors' use of culture-specific patterns of verbal and nonverbal expressions, for example certain sentences used for minimizing patients' fears, specific postures and gestures signalling distance or closure. These findings will be used for planning future investigations of doctor-patient interactions based on the collection and analysis of audio-visual material. Having a detailed knowledge of what patterns mostly affect communication in natural settings will provide important information to be implemented in digital devices.

Keywords: doctor-patient communication; verbal and nonverbal communication

1. Introduction

In the past patients had high levels of trust in health care professionals. Interpersonal doctor-patient relations were characterized by a sort of blind reliance in doctors; this developed as a result of a longstanding relationship between the patients and their personal physician as well as the patients' recognition of the physician's knowledge and medical expertise.

This relationship has been transformed by changes in the culture of health care. Public attitudes towards professionals and their authority as medical experts are changing and there is a decreasing deference to authority and trust in doctors and institutions. The lack of trust in doctors may be the result of various factors such as the increasing competence and confidence of the patients in their own personal judgment of risk (Beck, 1992; Hall, Roter and Rand, 1981), a wider level of education, and a growing level of available information (i.e. through media such as television or the internet). These factors allow patients to doubt what their personal physician says and, consequently, demand more. The change in the institution of medicine may be another cause for the decline in patients' trust in doctors: technological progress and the continuous search for a higher efficiency and rapidity of treatments lead patients to expect more of the health care providers and justify less any possible mistake. This is one of the leading causes for the rise of the so-called 'defensive medicine', referring to the doctor's practice of recommending treatments that are not necessarily the best option for the patient, but that mainly serve the function to protect the physician against the patient as potential plaintiff.

One of the effects of the complex changes in the culture of health care is that, as compared to the past, doctors generally adopt a more detached approach to their profession. That means, they bring an increased distance in the personal relationship with the patients and therefore a decreased sensibility and empathy for their patients during their encounters. Thus, even though the doctor-patient relationship has traditionally been recognized as a central aspect of medical care (Roter, 2000), with the rise of modern medical science doctors' ability to communicate with patients has been lost to a greater attention to the technical, purely biomedical detail. In fact, it has been suggested that medicine care has changed the nature of its communication culture: while before it was characterized by an attention to emotions, the unstated or the nuanced, it is today based on the verbally explicit. In other words, it has shifted from being a 'high-context' to a 'low-context' [1] communication culture (Roter, Frankel, Hall and Sluyter, 2006).

Good communication seems to be the basis of an effective relationship between doctors and patients. Patients consistently articulate their desire for a physician who they trust, has their best interests in mind, and understands and takes their feelings into consideration (Calnan and Sanford, 2004; Golin, Thorpe and Di Matteo, 2007). Good communication in doctor-patient interactions prevents misunderstandings, builds trust between physicians and their patients and, above all, brings better health outcomes. For example, patient-doctor interactions based on good communication have been shown to influence a variety of outcomes including adherence to treatment, recall and understanding of medical advice, and health improvements. On the other hand, the lack of trust in the

[1] These terms are used in anthropology to describe differences in people's cultural behavior as it relates to communication.

doctor-patient relationship, as a result of poor communication (both verbal and nonverbal), may lead to problems such as misunderstandings or the perception of the doctor as not reliable (Golin, Thorpe and Di Matteo, 2007; Taylor, 1992).

Doctor-patient communication has been the object of increasing attention in the recent years. However, in some countries, such as the US or the UK, there is greater attention to health care professionals' communication practices than in others. In Italy, health care providers do not generally receive training or formal instruction in aspects of doctor-patient communication, and, as a result, they are often unaware of the importance of communication in their interactions with the patients.

Research on doctor-patient communication in Italy is also scanty. Yet, there are several reason why doctor-patient communication needs investigation: In the first place, to understand how communication in health care environments is linked to certain Italian-specific cultural patterns; secondly, to understand how different kinds of communication can influence patients' satisfaction and level of trust in their doctors; thirdly, to recognize which communication barriers may arise when care professionals inadvertently use erroneous communication strategies in their interactions with their patients; and finally –and most importantly– to identify possible avenues for overcoming these barriers.

Understanding the dynamics of doctor-patient communication in natural settings can also provide valuable information that can be used in digital media applications to be used in disaster management, emergency relief and interactions with health care professionals.

2. The Present Study

This study presents a preliminary analysis of doctor-patient interactions in Italy. The aim is to set the basis for investigating instances of miscommunication between doctors and patients leading to lack of trust in doctor-patient encounters. The hypothesis is that the unconscious practice of some communication strategies may lead to misunderstandings between doctors and patients and affect the trust in the physician and the perception of his/her reliability.

2.1 Methods and Materials

To get a preliminary overview of how Italian doctors and patients interact, a corpus of existing video materials on doctor-patient interactions, as found on the web, was collected. All the scenes including actual doctor-patient, or doctor-patient's relatives' interactions, were analyzed with the intent to spot glitches in communication that could create possible barriers and lead to misunderstandings and/or contribute to decreasing the level of patients' trust in their doctors. Both authors separately examined the clips many times and identified the salient features of the observed doctor-patient interactions. The authors then compared and discussed their observation notes.

3. Preliminary Findings

The analysis was made on a variety of different communicative instances of both verbal and nonverbal communication between Italian doctors and patients. Three main categories of communicative barriers occurring during the analyzed medical encounters were identified. The categories were the following:

- external interruptions
- verbal barriers
- non-verbal barriers.

3.1 External interruptions

The first and most easily recognizable communication barrier observed was the presence of external interruptions occurring during the doctor-patient exchanges. Typically, these were personal phone calls received by the physician during the interaction with the patient, or people (for example, medical staff) entering the room where the physician was speaking with the patient. These interruptions caused a break in the conversation and in the flow of information exchange between the patient and the doctor. They represented a source of distraction for both people involved in the exchange, and at times made it difficult for the patient and/or the doctor to recall exactly what was said prior to the interruption. However, most importantly, it is possible that these interruptions may contribute to the patients' feeling of frustration: if the doctor does not switch his/her phone off and lets staff enter the consultation room when he/she is speaking with the patients, these may feel that the doctor is not respectful of their needs and is not taking their case in due consideration.

3.2 Verbal Barriers

Another important communication barrier in the doctor-patient interaction was identified in some aspects of the language used by the doctors in their conversations with the patients.

Most noticeably, doctors often used language that was too technical. If a doctor uses technical definitions to describe possible diseases, he/she is using a high level of verbal dominance (Bertakis and Roter, 1991; Kiesler and Auerbach, 2003) and the patients will not be able to fully understand what is being said to them. Patients will also probably feel not satisfied with the way they are being approached, and this will decrease their level of trust in the doctor.

The second relevant communicative barrier identified in the analysis was doctors' strategies for addressing the patients' reactions after the latter were told about their disease. The Italian physicians observed in the video clips often used sentences that minimized or involuntarily underestimated their patients' worries and fears. For example, as a strategy to cheer up the patient, the physicians would use sentences like: "Come on! Take it easy, it's not that bad! Try not to think too much about it!"[2] or "You worry too much!".[3] These sentences, far

[2] The original sentence was: "Su con la vita. Non è così grave.

from having the effect the doctors were hoping to achieve, i.e., to cheer up the patients, work in the opposite direction, that is, they make the patients feel judged or isolated in their fears. Once patients feel the doctor's lack of empathy or that their fears are not taken seriously, their trust in doctors decreases irremediably.

Studies have shown the advantages of patient-centered communication, that is, communication that takes the patient's individual factors (such as age, gender, race, past experience, costs and familiarity with the disease) into consideration and incorporates the patient's perspective and experiences in care planning and decision-making (Wissow et al., 1998). When adopting patient-centered communication physicians provide information, both biomedical and psychological, to patients spontaneously and in response to their concerns. By having a better understanding of the diagnosis and treatment and feeling actively involved in the decision-making process concerning their own health, patients feel that their needs and desires are being met and respond better to care managing their own appointments, filling prescriptions, taking medications, etc. (Hakim, 2011). The style used in patient-centered communication needs to be positive, and aimed at engaging patients in a shared decision making process, eliciting their preferences and understanding their perception of risk and benefit (Hakim, 2011). On the other hand, doctors' using a critical and judgmental attitude towards their patients during a medical consult can lead to problems of miscommunication and affect the relationship of trust and reliability between the two parts.

3.2 Non-Verbal Barriers

The last barrier to good communication and trust between doctors and patients identified in the analysis was represented by doctors' use of nonverbal language (and particularly their use of body language).

It has been claimed that 80% of communication between individuals is nonverbal (Mehrabian, 1968). Thus, doctors' nonverbal behavior and what they communicate to their patients through their bodies is very relevant in highly socio-emotional exchanges such as doctor-patient interactions (Pawlikowska et al. 2012). In general, specific physician behaviors viewed favorably by patients include eye contact, less time looking at medical charts, forward leaning, open body posture, head nodding, use of open hand gestures, and the maintenance of a closer interpersonal distance (Griffith, Wilson, Langer and Haist, 2003; Roter and Rand, 1981; Hall, Harrigan and Rosenthal, 1995).

In our analysis, the observation of doctors' nonverbal communication focused on their posture, hand gestures, gaze and facial expressions. The effect of any physical object that might work as a barrier, such as the doctors' desk, medical charts, etc., increasing the distance between the doctor and the patient, was also taken into consideration.

The analysis revealed that Italian physicians are generally unaware of the meanings of non-verbal language in their interactions with the patients. The physicians' were sometimes leaning backwards, increasing the distance and lack of empathy with the patients and communicating little interest in the patients' complaint. In addition to leaning backwards on their chairs, doctors might use other gestures that can be interpreted as lack of empathy or participation, as well as perplexity, evaluation, or closure. For example, while listening to the patient or the patient's relatives, they might use a variety of chin-stroking gestures, signs indicating that the listener has negative thoughts, is evaluating or is being critical about what the speaker is saying. Also, the physicians might often have their hands clenched in a raised or middle position, signaling –again– a negative attitude or little openness towards the interlocutors. The physicians might also often have their hands joined as in prayer, a typical Italian gesture used to express frustration –a sign that can have the same effect as the use of language minimizing the patient's feelings and fears. Finally, Italian physicians showed little awareness of the importance of gaze and eye-contact in communication: In their interactions they might use a top-down gaze to address their interlocutors, a gesture that may be interpreted as a signal of dominance and does not support a relationship based on partnership and empathy.

As for physical objects that might work as barriers in doctor-patient communication, all videos showed that Italian health care practitioners hold their consultations with their patients, or with the patients' relatives, sitting behind their desks. Doctors also often hold medical charts for extended periods of time while talking to their patients. Studies have shown that seating arrangements are an important factor in determining the patients' evaluation of the physician and that objects that are interposed between the physician and the patient may work as a distancing device to the detriment of communication (i.e., Griffith, Wilson, Langer and Haist, 2003; Roter and Rand, 1981; Hall, Harrigan and Rosenthal, 1995). In countries like the US health care professionals tend to avoid interacting with their patients across their work desk to reduce their distance from their patients; in Italy, on the other hand, the space in the doctor's office is still arranged in the traditional fashion, so that the physician and the patient speak at each other by sitting at the opposite sides of the desk, which amplifies the distance between the speakers.

4. Discussion and Conclusion

This preliminary analysis of the corpus of short videos on Italian doctor-patient interactions was aimed to identify and categorize health care providers' use of linguistic and non-linguistic strategies that might create communication barriers in the doctor-patient relationship. The purpose of this analysis was to gain information to plan a more systematic investigation of the verbal and

La prenda con filosofia, e cerchi di pensare ad altro".

[3] Translated from the original phrase: "Lei si abbatte troppo".

non-verbal communication strategies commonly used by health-care professionals in Italy as compared to the strategies used in other countries.

This analysis shows that a number of factors, both verbal and non-verbal, contribute to the transmission of meaning in doctor-patient interactions. Some of the strategies used by doctors when speaking to patients may result in unsuccessful communication, with detrimental effects in many aspects of the management and provision of health treatments. Some of the dynamics observed in this study may, in fact, be specific to the Italian culture and need to be investigated in depth. Our data suggest that the Italian doctor-patient communication is often univocal, and characterized by the doctors' dominance in both verbal and nonverbal behavior; also, doctors appear not to take into proper consideration patients' feelings and ideas.

These characteristics clearly do not favor a trust relationship. At the same time, while Italian doctors may be relatively unaware of the consequences of poor communication in their interactions with patients, Italian patients may have lower expectations, as compared to patients in other countries, from their communications with doctors, as they have generally not yet been exposed to communication that is less traditional, more patient-centered. At the same time, communication between doctors and patients might be successful –to some extent- in spite of doctors' poor communication skills, as it has been shown that in conversations, and particularly in certain contexts of urgency, collaborative processes are at work between the participants aimed at making communication flow and avoid miscommunication (Vogel, 2013). In fact, it has been recognized that the social process of interaction in conversation plays a central role in the cognitive process of mutual understanding. Listeners who participate in a conversational interaction go about understanding and thus trust their interlocutor much more than those who are excluded from it (Schober, Michael and Herbert, 1989). However, more needs to be known of the dynamics of doctor-patient interactions.

Future investigations of doctor-patient interactions will be carried out based on the analysis of audio-visual material collected on purpose. Having a detailed knowledge of what patterns mostly affect communication in natural settings will also provide important information to implement in digital devices.

5. Acknowledgements

The authors wish to thank Giampietro Vecchiato for useful suggestions about doctors' minimizations strategies.

6. References

Beck, U. (1992). *Risk society*. London: Sage.

Bertakis, K., and Roter D.L., (1991). The relationship of physician medical interview style to patient satisfaction. *Family Practice*, 32, pp. 175-81.

Calnan, M. and Sanford, E. (2004). *Public Trust in Health Care: An Agenda for the Future Research*. London: The Nuffield Trust.

Golin, C.E., Thorpe, C. and Di Matteo, M.R. (2007). Accessing the patient's world: Patient-physician communication about psychosocial issues. In J.L. Earp, E.A. French & M.B. Gilkey (Eds.), *Patient Advocacy for Health Care Quality: Strategies for Achieving Patient-Centered Care*. Mississagua, Ontario (CA): Jones and Bartlett Publishers Int., pp. 185-212.

Griffith, C., Wilson, J.F., Langer, S. and Haist, S.A. (2003). House staff nonverbal communication skills and patient satisfaction. *Journal of General Internal Medicine*, 18, 170-174.

Hall, J.A., Harrigan, J.A. and Rosenthal, R. (1995). Nonverbal behavior in clinician-patient interaction. *Applied & Preventive Psychology*, 4, pp. 21-37.

Hall, J.A, Roter, D.L., Rand, C.S. (1981). Communication of affect between patient and physician. *Journal of Health and Social Behavior*, 22, pp. 18-30.

Hakim, A. (2011). Perception of risk and benefit in patient-centered communication and care. *Patient Intelligence*, 3, pp. 39-48.

Kiesler, D., and Auerbach, S. (2003). Integrating measurement of control and affiliation in studies of physician-patient interaction: the interpersonal circumplex. *Patient Education Counselling*, 74, 1707.

Mehrabian A. (1968). Communication without words *Psychology Today*, 2, pp. 52-55.

Pawlikowska, T., Zhang, W., Griffiths F. and Van Dalen, J. (2012). Verbal and non-verbal behavior of doctors and patients in primary care consultations – How this relates to patient enablement. *Patient Education and Counseling*, 86(1), pp. 70-76.

Roter, D.L. (2000). The enduring and evolving nature of the patient-physician relationship. *Patient Education and Counseling*, 39, pp. 5-15.

Roter, D.L., Franke, E.M., Hall, J.A. and Suyter, D. (2006). The expressions of emotions through nonverbal behaviour in medical visits: Mechanisms and outcomes. *Journal of General Internal* Medicine, 21 (suppl1), S28-S34.

Roter, D.L. and Hall, J.A. (1992). *Doctors Talking to Patients/Patients Talking to Doctors: Improving Communication in Medical Visits*. Westport, CT: Auburn House.

Schober, M.F. and Herbert, H.C. (1989). Understanding by addressees and overhearers. *Cognitive Psychology,* 21, pp. 211-32.

Taylor, T.J. (1992). *Mutual Misunderstanding: Scepticism and the Theorizing of Language and Interpretation*. Duke University Press.

Vogel, C. (2013). Attribution of mutual understanding. *Journal of Law & Policy*, 21(2), pp. 377-420.

Wissow, L.S., Roter, D.L., Bausman, L., Crain, E., Kercsmar, C., Weiss, K., et al. (1998). Patient-provider communication during the emergency department care of children with asthma. *Medical Care*, 36, pp. 1439-1450.

Anonymous FreeSpeech

Carl Vogel

Computational Linguistics Group
Centre for Computing and Language Studies
School of Computer Science and Statistics
Trinity College Dublin
Dublin 2
Ireland
vogel@tcd.ie

Abstract

In public and private discourse, some may be heard to express disquiet about the supposed dangers of anonymity. Anonymous suggestion boxes may be classed with anonymous accusation of crime with the accusation forming the basis of legal proceedings. In some contexts, anonymity does appear to create danger. However, other contexts reveal that important benefits accrue from having the possibility of anonymous expression. Some of the literature on behavioral impacts of anonymity is reviewed with the aim of analyzing social media systems in light of their support for anonymous contribution. A new system is described. The new system supports anonymous communication, while thwarting some of the obvious risks that anonymity affords.

Keywords: social media, anonymity, free speech, sentiment analysis, meme analysis, ephemera, dynamic social groups

1. Introduction

Popular thinking is conflicted on the value of anonymity. The US Constitution's First Amendment protects anonymous free speech, and accordingly the US courts have established guidelines for discovering the identity of anonymous internet posters (Den, 2001; Mob, 2007; Kri, 2008). It is appears to be a coherent position to hold that anonymity has no social value, since it is an asocial concept. Davenport (2002, p.33) wrote "By allowing anonymous Net communication, the fabric of our society is at risk." However, it will be seen (§2.) that if anonymous communication is asocial, it is paradoxically so, since there are conditions in which group identity is more strongly present for anonymous rather than identified communicators. Neumann (1996) describes the uneasy availability of anonymity and argues that accountability must always compromise anonymity. The CEO of Facebook has called for the end of internet anonymity;[1] however, he has a commercial interest in the users of his company's system being identifiable. The UK Defamation Act 2013 provides means to preserve the anonymity of posters, but passes accountability to hosting website operators.[2]

It is a reasonable principle of law that if one is to be tried, one should know what the accusation is and who the accuser is, and this is contrary to the possibility of anonymity. On the other hand, society has not yet established effective means of protecting "whistle-blowers", who are in many cases neither directly harmed by the accused nor have any ground to benefit from a successful claim, from retribution by the accused or the system in which the accused operates. Social media have been attributed a role in enabling societal reform (Khondker, 2011; Eltantawy and Wiest, 2011), and prohibitions of anonymity in social media have been cited as an obstacle to reforms (Youmans and York, 2012). However, one mob's societal reform is another's treasonous rebellion. Nonetheless, a number of social media platforms have emerged in support of anonymous communication. While each of these has a niche, and acknowledging that the actual utility of any tool frequently diverges from its intended use, the architecture of a new platform for anonymous communication is detailed.

This article reviews some of the literature on the potential positive or negative impacts of anonymity (§2.). The impacts are mixed whether the interactions are online or in person, both with respect to likelihood of aggression and generosity. It also describes extant systems that support anonymous communication to varying degrees. The design of a new system is specified (§3.): it is a system that supports anonymous communication in its outward facing function, and it provides a testing ground for methods of text analytics and analytics of mult-modal content within its system architecture.

2. Background

2.1. Effects of anonymity

In a lab based setting Connolly et al. (1990) showed that anonymous groups working jointly on idea generation produced more ideas than identified groups, particularly when a critical mode of interaction was induced by a confederate. Postmes et al. (2001) demonstrated that anonymous groups would exhibit behaviors consistent with introduced norms, while identified groups did not – this finding alone is a powerful rebuttal to the notion that anonymity is inherently asocial. Smith et al. (2007) obtained results which showed that participating individuals with strong group self-identification behaved according to group norms even in anonymous conditions. Diener (1976) showed in a lab

[1] See a 2011 staff journalist article in *The Daily Mail* – www.dailymail.co.uk/news/article-2019544/Facebook-director-Randi-Zuckerberg-calls-end-internet-anonymity.html – last verified February 2014

[2] http://www.legislation.gov.uk/ukpga/2013/26/section/5/enacted – last verified February 2014.

setting involving male undergraduate students that group presence yielded decreased aggression, but that providing anonymity had no effect on aggressive behaviors. Piazza and Bering (2008) conducted a mixed gender study in which economic punishment behaviors were more pronounced in identified conditions than where anonymity was preserved.

On the other hand, Donnerstein et al. (1972) found aggression in lab situations invoking race conflict more likely in situations of preserved anonymity than where the aggressor could be identified. More recently, Ellison-Potter et al. (2001) found that people drive automobiles more aggressively when anonymous than when identified. In a natural field experiment Alpizar et al. (2008) found public donors to be 25% more generous than anonymous benefactors. Rule-breaking has been found in lab settings to be more likely where participants are not identifiable and not accountable for their actions (Nogami and Takai, 2008).

Lapidot-Lefler and Barak (2012) found that lack of eye contact was a greater contributor to toxic online communications than anonymity in computer mediated communications in which visibility and eye-contact were mediated by web-cams, and anonymity, by random online aliases.

Stone et al. (1977) showed that in student evaluations of lecturers, signed statements were more positive than anonymous feedback.

Kidder et al. (1977) demonstrated gender differences in behavior where decisions regarding generosity would become public or remain anonymous, with men choosing a more generous response than women when the decision would remain anonymous (the reverse of when decisions were made public). Durant et al. (2002) also noticed that anonymity guarantee yielded higher quality data in medical self reporting than where identification was supported.

The literature mentioned above demonstrates both support for anonymity as a social construct and potentially negative impacts of anonymity. Popular opinion on anonymity is perhaps conflicted because the value of anonymity is context dependent.

2.2. Extant systems supporting anonymity

I distinguish between systems that, perhaps despite stated terms and conditions of use, do not force identification of users by verifying the identity of the person registering but support the identification of posters or creation of online persona (e.g. Facebook, Twitter, Reddit, Tumblr), and those systems that purport to maintain the anonymity of posters. Kwikdesk[3] provides an outwardly anonymous messaging service like Twitter, with messages that last only a very short amount of time, but requires user registration. Self-created account names have been analyzed to assess the likelihood that account names that appear in sentiment lexica (e.g. SentiWS (Remus et al., 2010)) are likely to host content that anyone wishing to filter offensive content might wish to exclude, and found a correlation between negative sentiment user identifiers and such content (Vogel, 2013). While self-created account names do not reveal identity, they do lend well to establishing distinctive online persona. This latter category, systems that do not record details of posters, are described further below.

Some sites allow posters to post content via random identification numbers into threads of communication, for example `4chan.org`.[4] From third party perspectives, unless posters reveal themselves, they are anonymous. To the system managers, posters must reveal an effective email address, and presumably system logs record connecting IP address information — the terms and conditions suggest that the system managers have the means of identifying and permanently banning posters, which suggests a relatively high level of identification. Postings to this service have been studied in light of the fact that a high-volume of postings exist, despite the very ephemeral nature of some of the threads and estimated 90% anonymous posting (Bernstein et al., 2011). This demonstrates that a social niche is filled through the possibility of such a facility existing to support outward facing anonymity, in spite of the frequently offensive content (Knuttila, 2011). Schoenebeck (2013) points out that YouBeMom.com provides a positive means of expression of negative emotions through the dis-inhibition enabled by anonymity.

A service provided by smalltime.com[5] allows users to post messages anonymously to the next person who posts a message. The system functions something like an ephemeral message-in-a-bottle, not directly supporting a general view of sentiments expressed by many. A poster need not post self-identifying information, nor information that reveals who intended readers are. In contrast, hadtosay.com[6] allows users to send email messages to named individuals, but from an anonymous address, with the proviso that the content of the message is posted to a public forum. This can have potential negative consequences for recipients whose email contact details are deposited with the system; however, it is curious to see the preponderance of sentiment that emerges for occasions like Valentine's Day.

2.3. Observations

The review here has attempted to be suggestive rather than exhaustive. The literature reveals both benefits and risks associated with anonymous communication. Social media systems support anonymity to varying degrees, most depending on the construction of online social personae, if not actually identifying individual posters. Some that support greater levels of public anonymity still require system registration, and nonetheless have become a clearinghouse for content of dubious merit.

3. Design of a new system

This section provides a specification of FreeSpeech, a system for anonymous expression and perusal of ideas. The system is currently under prototyping.

3.1. Desiderata

As in any system design, some principles are formulated in positive terms and others are proscriptions.

[3] See `kwikdesk.com` – last verified February 2014.

[4] Last verified February 2014.

[5] See `http://www.smalltime.com/anon.html` – last verified Feb 2014.

[6] See `http://hadtosay.com/` – last verified Feb 2014.

As a matter of principle the system should afford users the possibility of posting text and reading texts. The system should provide comparable facilities for other modalities; however the primary focus should be the articulation of thoughts. Users may wish to tag the text in various ways to indicate perceived categories appropriate to the text. However, the system should, in the background, supply continuous indexing that re-categorizes submissions in additional, alternative manners.

The system should be built on a principle that posters and readers may retain anonymity. It should be possible for a user to express thoughts without registering an email, IP address or other identifying signals, and similarly it should allow other users to examine expressed thoughts, via an assortment of possible searches and indexing mechanisms.

The system should also respect the rule of law in the jurisdiction in which it is hosted, and to the extent possible given inconsistencies, internationally. It should facilitate the expression of thought and estimation of current thinking. Content that is not legal should be rejected by the system; content that is merely offensive or inappropriate should be quarantined.

Preservation of anonymity may prevent the system from allowing the geo-location of current thinking, but should facilitate analysis of the time course of popular thought.

The system should not provide support for the maintenance of persona, whether of actual individuals or constructed for an online presence. The primary function of the system is not maintenance of synchronous or asynchronous online dialog.

The system should support multi-lingual views on the content recorded.

Through the totality of its public facing functions as supported by the system-interior functions, the system should provide a testing ground for research associated with the system-interior functions, content analytics, in particular. It should be possible to easily adapt the system to include alternative classifiers and test the relative differences and similarities obtained by integrating alternative methods.

3.2. Public features

The individual posting content may do so into a window amidst extant content, and doing so will automatically tag the new content in relation to the extant visible content, or posting may be made from a top level portal that comes with no *a priori* tags. In either case, the poster may add additional tags to the content.

An individual seeking to monitor thought expressed in the system should have the possibility of seeing raw content or content aggregation summaries. Content should be visible in relation to the most recent postings or in relation to categories. The categories may be supplied as a query term from the monitor or dominant themes. Category related views should also be possible with respect to past temporal intervals.

It should be possible for users to register support or dissent with respect to ideas hosted by the system, and the time course of support and dissent should be available to monitors. Support or dissent may be registered in categorical terms or supplemented with textual comment. Similarly, search term use will also be aggregated and reported for the benefit of monitors wishing to sample global thought.

While it will not be possible in general to sample content relative to geo-location parameters, because of the multilingual support, it should be possible to view content with language of input as a parameter. Equally, it should be possible to view (via automatic translation), content made available cross-linguistically.

The system requires active monitoring because as a matter of principle it does not record email addresses of individuals wishing to be informed of updates.

The system should not declare which system-interior classifiers are at work in clustering postings together at any time of use.

3.3. System-interior features

On the system side, there are facilities for multi-dimensional indexing of content. From the perspective of research in content analytics and machine learning. The indexing of the content comes from the tags supplied by the user, or supplied contextually at the point of posting, or from background indexing processes.

Interfacing to machine translation will be necessary possibly as a remote online service. Language detection services may be used in order to tag input textual content (Cavnar and Trenkle, 1994; van Noord, 1998).

Background analysis of posted content is intended to support filtering out spam, defamatory claims, pornography and so on.[7] At the same time, it should be possible to classify and index posts according to multiple criteria: terms used, user-supplied tags, system detection of similarity to other posts, system detected anomaly of content. The same general techniques that are used for detecting effects of source language on translation and stylistic quality can be used to detect content based anomalies (Vogel et al., 2013; Moreau and Vogel, 2013). In fact, it is a design principle of the system that it should provide a scenario for testing out theories and methods of content classification.

The system will thwart posting of links to external content. Thus, pre-processing of posts will remove external links. Content beyond text that is uploaded will be filtered. While the system will support the posting of visual content, this will be secondary to text.

Indexing will be based on user-supplied key-words as well as system discovered relevant terms. Index expansion will be facilitated with reference to machine readable lexical resources like WordNet.

Tools for detection of named individuals and automated anonymization may be necessary in order to avoid compromising the system (or supporting defamation).

A database suitable for large scale data sets will accommodate the data and its multiple indices.

3.4. Interaction types

3.4.1. Perusal

A person may approach the site from its front end web presence, and through a default language setting will be presented with a few alternative views of the data recorded in

[7] See O'Brien and Vogel (2003), for example.

the back-end, the n most: recent posts; endorsed posts; denounced posts; anomalous posts; active user tagged topics; active system categorized topics. A reader may therefore browse posts through any of those mechanisms for ranking past postings, or a user may search for posts, which will be indexed by their content and also by tags (user supplied, and system generated, as described above). The possibility of posting will also be possible at this point (see §3.4.2.), and an alternative access point will support posting only, without a view of current topics.

In selecting a post to view, an additional interface context is created in which other posts related by topic, recency, etc. are also visible. Posts will have unique identifiers, but this will not include any system-generated signal of who the user is. Thus, readers should be able to quickly view linked (and linking) posts and their contexts. Exploring a post should give the impression that there is an unending potential for expanding links from posts to related posts, even if those links are not user-supplied links, but links based on shared index terms.

Tools for aggregate summary analytics will also be available. That is, it may be interesting to some users not to read posts directly, but to study the volume of postings on topics over time, or as a function of language of posting, etc. Tools for visualizing trend analysis will be supplied.

3.4.2. Posting

A person may post content into the context supplied by postings perused. This may, if attending to immediacy and certain contextual features, have the effect of sustaining an interaction between contemporaneous posters. However, postings all without user identifiers, will make it difficult to sustain a dialogue as in other online social media – direct quotation may be the easiest means of directly signalling such interactions, but post identifiers may link posts, too. Rather, postings will be amid other postings, creating the effect of ideas interacting with other ideas, rather than personalities interacting with personalities. On entering a contribution, a user posting will have the option of including multi-media files in the post (knowing that these will not appear with the post immediately, as the media file filtering is expected to take longer than text filtering) as well as key phrases that contribute to the index of the post. A post may link to the post-id of another post, but not outside the system.

A user will not have a facility to edit or delete any post after accepting submission. Some edit facilities, including draft preview, are supported. Additionally, a feature allows a user to decline to submit a draft, and in that case it is not recorded in the system except as a count of aborted drafts. Once the draft is committed to the system, the post ceases to be controlled by the user posting content.

Once a post passes text filters, it is available for perusal to others in the original context, and in other contexts linked by time, user-supplied key phrases, and key phrases generated from the text, and through index expansion. Thus, a post may appear to a user to be part of one topic of discussion, but may, through indexing be visible to others in relation to topics of discussion that the original poster perhaps did not consider.

3.5. Anticipated use

Given that social media exist for interactive social communication, it is not anticipated that the FreeSpeech system will occupy this dimension of interaction. Rather, it is foreseen that the system will be a living meme monitor, serving to digest the global pulse in a way that extensively supplements information about trending topics or top search terms. It is imagined that the system will attract use more general than the automatic mood detector supplied by WeFeelFine.org which trawls the internet for content that contains the expressions "I feel" and "I am feeling".[8] Yet neither is the system likely to emerge as a crowd-sourced compendium of knowledge in competition with Wikipedia (not least because wiki functions such as the potential for user editing or deleting of posts will not be supplied – not for the others' posts, nor their own).

Some individuals may try to subvert the spirit of anonymity either by naming themselves in their posts or by creating trademark identifiers that effectively tag their contributions and establish a public persona. Nonetheless, it is hoped that the system will facilitate debate of ideas in their own sake, and without reference to their proponents. Within this particular social medium, the role of "key movers" will be diminished through anonymity, so that ideas may be applauded or detracted because of the ideas themselves rather than because of the personalities expounding them or endorsing them.

It is foreseen that in order to overcome system obstacles to identifying users will make it cumbersome to use the system for social chat. Therefore, while it might be feasible to use the system to assemble a "flash mob" (Gore, 2010), it would be very difficult to foster a pen-friendship through the system.

Background filter that remove obnoxious content are meant to have the effect of deterring the posting of such content and use of the site for sharing such. Similarly, it is not expected to provide affordances for copyright violation.

The system is not proposed with any model of economic exploitation in mind as a driving principle; rather, the driving principles are in the provision of a tool for expressing thoughts, and for testing out methods of content analytics. However, it might be noted that the system, as described, is compatible with having a side panel that displays relevant advertizing. Moreover, it is compatible with the goal of the system in providing a testing bed for research into methods of content analytics that one might explore study of relations of compatibility (as opposed to inappropriateness) between content displayed in the primary content areas and any ads displayed in such an advertizing area, as an instance of the general problem of monitoring compatibility of message across types of media (Pastra, 2006; Pastra, 2008). Such ads would presumably have to respect the constraints imposed on regular posts.

4. Evaluation

The interactive elements of the sytem will be evaluated through its uptake: the system may be deemed successful if it demonstrates growth in usage. By construction, it will

[8] See http://www.wefeelfine.org – last verified February 2014.

not be possible to measure the number of distinct users.[9] Rather it will be necessary to measure the quantity of posts, diversity of topics, and depth within topics.[10] Empirical hypotheses underlying the system design may be simultaneously tested. It is a hypothesis of this work that sufficiently many alternative systems exist to support fostering relationships or hosting particular kinds of multi-modal content. It is not imagined that this system will compete with those on the same ground, because this system is not designed to facilitate that sort of content managment. Nonetheless, it will be possible to evaluate the quantity of postings that are attempted, but filtered through the various possibilities of excluded contexts. The ratio of desirable to undesirable content in this system will be interesting to monitor in itself and in relation to the same ratios for other social network systems.

The capacity of the system to support research into analysis of the competition of ideas as a dynamic system, meme analysis, is also to be evaluated. This will require study once sufficient content accumulates within the system.

The back-end facilities are designed in order to provide a sandbox for developing and evaluating methods of content analytics. Evaluation of the proposed system in this dimension relates to the ease with which alternative classifiers may be correlated with each other and differentiated from each other, for example.

5. Final remarks

This article has defended the role of anonymity in online communication and expression of ideas. It has suggested a new system for facilitating anonymous contribution of ideas into public debate. The system provides features for users to express their ideas and for others to monitor ideas that are current (or which have subsided in general interest). The system itself provides a locus for research in multilingual multi-modal content analytics. It may be argued that the design described here is ambivalent in approaching content beyond the textual mode. It is true that the prohibition of external links and severe filtering of other content provides a greater level of visual filtering than is applied to text. These features are seen as limiting the potential for an anonymity preserving communication device to be used in ways that other online facilities already appear to support with ample resources. That should leave interest in the system presented here primarily for expressing thoughts that require consideration.

The value of anonymous communication during disaster management remains for further consideration. Manifestly, during disaster management, the provenance of messages matters to citizens a great deal, and trust is accorded to messages seemingly as a direct function of the trust accorded to the public persona of the source. Without automatic geolocation of postings in the system described, it would be challenging to use the system in order to assess the situation "on the ground" as reported by citizen observers, and it would be difficult to convey messages of urgency to the public one would hope to respond, if one were using the proposed system as a channel for such communications. However, in the case of emergencies that are more politicized, it is easy to see the value that could be provided by anonymous communication mediated through FreeSpeech.

[9] Arguably, systems like Facebook are not in a position to provide reliable methods for estimating unique user numbers accurately, either; however, here, it is in principle not possible.

[10] Developing metrics for topic diversity and depth presents an interesting scientific challenge.

6. Acknowledgements

I am grateful to Eamonn Lawlor for prototyping some of the features of FreeSpeech suggested here in the context of his final year project on the undergraduate of honors course in Computer Science, Linguistics and German at Trinity College Dublin; he also alerted me to the existence of 4chan and Kwikdesk.

7. References

Francisco Alpizar, Fredrik Carlsson, and Olof Johansson-Stenman. 2008. Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in costa rica. *Journal of Public Economics*, 92(56):1047 – 1060.

Michael Bernstein, Andrs Monroy-Hernndez, Drew Harry, Paul Andr, Katrina Panovich, and Greg Vargas. 2011. 4chan and /b/: An analysis of anonymity and ephemerality in a large online community. In *Fifth International AAAI Conference on Weblogs and Social Media*.

W. B. Cavnar and J. M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175. Las Vegas, NV, UNLV Publications/Reprographics.

Terry Connolly, Leonard M. Jessup, and Joseph S. Valacich. 1990. Effects of anonymity and evaluative tone on idea generation in computer-mediated groups. *Management Science*, 36(6):689–703.

David Davenport. 2002. Anonymity on the internet: Why the price may be too high. *Communications of the ACM*, 45(4):33–35.

2001. Dendrite International v. Doe, 775 A2d 756 (N.J App. Div.).

Edward Diener. 1976. Effects of prior destructive behavior, anonymity, and group presence on deindividuation and aggression. *Journal of Personality and Social Psychology*, 33(5):497 – 507.

Edward Donnerstein, Marcia Donnerstein, Seymore Simon, and Raymond Ditrichs. 1972. Variables in interracial aggression: Anonymity, expected retaliation, and a riot. *Journal of Personality and Social Psychology*, 22(2):236 – 245.

Lauren E. Durant, Michael P. Carey, and Kerstin E.E. Schroder. 2002. Effects of anonymity, gender, and erotophilia on the quality of data obtained from self-reports of socially sensitive behaviors. *Journal of Behavioral Medicine*, 25(5):439–467.

Patricia Ellison-Potter, Paul Bell, and Jerry Deffenbacher. 2001. The effects of trait driving anger, anonymity, and aggressive stimuli on aggressive driving behavior. *Journal of Applied Social Psychology*, 31(2):431–443.

Nahed Eltantawy and Julie Wiest. 2011. The arab spring—social media in the egyptian revolution: Reconsidering

resource mobilization theory. *International Journal of Communication*, 5(0).

Georgiana Gore. 2010. Flash mob dance and the territorialisation of urban movement. *Anthropological Notebooks*, 16(3):125–131.

Habibul Haque Khondker. 2011. Role of the new media in the arab spring. *Globalizations*, 8(5):675–679.

Louise H. Kidder, Gerald Bellettirie, and Ellen S Cohn. 1977. Secret ambitions and public performances: The effects of anonymity on reward allocations made by men and women. *Journal of Experimental Social Psychology*, 13(1):70–80.

Lee Knuttila. 2011. User unknown: 4chan, anonymity and contingency. *First Monday*, 16(10).

2008. Krinsky v. Doe 6, 159 Cal.App. 4th 1154 (Cal. Ct. App.).

Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, 28(2):434 – 443.

2007. Mobilisa v. Doe, 170 P.3d 712, (Ariz. Ct. App.).

Erwan Moreau and Carl Vogel. 2013. Weakly supervised approaches for quality estimation. *Machine Translation*, 27:257–280.

Peter G. Neumann. 1996. Inside risks: Risks of anonymity. *Commun. ACM*, 39(12):162–, December.

Tatsuya Nogami and Jiro Takai. 2008. Effects of anonymity on antisocial behavior committed by individuals. *Psychological Reports*, 102:119–130.

Cormac O'Brien and Carl Vogel. 2003. Spam filters: Bayes vs. chi-squared; letters vs. words. In Markus Alesky et al., editor, *Proceedings of the International Symposium on Information and Communication Technologies*, pages 298–303.

Katerina Pastra. 2006. Beyond multimedia integration: Corpora and annotations for cross-media decision mechanisms. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*, pages 499–504.

Katerina Pastra. 2008. Cosmoroe: A cross-media relations framework for modelling multimedia dialectics. *Multimedia Systems Journal*, 14(5):299–323.

Jared Piazza and Jesse Bering. 2008. The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, 6(3):487–501.

Tom Postmes, Russell Spears, Khaled Sakhel, and Daphne de Groot. 2001. Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, 27(10):1243–1254.

R. Remus, U. Quasthoff, and G. Heyer. 2010. SentiWS - a publicly available German-language resource for sentiment analysis. In *Proceedings of the 7th International Language Ressources and Evaluation (LREC'10)*.

Sarita Yardi Schoenebeck. 2013. The secret life of online moms: Anonymity and disinhibition on youbemom.com. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Joanne R. Smith, Deborah J. Terry, and Michael A. Hogg. 2007. Social identity and the attitudebehaviour relationship: effects of anonymity and accountability. *European Journal of Social Psychology*, 37(2):239–257.

Eugene F. Stone, Mark D. Spool, and Samuel Rabinowitz. 1977. Effects of anonymity and retaliatory potential on student evaluations of faculty performance. *Research in Higher Education*, 6(4):313–325.

Gertjan van Noord. 1998. http://odur.let.rug.nl/ vannoord/TextCat/. last verified 25 June, 2001.

Carl Vogel, Ger Lynch, Erwan Moreau, Liliana Mamani Sanchez, and Phil Ritchie. 2013. Found in translation: Computational discovery of translation effects. *Translation Spaces*, 2(1):81–104.

Carl Vogel. 2013. Multimodal conformity of expression between blog names and content. In Péter Baranyi, Anna Esposito, Mihoko Niitsuma, and Bjorn Sølvang, editors, *4th IEEE International Conference on Cognitive Infocommunications*, pages 23–28.

William Lafi Youmans and Jillian C. York. 2012. Social media and the activist toolkit: User agreements, corporate interests, and the information infrastructure of modern social movements. *Journal of Communication*, 62(2):315–329.

Trust in Social Media for Disaster Management

Ms. Sadhbh McCarthy (CIES), Mr. Martin Mackin (Q4PR)
Centre for Irish and European Security (CIES),Q4PR
www.cies.ie; www.q4pr.ie
sadhbh@cies.ie

Abstract

Social media use in times of crisis and disaster has the potential to deliver many benefits. Uniquely, social media is a bilateral communication medium, which allows information to be conveyed and solicited between both the citizen and state. Social media is also a domesticated technology, which creates potential paths for surveillance crossover from the public into the private sphere. Any indications that state institutions (revenue commissioner, police, social welfare department) are abusing the technology - for example, by prying into citizens' behaviour - will have negative implications on the trust relationship between state and citizen. Crucially, it is this trust relationship that needs to be fostered, as those very state institutions will call upon it during times of crisis.

Social media has also become deeply politicised, wherein parties, for their own ends, either valorise it as a societal innovation, an important component and contributor to the digital economy, or demonise it as the cause of societal ills (e.g. London riots, cyber bullying, harmful online fads) - of course, both are reductive, serving only to undermine and distort the true societal impact of social media.

At its worst, social media can be can be seen as an engine of moral panic, occupying a space where headline driven journalism and opportunistic politics collide, which inevitability leads, over time, to further erosion of the bonds of trust between government stakeholders and the citizenry.

In the absence of transparent, structured social media strategies, as part of a piece of coherent government policy and implemented at institutional level, the successful use of social media in disaster management (response and recovery) will be severely restricted. As it stands, government or party political interventions and engagement with social media tend to be focussed on narrow matters of voter mobilisation and engagement. While this approach can, in the short term, serve to deliver – as part of a set of campaigning tools – a temporary dividend in support, the cynical motives of it will, over time, only serve to undermine citizen trust. The power to harness the real benefits of social media will continue to be undermined as long as the societal impact of government constructs around social media do not recognize the important dynamics of such a trust relationship.

A new paradigm of engagement, one that is rooted in trust, transparency and open dialogue, must be engendered before social media can fully realise its potential as a robust, reliable, effective communications component within crisis management strategies for periods of instability.

The role of Trust

Why Trust Matters

At a time of crisis trust is critical, and in particular the trust dynamic between the citizenry and political / governmental stakeholders. Trust is the bond whereby governmental stakeholders can direct and shape behavior at moments of crisis. Where trust is weak or does not exist, this critical bond is eroded and the implementation of disaster management strategy can be compromised. Trust enhances solidarity. Distrust fosters fragmentation. Trust matters. But so do governments. It is governmental stakeholders who must manage disasters and it is government actors who must often drive this process.

An examination of this concept is timely, in light of revelations by the former C.I.A contractor, Edward Snowden, of the extent to which social media surveillance has been co-opted by intelligence agencies world-wide.

Trust Capital and the Government Dynamic

The accumulation of trust capital requires consistency and commitment on the part of governmental stakeholders. Governments however can be anything but consistent. Institutions of the state are shaped by the demands of the political class. Politics is buffeted by an increasingly frantic media spin cycle. It can be driven by clientelism and shaped by elusive ideologies. On a day to day basis it is subject to what Harold McMillan referred to as 'events dear boy', that is to say politics is permanently prey to the exigencies of the day to day. Truth may be the first casualty of war, but trust can be a daily casualty of the political fray. For successful implementation of social networks in disaster management we must first understand this concept of trust capital, before devising and managing a comprehensive government communication strategy that embeds the dynamic.

Description of social networks (media) as a unique technology with specific attributes

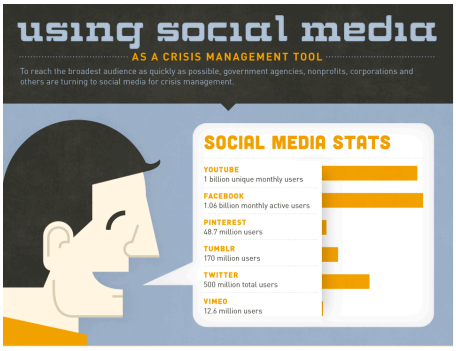
Societal Dimension of Social Media

When we look at social media as a tool for crisis management, the predominant considerations are related to the broadcast features of the technology or the crowdsourcing potential. As such, we aspire to use the amplification dimensions of the technology in a manner that is intended to be benign.

However, social media is not just a passive technology that has one controller. Social media is truly a "societal" technology.

To understand the societal aspects of this we need to address the following issues:

Who are the social media users? [i]



The assumption here is that we are connecting with (broadcasting) or pulling from (crowd sourcing) society as a whole, comprising:

- Individuals
- Communities
- Anyone who is a member of a group or club with a social media profile
- Professional users (bloggers; journalists)
- Personal users (Facebook; Twitter; Instagram; Tumblr)
- **Anyone** who uses a Social Network

For disaster management we need to ask who are the social media controllers?

- First responders (e.g. fire & ambulance)
- Emergency services (e.g. hospitals; quarantine centres)
- Civil /society crisis organisations (e.g. Red Cross)
- Government agencies (schools)
- Police services (local, regional & national)
- Defence services (civil- military response e.g. flooding etc.)
- Government & state (political & institutional)

While technological mediation of interactions between citizen and state during times of crisis is not particularly novel, we argue that social media deployment in this context should carry with it special considerations of societal acceptance & sensitivity.

To develop this point, we will touch on the unique aspects of social media that impact on its use in this context.

Social media attributes

Media is a bilateral communication medium, in so far as it can be used to both convey information from the institutions of the state, and to solicit information from citizens. Traditionally we broadcast, using TV, radio, news media, or we garner information through surveys, face-to-face canvassing, or interpretation of actions.

In a more strictly academic sense:
"social media is a unique form of communication that converges three distinct modes - cognition, communication and cooperation - of sociality into an integrated form" [ii].

Fuchs further illustrates these three concepts and their relationship using the following example,

"[O]n Facebook, an individual creates a multi-media content like a video on the cognitive level, publishes it so that others can comment (the communicative level), and allows others to manipulate and remix the content, so that new content with multiple authorship can emerge." (ibid)

While Fuchs cautions that one step does not necessarily lead to the next, the potential exists in the technology for these three activities to combine in the one space (ibid).

So we can see that not only can the crisis management user broadcasts information, but can also ask for responses and merge the information for re-broadcast and citizen commentary or information on fellow citizens in real-time and through limited media channels (e.g. Facebook; Twitter).

When considered in this context, it is not surprising therefore that for citizens to share information and to believe or accept the information they are receiving will require high levels of trust.

In typical unmediated social networks, the element of trust resides in the information provider or data controller, which in this instance will be an institution of the state (see crisis management users above).

Logically we can deduce for effective use of social media in crisis or disaster management a trust relationship must exist between and amongst the parties to the networks.

How social networks are currently instrumentalised by the State

To date in most member states in Europe and at an overall European level, there has been very little effort to create trust networks or understand the importance of creating such networks in the digital sphere.

Trusted networks are seen to relate more to issues of encryption, firewalls, data protection and privacy, rather than the trustworthiness of the content itself.

The Political State

In most political arenas (from European Parliament to national member states) digital technology, in this case social media, is instrumentalised for largely political or policy goals.

Social networks are valorised as tools of global value for everything from popular uprisings in repressive states (e.g. Arab spring), facilitation of micro-financing in under-developed countries, enablers of domesticated activism to the ultimate democratization of society by giving everyone an uncensored voice.

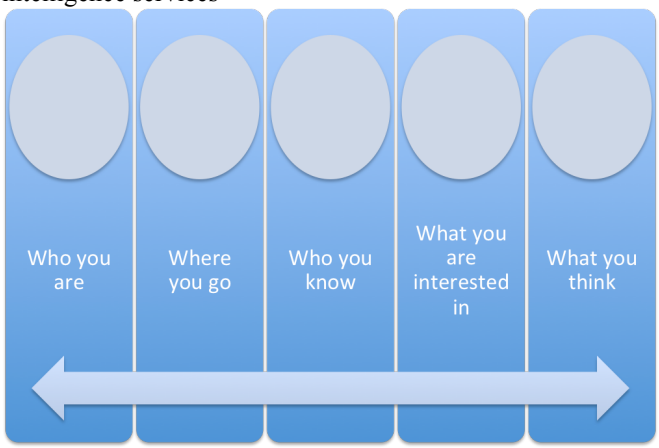
On the other side, social networks are demonised as facilitators (even casual) of civil unrest and rioting (London riots of 2011), mechanisms for self-radicalisation, providers of communication channels between global terrorist cells, promoters of 'slacktivism', and everything from cyber bullying to child pornography.

Of course neither valorisation nor demonisation of a technology is useful. It demonstrates a lack of understanding of what social networks can truly achieve - be it good or bad - and the extent to which it is the society (citizens) behind them who will have the most influence over the outcomes.

The Institutional State

The other way in which the state or governments instrumentalise social media is more operational. That is to say, how they use the technologies for their own advantage. For the most part, governments use social networks to broadcast information or effect efficiencies, e.g. to tell citizens how or when to make their tax returns, or broadcast tweets from traffic police etc. This, of course, can be relatively benign, although again the difference between government information & government propaganda is largely dependent on the political society.

Egregious use of social networks, on the other hand, can be seen played out in citizen surveillance by government agencies, which, of course, will use those same social networks in times of crisis (e.g. police services, intelligence services



The surveillance of individuals, communities, ethnics groups and demographic groups (e.g. men between the ages of 17 & 23) is a common feature of current policing methods. Combining predictive and content analytic tools with Web crawling, site scraping, and (questionably robust) semantic analytics is not unusual. The term Open Source Intelligence (OSINT) has even been readapted to reflect its latest iteration: Social Media Open Source Intelligence (SOCMINT) [iii]

The cost of doing this is low, and with little legal precedent, the over-arching sense of everything you say or do online through social networks being monitored faces little opposition.

This of course brings us back to the crux of the problem: during periods of relative societal harmony, social network users can be subject to propaganda, politicization of the technology and illegitimate (albeit not strictly illegal) surveillance by governments and its agencies.

The Erosion of Trust

Why then would citizens trust those same government agencies in a time of crisis? Are they not more likely to undermine or subvert the process rather than participate and support it?

In addition when the political dynamic engages with social media the results can be unlovely. In this context social media can inhabit a space where moral panic media

and opportunistic politics collide and the bonds of trust with the citizen are potentially eroded further.

Developing and maintain the trust relationship

Maintaining trust means developing a new paradigm of engagement that involves the following key considerations, which taken collectively could be seen as representing a radical, if not naïve, approach:

- Accepting that building trust is never static and that the process cannot be contained within one strategy
- The construction of trust based communications structures
- Strong policies that drive transparency and accountability
- Working to the ultimate creation of a new paradigm of engagement with the citizen

Paradoxically, political actors continuously seek means by which to engender trust as a means of harnessing citizen support. Social media has been embraced to achieve politically beneficial citizen dialogue. However social media is radically different from traditional communications platforms in politics. The communications dynamic is bilateral, in that it can convey messages from institutions of state but it can elicit a response, wanted and otherwise, from the citizenry. And this is the challenge. Social media is rooted in a freeform dialogue with which government continues to struggle. Social media demands an 'always on' dialogue with political stakeholders as opposed to the traditional top-down paradigm. The potential for criticism is huge, the capacity to control limited. And this is still relatively new territory.

Conclusion

In the absence of transparent, structured social media strategies, as part of a piece of coherent government policy and implemented at institutional level, the successful use of social media in disaster management will be severely restricted.

As such, the power to harness the real benefits of social media will continue to be undermined if the societal impact of government constructs around social media do not understand the important dynamics of such a trust relationship.

A new paradigm of engagement, one that is rooted in trust, transparency and open dialogue, must be engendered before social media can fully realise its potential as a robust, reliable, effective communications component within crisis management strategies during periods of instability.

REFERENCES

[i] http://www.emergency-management-degree.org/crisis/

[ii] Research Paper Series, Number 8, Fuchs and Trottier, 2013, research conducted in the project "PACT - Public Perception of Security and Privacy: Assessing Knowledge, Collecting Evidence, Translating Research into Action", funded by EU FP7 SECURITY, grant agreement no. 285635

iii Omand D., Bartlett J., Miller C. , "Introducing Social Media Intelligence (SOCMINT)," Intelligence & National Security, 27, no. 6 (2012): 809, accessed March 12, 2013, http://search.proquest.com/docview/1239085575?accountid=322

Terminological Ontologies for Risk and Vulnerability Analysis

Bodil Nistrup Madsen, Hanne Erdman Thomsen
Copenhagen Business School
Dalgas Have 15, 2000 Frederiksberg, Denmark
E-mail: bnm.ibc@cbs.dk, het.ibc@cbs.dk

Abstract

Risk and vulnerability analyses are an important preliminary stage in civil contingency planning. The Danish Emergency Management Agency has developed a generic model and a set of tools that may be used in the preparedness planning, i.e. for identifying and describing society's critical functions, for formulating threat scenarios and for assessing consequences. Terminological ontologies, which are systems of domain specific concepts comprising concept relations and characteristics, are useful, both when describing the central concepts of risk and vulnerability analysis (meta concepts), and for further structuring and enriching the taxonomies of society's critical functions and threats, which form an important part of the model. Creating terminological ontologies is a time consuming work, and therefore there is a need for automatic tools for extraction of terms, concept relations and characteristics. Terminological ontologies must adhere to a number of constraints, and therefore tools for automatic validation of these ontologies are also needed. Methods and tools for automatic ontology construction are being developed by researchers at Copenhagen Business School. The tools developed may also be used for extracting information on disasters from various media, and terminological ontologies may be used for enhancement of retrieval of information about disasters and for choosing the relevant countermeasures.

Keywords: Terminological Ontology, Automatic Knowledge Extraction, Disaster Management

1. Introduction

The Danish Emergency Management Agency (DEMA) has developed a generic Model for Risk and Vulnerability Analysis, the RVA Model (DEMA 2006). In the introduction to the user guide it is stated that: "It will never be possible to completely prevent large-scale disturbances, accidents and catastrophes, but they can be countermanded using timely and comprehensive preparedness planning. The ability to uphold and continue society's critical functions will thereby be improved. Risk and vulnerability analyses are an important preliminary stage in such preparedness planning."

The RVA model is designed for team-based analysis work, and DEMA (2006) emphasizes that "before starting the analysis work, it is important that the working group thoroughly discusses the methodology and terminology. What may appear, on the surface, to be purely theoretical differences between terms like "threat", "risk" and "vulnerability" can, in practice, have great impact on the conclusions of the analysis and thus on the working group's proposals for specific countermeasures."

Here terminological ontologies come into the picture. A terminological ontology is a domain specific ontology, cf. for example the categorisation of ontologies by (Guarino 1998) and (Madsen & Thomsen 2008). We use the term terminological ontology as synonym to the term concept system, which is normally used in terminology work, cf. for example (ISO 704 2000).

Concept clarification based on the principles of terminological ontologies has proven very successful in many kinds of terminology work and in many domains. Building terminological ontologies renders a solid foundation for reaching consensus about definitions of concepts, e.g. in a standardization process.

First we will introduce the basic principles of terminological ontologies and briefly describe DEMA's RVA Model.

Then we will give examples of the usefulness of terminological ontologies for clarification of central concepts, i.e. the meta-language used in risk and vulnerability analysis, and illustrate how terminological ontologies may also be useful when describing society's critical functions and the categories of threats.

Developing terminological ontologies is very time consuming, and therefore there is a need for automatic ontology construction based on various kinds of input: internet text and text files but also structured data, taxonomies, etc. We will briefly mention ongoing work on automatic ontology construction.

Finally we will mention some perspectives for the use of ontologies in disaster management and disaster recovery.

2. Terminological Ontologies

One of the main threat categories in the RVA model is *Destruction, interruption or other failure of society's critical functions*. It has the sub-category: *Water* with examples: *Drink water supply, transport and treatment of waste water,* cf. Appendix A.

As an example of a terminological ontology we have chosen a simplified part of a terminological ontology of *wastewater treatment plants*, cf. Figure 1[1]. Each coloured box corresponds to a concept represented by the preferred term, e.g. *chemical treatment plant*. Lines between concepts correspond to type relations (also known as ISA relations). Other relations (part-whole, temporal and associative relations) may be represented with different line types, as illustrated in Figure 5.

[1] The terminological ontology was developed by using the terminology and knowledge management system i-Term, developed by the DANTERM Centre of Copenhagen Business School, www.iterm.dk.

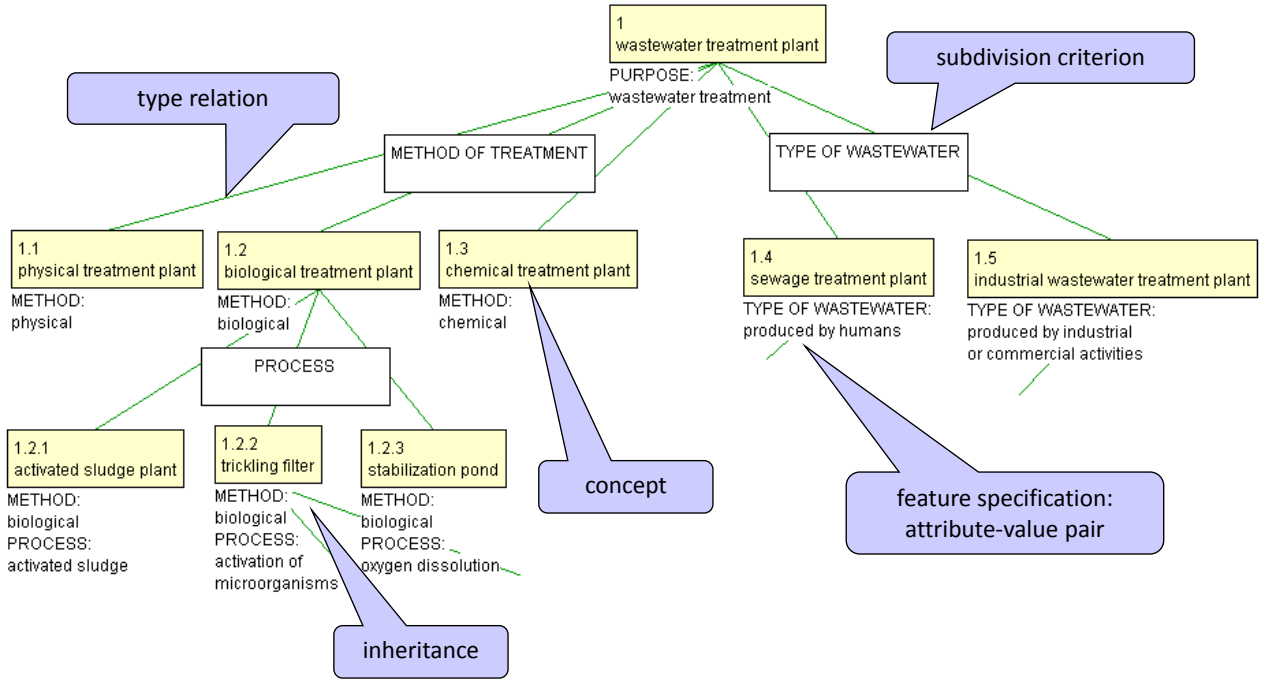


Figure 1: Draft terminological ontology for wastewater treatment plants

The characteristics of the concepts are represented below the boxes as feature specifications in the form of attribute-value pairs, e.g. *METHOD: biological*. On the basis of these feature specifications, subdivision criteria are introduced, i.e. the white boxes with text in capital letters. Subdivision criteria give a good overview and help the terminologist in writing consistent definitions.

The concept *trickling filter* is a type of *biological treatment plant*, and it inherits the characteristic *METHOD: biological* from this super-ordinate concept. The characteristic: *PROCESS: activation of microorganisms* differentiates *trickling filter* from the co-ordinate concepts *activated sludge plant* and *stabilization pond*. When clicking the concept in the ontology, it is possible to see e.g. definitions, examples, pictures and information in other languages.

The use of feature specifications is subject to a number of principles and constraints, some of which are described in detail by (Madsen, Thomsen, & Vikner, 2004).

As a background for the examples of terminological ontologies of central concepts used in risk and vulnerability analysis, as well as concepts that are relevant when describing society's critical functions and the categories of threats, we will introduce some elements of DEMA's RVA Model. The description is to a large extent directly based on DEMA's publications, e.g. DEMA (2005; 2006) and information on DEMA's web-site, http://brs.dk/eng/.

3. DEMA's Generic Model for Risk and Vulnerability Analysis – the RVA Model

According to the Danish preparedness act, the individual ministers are responsible for planning for the continuity of 'society's functions', each within their respective areas. However, the act does not specify exactly what these functions are. The Danish National Vulnerability Evaluation – an inter-departmental, cross-sector evaluation from 2004 – aimed to make systematic use of risk and vulnerability analysis an integrated part of the wider civil contingency planning responsibilities of central government institutions in Denmark, and one of the 33 specific recommendations in this evaluation was that a generic risk and vulnerability analysis model should be developed for civil contingency planning.

This project was assigned to DEMA, and the model was completed in late 2005. Consequently specific models for risk and/or vulnerability analysis were planned among local fire and rescue services, harbour authorities, electricity and natural gas suppliers, the Danish central bank, the police Security Intelligence Service, the National Centre for Biological Defence, and the National IT-and Telecom Agency (now the Danish Agency for Digitisation).

DEMA's model, the RVA Model, focuses on the need for continuity of 'critical functions' in case of large-scale disturbances, accidents or outright catastrophes. By critical functions the model refers to activities and services that are indispensable for society. Their importance is such, that any entire or partial loss could have grave consequences for life, health, property, or the environment.

The RVA model is divided into four parts:

Part 1 - Starting point for the analysis

Part 2 – Identification of threats

Part 3 – Analysis of threat scenarios

Part 4 – Risk and vulnerability profile.

To support the analysis, a template for each part has been prepared in Microsoft Word. During the analysis, one document is to be filled in for parts 1 and 4, respectively. For parts 2 and 3, one document is filled in for each threat scenario included in the analysis.

In Part 1, the user is asked to briefly describe the critical functions, c.f. Figure 2, and is referred to *Appendix A* in the user guide: *Overview of society's critical functions*, which is found in Appendix A.

In Part 2, the user must select a threat category for the formulation of threat scenarios by using a pick list with the main categories, see Figure 3. For a more detailed description of the scenarios, the user will need the sub-categories found in Appendix B. This may pose a problem, since the individual sub-categories are not defined, and it may be difficult to differentiate them, c.f. for example the category *Atmospheric threats* in Appendix B.

In Part 3, the user must assess the possible consequences of the type of incident for his or her particular organisation/area of responsibility and for society in general, c.f. Figure 4. In both cases, the following index is used: 1 = Limited, 2 = Moderate, 3 = Serious, 4 = Severe, 5 = Critical. The difference between e.g. '3 Serious' and '4 Severe' is not defined.

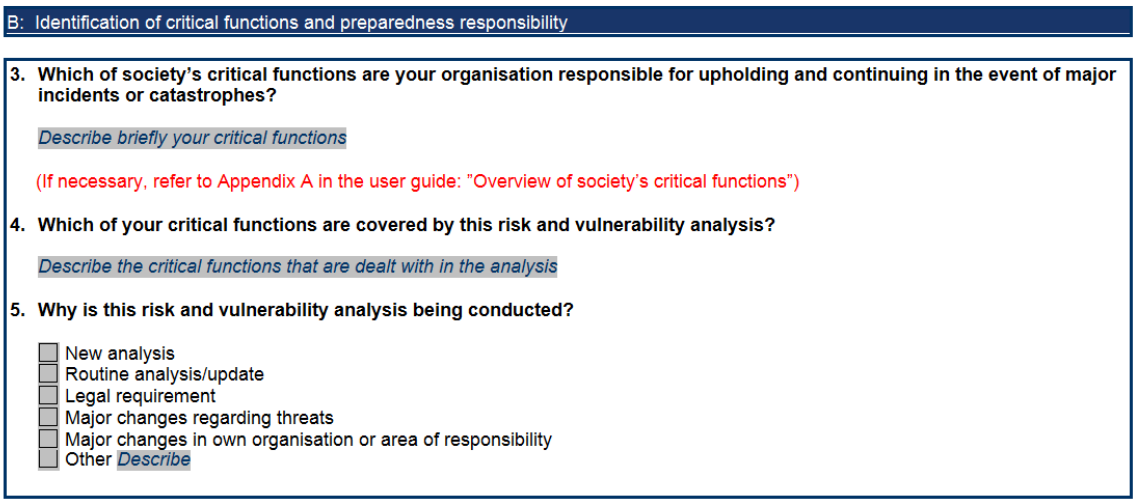


Figure 2: The RVA Model - Identification of critical functions and preparedness responsibility

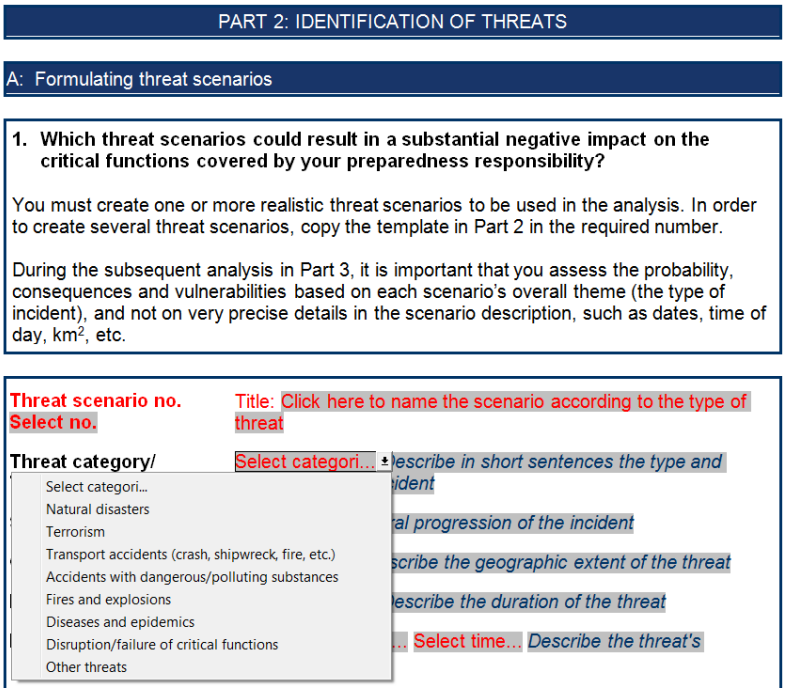


Figure 3: The RVA Model – Identification of threats

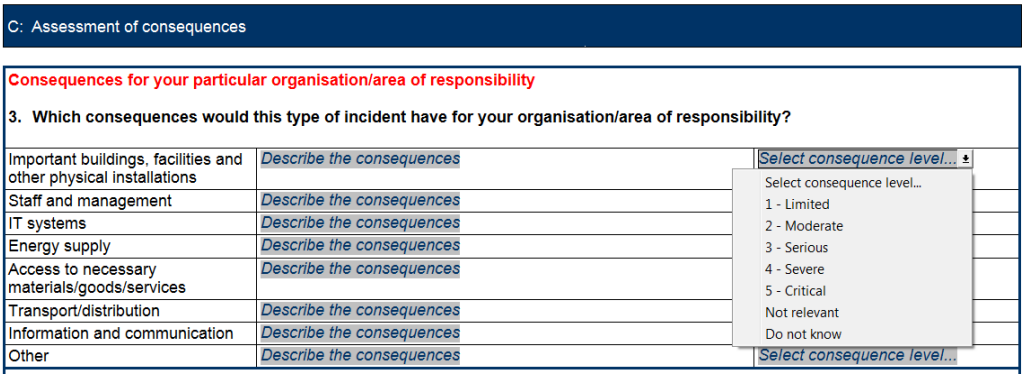


Figure 4: The RVA Model – Analysis of threat scenario: Assessment of consequences

4. Terminological Ontology for Central Concepts of the RVA Model

DEMA (2006) comprises explanations of central concepts related to risk and vulnerability analysis. Developing a terminological ontology which clearly shows the relations between these concepts and for each concept gives a characteristic in the form of a feature specification, furthers the elaboration of brief and consistent definitions, and thus provides a good basis for a common understanding and use of the concepts in question.

Figure 5 shows a manually constructed draft of a terminological ontology comprising some of the central concepts related to risk and vulnerability analysis. By clicking a concept in the ontology, the user can see information registered on the concept in the underlying term bank, such as synonyms, definition, comments, and – if a parallel ontology is developed in e.g. Danish – equivalent information in Danish.

A proposal for the definition of the concept *probability assessment* written on the basis of the terminological ontology could be: *part of risk analysis which has the purpose of assessing the probability of an undesired incident*, which is clearer that the description in DEMA (2006): *Probability and consequences can be assessed either quantitatively or qualitatively. Consequence assessment refers to the scope, extent and duration of loss and damage to life, health, critical functions, property, environment or other assets. Probability assessment can be carried out using indicators of likelihood (i.e. frequency) or "plausibility" (i.e. qualified guesses). The latter is usually the case with respect to threats caused by humans that are difficult or impossible to predict with any accuracy.* More specific information could be entered in the term bank as a supplement to the definition.

5. Terminological Ontologies for Categories of Threats

As mentioned above, the RVA Model Appendix B comprises a catalogue (taxonomy) of threats. One of the categories of the first main category *Extreme natural phenomena* is the category *Atmospheric threats*. Sub-categories (Examples) are: *hurricane, cyclone, tornado, blizzard, ice winter, glazed ice, dense fog, cloudbursts, hail storms, lightning*. Here a terminological ontology would help the users of the RVA Model, since it would provide information not found in the taxonomy, such as characteristics and definitions of the sub-categories to explain the differences between the various types of atmospheric threats. It would also constitute a knowledge base of great value for new employees in DEMA or new collaborators. The terminological ontology can be enriched with concepts on countermeasures.

Such an ontology can also be employed in information retrieval in cases of emergency, c.f. section 7 below and for decision on relevant countermeasures. The process of developing a terminological ontology in itself helps clarify concepts in the domain and leads to identification of sub-categories originally left out, thus ensuring completeness.

6. Automatic Ontology Construction

It is imperative that a term bank covers the terminology of a given domain completely and that the stored information is of a high quality. If a term bank does not contain a sufficient number of terms, or if it contains a large amount of terms with only little or poor quality information attached, users will tend not to use it. One way of increasing the amount of terms in a term bank and to ensure the quality of information related to the terms is to automatically extract terms and information about terms from texts and to automatically construct and validate terminological ontologies.

Here we will briefly introduce on-going work carried out in the framework of the DanTermBank research project, which aims at establishing the foundations for a National Terminology and Knowledge Bank, c.f. (Madsen et al., 2010). In this project, prototypes for automatic knowledge extraction, ontology construction and validation are under development.

6.1 Automatic Knowledge Extraction

The first prototype is a web crawler for domain corpus compilation, which, on the basis of exemplary texts, predefined seed lists of terms, statistical measures and set operations on the results, collects domain texts from the internet.

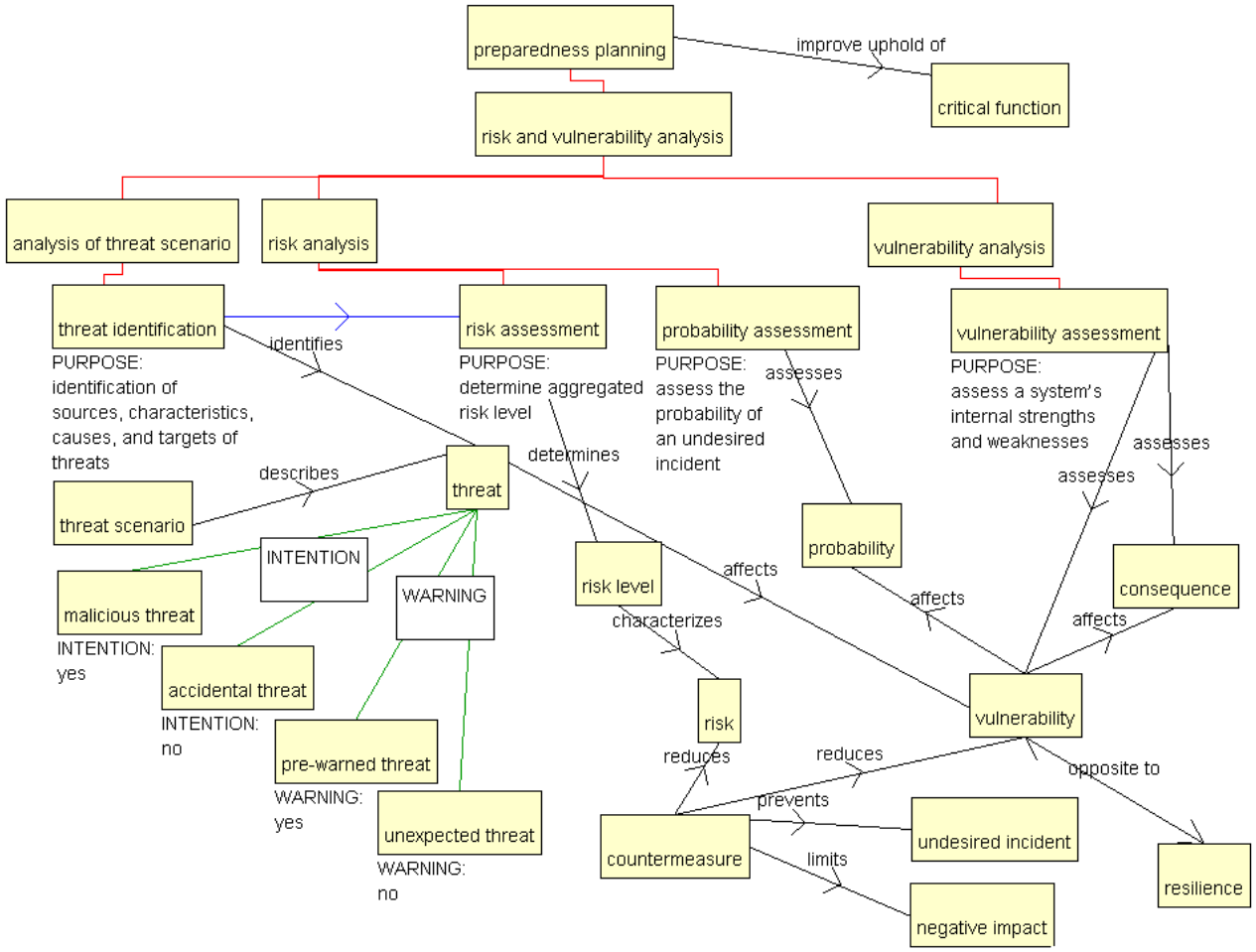


Figure 5: Draft terminological ontology of some central concepts related to risk and vulnerability analysis

The second prototype is a POS tagger, i.e. the TreeTagger (Schmid 1994, 1995) trained for Danish in the DanTermBank project.

The third prototype is a term extractor which takes as input the POS-tagged corpus files and outputs term candidates on the basis of syntactic and morphosyntactic patterns and statistic measures. Morphosyntactic patterns match closed compounds, e.g. *sårbarhedsanalyse-vejledning* (directions for vulnerability analysis), without blanks. This type of compounding is highly productive in Danish. At present, in our prototype, we apply co-occurrence scores, e.g. Pointwise Mutual Information (Church & Hanks, 1993) and Dice coefficient (Smadja, 1993), as well as 'termhood' scores, e.g. Log Odds Ratio (cf. e.g. Everitt, 1992) and weirdness (Ahmad et al., 1999)'.

The fourth prototype, a relation extractor, is currently under development. Its purpose is to identify concept relations in the corpus and produce draft ontologies as output. A boot-strapping approach is foreseen, where term pairs between which a known relation exists are used as input to generate knowledge patterns (cf. e.g. Ahmad & Fulford, 1992; Halskov & Barriere, 2008; Hearst, 1992; Meyer, 2001), expressing the given relation. Subsequently the corpus is searched for occurrences of the patterns to identify relations holding between known concepts (e.g. identified by the term extractor) or between a known concept and a potential new concept.

A detailed description of the extraction prototypes is found in (Lassen, 2012).

6.2 Automatic Ontology Validation

The fifth prototype is an ontology validator. The extraction of concept relations results in draft ontologies, which are unlikely to comply with the principles and constraints that apply to terminological ontologies and therefore validation is needed. The current prototype identifies violations of principles, later versions will also propose how violations can be remedied. For example, if the draft ontology contains two concepts that have been placed in a direct type relation, but where the feature specifications imply that a concept should in fact exist between them, the system can introduce a new concept in order to make the ontology valid. At a later stage, domain experts must finalise the ontology and fill in terms for automatically created concepts, c.f. (Lassen, Madsen & Thomsen, 2011).

7. Perspectives for the Use of Ontologies in Disaster Management and Recovery

The knowledge extraction tools can be used for surveillance based on information extraction from social

and formal media. Information extracted in this way can then be mapped to already validated terminological ontologies by employing techniques for ontology-based information retrieval c.f. (Andreasen et al. 2004, 2011). Such an approach would allow immediate links to potential risks and relevant countermeasures.

8. Acknowledgements

The DanTermBank project is funded by the VELUX FOUNDATION.

9. References

Ahmad, K. & Fulford, H. (1992). *Knowledge Processing: 4. Semantic Relations and their Use in Elaborating Terminology.* (Computing Sciences Report). Surrey: University of Surrey.

Ahmad, K., Gillam, L., Tostevin, L. (1999). University of Surrey participation in TREC8: Weirdnessindexing for logical document extrapolation and retrieval (WILDER). In: *Anonymous The Eighth Text REtrieval Conference (TREC-8)*.

Andreasen, T., Jensen, P.A., Nilsson, J. F., Paggio, P., Pedersen, B.S. & Thomsen, H.E. (2004). Content-based Text Querying with Ontological Descriptors, in *Data & Knowledge Engineering 48 (2004)*, Elsevier, pp. 199-219.

Andreasen, T., Bulskov, H., Lassen, T., Zambach, S., Jensen, P. A. Thomsen, H. E., Madsen, B. N. & Nilsson, J.F., A Semantics-Based Approach to Retrieving Biomedical Information. In: *Lecture Notes in Computer Science, Nr. 7022, 2011*, pp. 108-118.

Church, K.W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16(1)*, pp. 22-29.

DEMA (2005). *DEMA's Approach to Risk and Vulnerability Analysis for Civil Contingency Planning*. Danish Emergency Management Agency.

DEMA (2006). *Introduction and User Guide - DEMA's Model for Risk and Vulnerability Analysis.* Danish Emergency Management Agency.

Everitt, B. (1992). The analysis of contingency tables. *Chapman & Hall/CRC Mono-graphs on Statistics & Applied Probability), 2nd edition*.

Guarino, N. (1998). Formal Ontology and Information Systems. In: *Formal Ontology in Information Systems*, Proceedings of the First International Conference (FOIS'98), June 6-8, Trento, Italy. Ed. Nicola Guarino. Amsterdam: IOS Press, pp. 3-15.

Halskov, J., & Barriere, C. (2008). Web-Based Extraction of Semantic Relation Instances for Terminology Work. *Terminology, 14*, pp. 20-44.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of COLING-92*, pp. 53-545.

Helmut, S. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing,* Manchester, UK.

Helmut S. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

ISO 704. (2000). *Terminology work — Principles and methods*. Genève: ISO.

Lassen, T. (2012). A Corpus Compilation and Processing Prototype for Terminology Work. In: Aguado de Cea et al. (Eds.): *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012),* 19-22 June 2012, Madrid, Spain, pp.218-230

Lassen, T., Madsen B.N. & Thomsen, H.T. (2011). Automatic Knowledge Extraction and Knowledge Structuring for a National Term Bank. In: *NODALIDA 2011 workshop: Creation, Harmonization and Application of Terminology Resources*. NEALT Proceedings Series Vol. 12. http://hdl.handle.net/10062/17274 (2011-05-09), pp 23-26.

Madsen, B.N. & Thomsen, H.E. (2008). Terminological Principles used for Ontologies. In: Madsen, B.N. and H.E. Thomsen (eds.): *Managing Ontologies and Lexical Resources.* Litera. ISBN: 87-91242-50-9, pp. 107-22.

Madsen, B.N., Thomsen, H.E., Halskov, J. & Lassen, T. (2010). Automatic Ontology Construction for a National Term Bank. In: Úna Bhreathnach & Fionnuala de Barra Cusack(eds.): *Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges.* Nicolson & Bass, Ireland. ISBN:978-0-9566314-0-4, pp.502-532.

Madsen, B.N., Thomsen H.E. & Vikner, C. (2004). Principles of a system for terminological concept modelling. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Vol. I. Lisbon, pp.15-18.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. In: D. Bourigault, C.J. and M.-C. L'Homme (eds.) *Recent Advances in Computational Terminology,* John Benjamins, Amsterdam/Philadelphia, pp. 279-302.

Smadja, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics, 19,* pp. 143-177.

10. Appendices

Appendix A. Overview of society's critical functions

"Society's critical functions" denotes those activities, goods and services that comprise the basis for the ability of society to function, and, therefore, must be upheld and continued during major accidents or catastrophes.

Sector	Critical functions	Sector	Critical functions
Energy	▫ Electricity supply ▫ Gas supply ▫ Oil and petrol supply	**Preparedness**	▫ Sending and receiving alarms and warnings ▫ Policing tasks ▫ Fire fighting ▫ Ambulance services and other pre-hospital tasks ▫ Rescue operations (land/sea/air) ▫ Evacuation, reception, accommodation and catering ▫ Chemical preparedness ▫ Biological preparedness ▫ Radiological preparedness ▫ Nuclear preparedness ▫ Ammunitions clearing ▫ Storm surge preparedness ▫ Environmental preparedness ▫ Military aid for civil authorities
Communication and IT	▫ Landline telephony ▫ Mobile telephony ▫ Data processing and data transmission ▫ Information networks ▫ Internet access ▫ TV, satellite and radio transmission ▫ Navigation ▫ Post and courier services		
Transport	▫ Management, monitoring and control of traffic and goods transport (road, rail, air and sea) ▫ Surveillance and control of infrastructure (bridges, tunnels, airports, stations, ports, etc.)		
Finance and economy	▫ Payments and money transfers ▫ Banking and insurance ▫ Securities ▫ Central bank functions	**Health**	▫ Primary health services ▫ Hospital services ▫ Care of vulnerable people ▫ Monitoring infectious diseases ▫ Medications preparedness ▫ Medications production
Foodstuffs	▫ Food supply ▫ Monitoring food safety ▫ Monitoring infectious livestock diseases	**Public administration**	▫ Crisis management capacity ▫ Upholding and exercising the authority of parliament, the government and central administration, the courts of law and the municipalities.
Water	▫ Drinking water supply ▫ Transport and treatment of waste water	**National security**	▫ Guarding and surveillance of key points and borders. ▫ Military defence and enforcement of sovereignty ▫ Intelligence tasks ▫ Counter-terrorism ▫ Bodyguard services
Dangerous substances	▫ Control of production, storage, and transport of dangerous substances (chemical, biological, radiological, nuclear)		

Appendix B. Catalogue of threats

Threat category/type	*Examples*
Extreme natural phenomena	
Atmospheric threats	Hurricane, cyclone, tornado, blizzard, ice winter, glazed ice, dense fog, cloudbursts, hail storms, lightning
Geological threats	Earthquake, volcanic eruption, avalanche, landslide
Oceanographic threats	Tsunami, storm surge, flooding, sea ice
Terrorism	
Terrorist actions against authorities, critical infrastructure assets, employees, the wider population, etc.	Conventional weapons, CBRN weapons (chemical, biological, radiological and nuclear) cyber terrorism
Terrorist actions abroad	Conventional weapons, CBRN weapons, cyber terrorism
Transport accidents (crash, shipwreck, fire, explosion,)	
Sea	Passenger ships, bulk, container and tank ships, military vessels
Air	Passenger planes, freight planes, military planes
Railway	Passenger trains, goods trains
Road	Cars, busses, trucks, lorries
Accidents with or emissions of hazardous/polluting substances	
Chemical substances	Chemicals, gas, oil and oil products, petrol, toxins
Biological substances	Bacteria, virus, toxins
Radiological and nuclear substances	Radioactive radiation
Explosives	Explosives, fireworks, ammunition
Fires and explosions	
Buildings/areas with many people	High buildings, shopping malls, theatres, cinemas, discotheques, sports halls, stadiums, conference centres, hotels, nursing homes, hospitals, prisons, institutions, offices, festivals, markets
Industry (production, distribution, storage, etc)	"Seveso companies", environmentally/fire hazardous operations, storage of inflammable/explosive substances
Infrastructure	Railway stations, airports, tunnels, ports
Countryside	Forests, moor land, fields
Cultural assets	Castles, museums, preserved buildings, churches, old town areas

Ontology and Terminology of Disaster Management

Xiubo Zhang, Khurshid Ahmad
Trinity College Dublin
Dublin, Republic of Ireland
xizhang@scss.tcd.ie, kahmad@scss.tcd.ie

Abstract

The spate of natural disasters in the USA and in the European Union, especially floods and hurricanes, has led to the creation of specialised government agencies for dealing with such disasters in a unified command and control mode. Disaster management involves above all an inter-agency communication strategy that is at once transparent and deals in a timely manner with a set of unfolding events with the highest degree of professional care. Experts in various branches of engineering have to work together experts in health, in logistics, and in civil administration. Each has its own terminology generated from discipline-specific experiential knowledge. This experiential knowledge is highly codified for pedagogical purposes yet it is not accessible for the purposes of building information systems. We outline a corpus based method for building the ontology and terminology of natural disasters that relies in part on a legacy glossary of specific disasters and the structures and mnemonics used to encode the glossary are regarded as prototypical ontology of the disaster domain. Our methods and techniques were developed for looking at emerging specialised sciences, nanotechnology, breast cancer treatment, and more recently the inter-agency response to the 2008-financial debacle.

Keywords: Disaster Management, Ontology Learning, Text Analysis

1. Introduction

Disaster management (DM) has been carried out for almost as long as the human beings have existed, however, the agencement – an almost infinite number of connections between machines and people, provided by ubiquitous IT systems that now exist and will exist in more sophisticated forms in times to come, has transformed the subject. The agencement means that different stakeholders in DM can send and receive life- and mission-critical information to each other. Each of the stakeholders, including first responders, disaster victims and their wider community, business continuity experts, has a different vocabulary of what exists in the disaster impacted environment and in their own enterprises. Place names may be deprecated making locating victims and damaged infrastructure difficult; names of life saving objects may be spelt differently or different stakeholders may name the same object with a broad or narrow term. The inter-relationship between different objects, for instance, cause-effect, container-contained, superordinate-subordinate relationship may not be clearly understood or articulated.

The organisation of terms, standardisation of names and relationships, is of critical import for effective and transparent communication. Terminology standardisation work is carried out by standards organisations. However, each new domain of knowledge evolves its own terms and upon achieving a degree of success the domain knowledge is archived by standards organisations. Each domain produces a description of the conceptual organisation of the terminology within its area – first almost by word of mouth. Then, over a period of time, after achieving a consensus within the domain, formalisation of the conceptual organisation takes place in the form of a thesaural dictionary or encyclopaedic document. It should be remembered here that each term, its definition, and its relationship with others, is first documented in a text of the domain. The construction of the conceptual organisation of a domain sometimes can be initiated by examining a diachronically organised collection of texts in the domain. This is our preferred method of constructing conceptual organisation for building information systems for a given domain.

The text based method is motivated, in part at least, by the notion of agencement: The agencement makes this process of terminology standardisation quite difficult and developers of information systems rely on protocols and methods within computer science to capture the conceptual organisation of a domain. The term increasingly used in this context is ontology.

There are a number of definitions of the term ontology in computing literature. The term has been borrowed from philosophy and these definitions show that a consensus is yet to be reached on what the definition of the term is. It has fallen onto a philosopher, with a number of key papers in computational ontology, especially of plant and social reality ontology, Brian Smith, to provide a concise definition of the subject of ontology: "Ontology as a branch of philosophy is the science of what is, of the kinds and structures of objects, properties, events, processes, and relations in every area of reality." (Smith, 2004). Given the considerable amount of successful intellectual and practical strides in the area of plant ontology and gene ontology, both dealing with animate and inanimate objects that are evolving in time and within an information-rich environment, perhaps, it is useful to start the discussion of disaster ontology in the context of ontology development with reference to plant and animal ontology.

The purpose of plant ontology is to "develop and maintain a controlled, structured vocabulary ('ontology') of terms to describe plant anatomy, morphology and the stages of plant development" (Cooper et al., 2013). The plant ontology draws upon the work of gene ontology and attempts "to link (annotate) gene expression and phenotype data to plant structures and stages of plant development, using the data model adopted by the Gene Ontology" (ibid).

Given that disaster management is an evolving subject involving animate and inanimate objects, we report on some work on creating an ontology of this domain through an examination of news and reports that focus on disasters. We begin by a survey of current work in the terminology and ontology of disaster management. We then outline a text-based method for identifying single and multi-word terminology of the domain leading onto the identification of the conceptual structure of our domain within the limitations of the texts. A case study is then presented and future work is outlined.

2. Motivation

Published texts provide a snapshot of key events, important people, places and objects, for the duration of the snapshot. More importantly for us is the notion that in order to understand the meaning of the building blocks of a text – words, especially terms specific to the events, people, places and objects – one should first see how others have used the words. Modern dictionaries, especially learners' dictionaries comprise examples of usage drawn from a systematically organised collection of carefully selected texts drawn from a variety of texts, called a representative corpus of texts (Ahmad, 2007). Disaster events, and the news flow associated with such events, that is the number of news and learned items comprising key terms, appear to leave their signature on a systematic organised text corpora related to the disaster. This can be demonstrated in an examination of our own 10 million words corpus of news reports (published between 1987-2014), and looking at frequency of keywords related to disaster – terms like tidal waves, tsunami, flooding and so on – together with the total number of news reports published on a given day comprising one or more of the keywords, we see that this daily time-series of keywords and number of news reports per day, correlates well with major hurricanes in North America and Indian Ocean (see Figure 1).

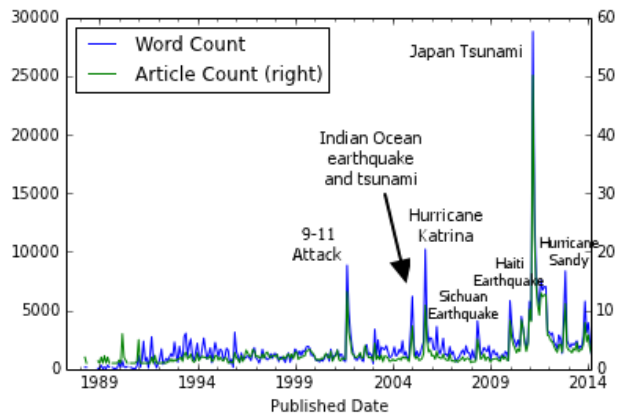


Figure 1: Time series of monthly occurrences of disaster news flow over a 27 year period.

2.1. Ontology and Disaster Management

Ontology has many uses in disaster management: At a practical level of managing information, that is enabling the various stakeholders in a disaster event to search and send important information to each other avoiding ambiguities as much as possible. The place names can cause much confusion during a disaster event when people use synonyms and deprecated names to describe a place instead of an officially designated name: an ontological description will have facilities to map the heterogeneous sets of names onto the designated name to lower the ambiguity of description if one existed. Some authors have argued that "geographic information is the key to effective planning and decision-making in a variety of application domains – semantic heterogeneity can be avoided by a well-constructed ontology" (Klien et al., 2006). Still in the practical vein, it has been suggested that the design of the disaster management websites should take cognisance of the disaster management phases, essentially an ontology comprising "five phases in disaster management: signal detection, preparation/prevention, containment/ damage limitation, recovery, and learning" (Chou et al., 2011). The authors have selected 100 disaster management websites and examined over 6,000 web pages to study the organisational principles, and methods and techniques of web site design and implementation. They have developed an illustrative ontology of 'supplies kit preparation' for disaster management after having looked at the websites of US FEMA and the International Red Cross and other major agencies (2011:57) (See Figure 2).

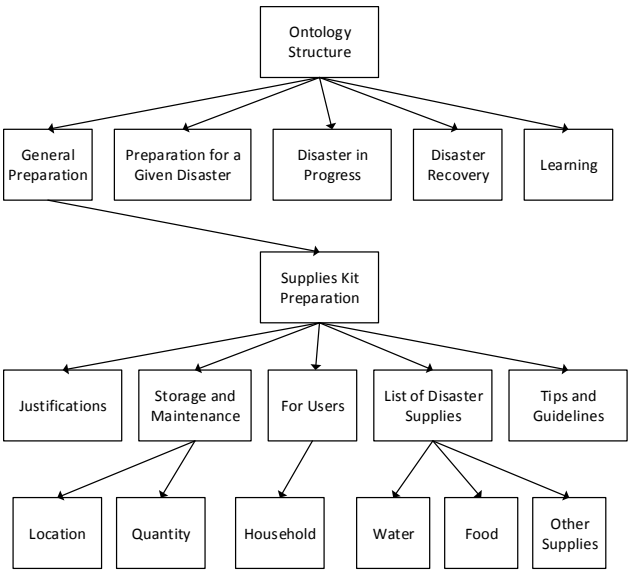


Figure 2: A merged ontology of 'supplies kit preparation' based on a synthesis of web sites belonging to major disaster management and disaster preparedness agencies developed by Chou et al. (2011)

Information sharing in the geological hazards domains can be mediated and facilitated by using an ontological approach and through the use of special architecture designed for information sharing (Wang and Lei, 2013). Documents released by disaster managers contain critical information about the status of transportation and energy infrastructure and literally hundreds of documents can be released in a short space of time. The stakeholders in a disaster event have to search for this information – here it has been sug-

gested that an ontology based search can help in the retrieval of key documents with greater accuracy. Wu et al. (2013) looked at press releases produced by the Miami Dade County and the Department of Homeland Security during Oct/Nov 2005 in the midst of Hurricane Wilma. The documents were classified according to an ontology of transportation objects that could be impacted by a hurricane including air, rail, bus and other services; it was found that an ontology based search improved retrieval of relevant documents by more than 50%.

At a policy level especially ontological descriptions have been used to critique the policy disaster preparedness and disaster management agencies. Martin and Simon (2008) have argued that the US DHS, which together with US FEMA, is the principal agency for disaster preparedness and disaster mitigation, uses a so-called virtual ontology of (ever present) threat and that has implications of security of people, places, infrastructure and other objects of value to a people. The authors expressed the opinion that "the everyday, emerging circulations of goods and people, present DHS with a terrain of shifting threats from which both emergencies and preparedness may materialise.".

Core data: The WWW Consortium is developing an ontology of disaster management and seek to distinguish the enterprise of disaster management from that of the management of unplanned accidents – commonly referred to as emergency management . We will refer to this the WWW enterprise as W3DM. The argument of W3DM is that it whilst it is true that there are different types of disaster, and there inevitably different systems used by different types of organization using different sorts of data, it is a possibility that a common set of data can be used across different disaster scenarios by different agencies. W3DM is critical of the current the Sahana system, which has been proposed as an experience-based conceptual basis for disaster management and is not too focussed on standardisation and "is extremely poorly integrated with existing W3 standards".

2.2. Review of Disaster and Risk Terminology/Ontology Resource

2.2.1. Glossaries, Terminologies and Disaster Management

A number of national, federal and trans-national organisations collate, analyse, and disseminate terminology of specialist subjects for information dissemination in whole range of human activities – there are terminology data banks for research, development, teaching and learning. A terminology data base is typically implemented using conventional data base architecture and each term has a set of attributes (spelling, definitions, deprecated terms and links to other terms) and each attribute may have one or more values. Recently, the world of digital repositories uses (a subset of) terminology data banks , that is a glossary of subject terms created for indexing documents in a repository such that these terms will match those of query terms used by end-users.

The disaster management community, within broader environmental communities and within risk, security and safety communities, has been collating terminology of related disciplines for use for communicating information critical for disaster management. There are terminologies/glossaries are highly relevant to the topic of disaster and risk management while others concern themselves more with environmental hazards There are a number of terminology data banks and many of these are available online: the main characteristic of these English-dominated data bases is that definitions are provided online and the terms are presented in an alphabetical order. The UN Office of Disaster Risk Reduction has produced "basic definitions [of terms] on disaster risk reduction to promote a common understanding on the subject for use by the public, authorities and practitioners." And the International Panel on Climate Change has a related glossary on climate change .The Italian Protezione Civile was similarly motivated to produce an Italian glossary of terms with English translation. (See, for example, Table 1).

Consider the GEneral Multilingual Environmental Thesaurus (GEMET) created by Eionet. Eionetis a partnership network of the European Environment Agency (EEA) and its member and cooperating countries. The purpose of this network is to ensure access to high-quality environmental data to the EEA members and their associates. The partners are in 32 countries and to ensure smoother communication across linguistic barriers the Eionet has created a term bank of 4000 terms covering 22 subject topics. An entry term in the Eionet term bank has a broader and narrower term, a definition and a textual source link that may provide further elaborations of the entry term. The Eionet subject topics comprise the traditional environmental subjects like agriculture, climate, ecosystems, energy, environmental policy, noise, radiation, transportation and water; and related topics like economics, geography, human health, and social aspects. The thesaurus has a topic called disasters, accidents, and risk. Consider, for example, the entry term *disaster contingency plan* (DCP) and its elaborations within the term bank. The definition is given as "an anticipatory emergency plan to be followed in an expected or eventual disaster, based on risk assessment, availability of human and material resources, community preparedness, local and international response capability, etc." The Eionet term bank suggests that DCP is a narrower term (subordinate concept) for the term *environmental contingency planning*. The broader term (or the super-ordinate concept) is safety measure, and that some of the attributes of DCP are shared with the term (concept) plan.

The organisation of terms in a structure that relates superordinate terms to subordinate terms is generally a first step in constructing an ontology. The relationship between planning and safety is annotated manually in terminology databanks. A formally constructed ontology system will use logical relationship to express hierarchies, part-whole, and co-ordinations between domain object, e.g. disaster, safety, planning. Note also that for further elaboration of terminology the terminology databanks employ hyperlinks to key documents where a term in question will be elaborated by the context of their usage.

We are interested in using these legacy terminology data bases as a key input to the construction of the ontology of a domain. We look at a collection of domain texts as a source of terms and show how automatic term extraction will be

Originator	Domains	Language	Terminology Coverage
Italian Civil Protection Department	Risks & emergency management	Italian	c. 260
UN Office for Disaster Risk Reduction (UNISDR)	Disaster risk reduction	6 UN languages	54
International Panel on Climate Change	Risk management of extreme events w.r.t. climate change (SREX)	English	
World Association for Disaster & Emergency Medicine	Risk & disaster management w.r.t medicine	English	200
CNR Italia	Natural & technological risks in the Environment	English/Italian	15000
GEMET (EEA)	Environmental concerns	32 EU languages	4000
Management of a Crisis Vocabulary (MOAC)	Geospatial & related crisis management	English	

Table 1: Rapisardi and Franco (2004): Disaster Resilience and the Babel of Semantic

helpful in complementing the information in legacy data bases.

3. Methods

Most modern ontology learning systems generally divide the task of learning ontologies from text into two subtasks, namely the learning of significant domain terminologies/concepts and the learning of the relationships/axioms between these concepts. In this paper, we present CiCui, a text analysis system which could assist the formulation of domain ontologies by providing knowledge engineers with insights such as terminologies and relationship predicates gained from the automatic analysis of large bodies of domain specific text.

Using this system, we purpose a method which can automatically learn from unstructured text: *1*) the important domain terminologies/concepts by exploiting various text analytic metrics and linguistic patterns, and, *2*) potential relationships between the learnt concepts by examining *local grammars* extracted from the texts. The output artefacts of the system could be used to assist domain experts and knowledge engineers in the accelerated building of domain ontologies.

The core of the CiCui system is an indexation engine which facilitates the computations of various text analytic metrics. Raw texts can be retrieved from both dynamic (such RSS feeds) and static (local file system) sources and then converted into CiCui's XML format for storage. The process then invoke Stanford's CoreNLP package (Toutanova et al., 2003) to perform some pre-processing tasks including tokenising, lemmatising and Part-of-Speech tagging. The indexation engine then build a positional inverted index database of the processed corpus for further analysis.

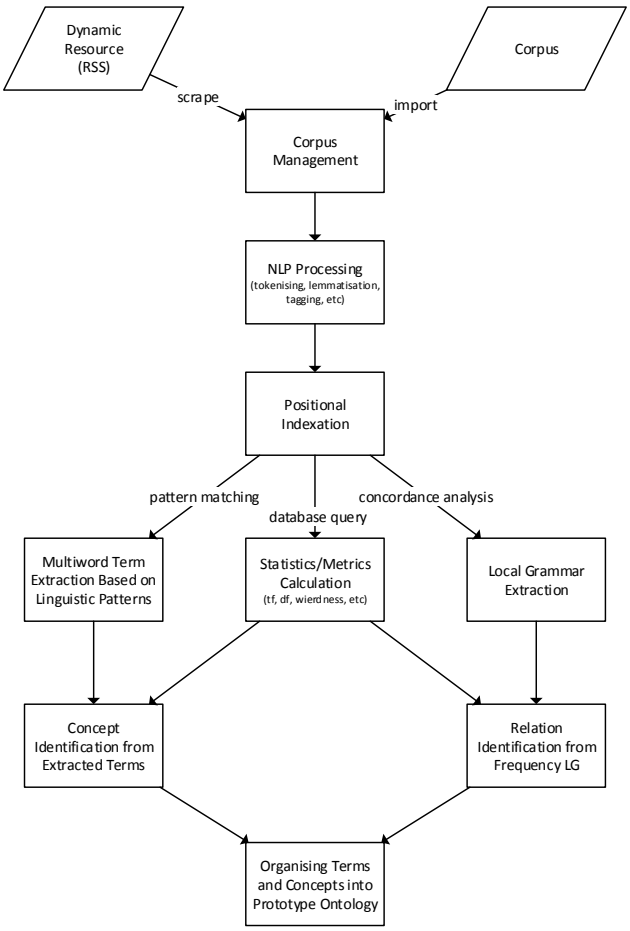


Figure 3: A high-level overview of the processes involved in the ontology extraction system.

3.1. Text Analytics

The CiCui system begins its analysis of the corpus by computing four key statistics for each word w in the corpus' lexicon, as described below:

Term Frequency (tf) The raw frequency of word t in the corpus.

Document Frequency (df) The number of documents in which word w occurred.

TF-IDF Term frequency-inverse document frequency of a word w is calculated as follows:

$$tfidf = \log(tf) \times \log(\frac{N}{df})$$

where tf is the term frequency of word w; df is the document frequency of t; N is the total number of documents in the corpus.

Weirdness The ratio between the relative frequency of the word w in the current corpus and the relative frequency of the same word in the reference corpus (i.e. keyword-ness). See Section 3.2.2. for details.

These statistics are used later in the subsequent phases of the extraction process to identify significant domain keywords.

3.2. Terminology Extraction

3.2.1. Extracting Multi-word Terms

Multi-word terminology extraction constitutes the core of any ontology learning system and has been examined in many previous studies. Frantzi et al. (2000) combined linguistics pattern matching using Part-of-Speech regular expressions and statistical filtering to extract potential domain terminologies. Later studies on multi-word terminology extraction and ontology learning more or less followed the same path. In the CiCui system, we extended this method, producing candidate multi-word phrases by matching the text against a set of linguistic patterns specified in the form of a hand-crafted finite state machine (as illustrated in Figure 4). The FSM utilises both the words' lemma and the Part-of-Speech information to define the linguistic patterns. An overview of the transition conditions used in our finite state machine is given by Table 2.

Most candidate terms come in two fashions – as noun combinations (e.g. "disaster relief", "death toll", etc.) or as adjective-noun combinations (e.g. "clean water", "central bank", etc.). Noun combinations often form terms of good quality by themselves, while adjective-noun combinations are less reliable terms due to the presence of certain quantifiers and determiners such as "other", "many", "last", "such", etc. We therefore introduced a blacklist containing common quantifier adjectives to eliminate trivial adjective-noun combinations.

3.2.2. Keywords/Concepts Identification

To make use of the multi-word terms extracted from the previous phase, it is often necessary to organise them with respect to the ontological concepts they are associated with. Buitelaar et al. (2005) pointed out that "the extraction of concepts from text is controversial as it is not clear what exactly constitutes a concept", but they did suggest that concept induction should provide: *1*) an intensional definition of the concept, *2*) a set of concept instances, i.e. extension, and, *3*) a set of linguistic realisations, i.e. terms for this concept. In this paper, we focus on the third point, namely to extract single-word keywords which capture the higher-level abstractions of the many aspects of the domain.

Most contemporary ontology learning systems typically follow an aggregation-threshold paradigm when it comes to keyword identification. Existing methodologies such as Text-to-Onto (Maedche and Staab, 2001) and Text2Onto (Cimiano and Völker, 2005) often employed tf-idf score to evaluate the likelihood of a term being a top level concept in a domain specific ontology. While being a very effective metric for the task of information retrieval, tf-idf does not fit the purpose of concept identification in a domain corpus very well. In a general and balanced corpus, words that are good indicators of significant topics often occur frequently only in a small portion of the documents, thus bearing high tf-idf scores. In a domain-specific corpus, however, such topic concepts would receive a lower tf-idf weight because they usually appear in the majority of the documents (i.e. low inverse document frequency) – yet it is exactly these "prevalent" terms that constitute the conceptions of the domain.

Ahmad (1995) and Ahmad and Gillam (2005) examined the use of weirdness score as the key metric for extracting key domain specific concepts. In its most basic form, the weirdness score is defined as:

$$weirdness(w) = \frac{\frac{f_{domain}(w)}{N_{domain}}}{\frac{f_{general}(w)}{N_{general}}}$$

$f_{domain}(w)$ is the frequency of term w in the domain corpus; N_{domain} is the total number of words in the domain corpus. Similarly, $f_{general}(w)$ and $N_{general}$ denotes the frequency of term w in the reference corpus of general language and the total number of words in the general language corpus respectively. In other words, the weirdness score of a term characterise how many times more/less likely to see the term in a domain specific context than in a reference corpus containing general language usages. In practice, the word frequency distribution usually was treated with "add-one" smoothing.

In more recent works, "weirdness"-style score as a metric statistic for domain concept extraction has seen furthter development and adoption (Jiang et al., 2010; Gelbukh et al., 2010). In our method, we extend the statistics-centric approach adopted in previous studies by introducing unsupervised clustering and multivariate analysis techniques to help us separate domain concepts from other trivial words. The concept identification procedure begins with the multi-word phrases obtained from the terminology extraction process. Each word w in the multi-word term list is associated with the following four statistics[1]:

Concept Frequency The total frequency of word w within the scope of the multi-word terms list, namely:

$$cf(w) = \sum_{t \in T} f(w, t)$$

where T is the entire set of multi-word phrases extracted from the corpus. $f(w, t)$ is defined as:

$$f(w,t) = \begin{cases} f(t) & w \in t \\ 0 & w \notin t \end{cases}$$

where $f(t)$ is the frequency of term t in the entire corpus.

Document Frequency The total number of documents in which the word w occurred.

[1]This statistic is transformed using logarithm function and then treated with min-max normalisation.

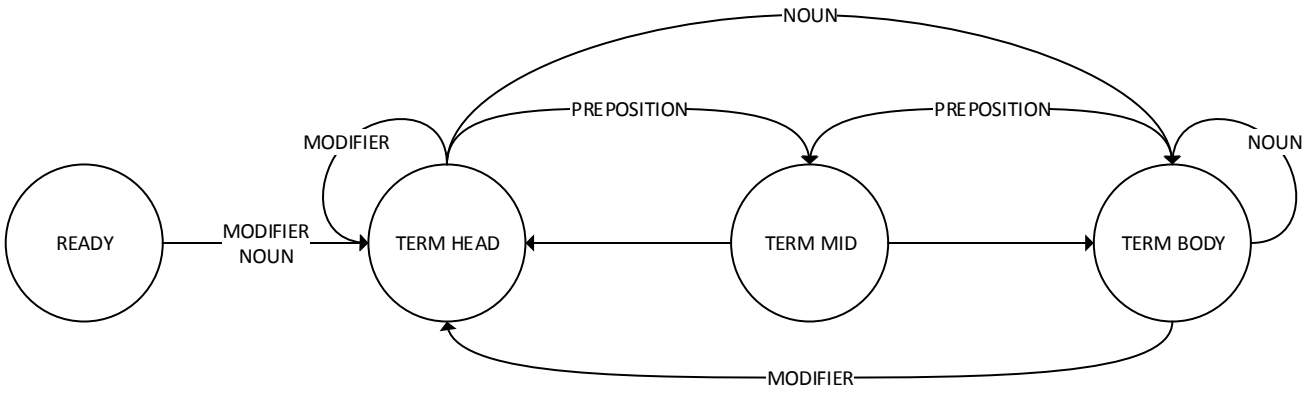


Figure 4: The finite state machine for multi-word term extraction. Note that for the sake of brevity, all `OTHER` transitions have been omitted from the diagram. Any such transition will reset the state machine to the `READY` state.

Priority	Condition	Resulting Symbol
7	the word is in the blacklist	OTHER
6	the word is "of" or "for"	PREPOSITION
5	the word is shorter than 3 characters or not a proper word	OTHER
4	the word is an adjective	MODIFIER
3	the word is a noun	NOUN
2	the word is "the"	THE
1	anything that is not one of the valid transitions above for the state in question	OTHER

Table 2: An overview of the transition conditions used in the FSM for multi-word terminology extraction. Note that if multiple conditions are met for a single word token then the rule with the highest priority wins. For example, if the incoming word is "of", both the second and third rules are triggered, in which case the second rule would win and the word is considered a `PREPOSITION` rather than `OTHER` as the second rule has a higher priority than the third (i.e. $6 > 5$).

Terminology Frequency The total number of multi-word terms in which the word w occurred.

Weirdness The weirdness score of the word w as defined previously.

A clustering analysis technique, fuzzy c-means, is used to see how the concept four-dimensional vectors cluster[2]. The clustering analysis computes for each candidate concept a vector of probabilities that quantifies its memberships for each of the clusters. Each candidate concept word w is assigned to its winning cluster: a cluster C is considered to be the "winning cluster" for a candidate concept word w if C's membership probability is the highest for w among all the clusters.

To select the clusters that better captured domain concepts, the clusters are ranked with respect to their centres' locations in the vector space. A new four dimensional dataset in the same vector space of the candidate concepts' is constructed using the cluster centres' coordinates, after which the new dataset is treated with principal component analysis. We then project each cluster's centre coordinate on the first principal dimension obtained from the PCA and use the projection as a measure to rank the clusters[3].

The number of clusters to find during the clustering analysis affect the "granularity" of the concepts discovery process – increasing the number of clusters to find means we can make a more fine-grained cut when separating the concept words from the non-concept ones, but it also makes it difficult to decide the number of clusters to be kept.

3.3. Extraction of Ontological Relationship

Relationships identification in ontology has been the focus of many studies in the past. In general, most of the research effort focused on the automatic extraction of two types of relationships: *1*) superordinate/subordinates relationships, i.e. taxonomical relationships, and, *2*) causal relationships. In the solving of these two problems, recent studies often featured extensive use of natural language processing techniques, especially phrase chunking and dependency parsing, as well as the incorporation of extrinsic knowledge in the form of common ontologies in order to deduce domain taxonomies from text (Navigli and Velardi, 2004; Cimiano et al., 2005; Cimiano, 2006; Zouaq and Nkambou, 2009; Zouaq et al., 2011; Wong et al., 2012). However, fully automated extraction of ontological relationships utilising sophisticated techniques is still in its early stage of development; its output often requires examination and correction by human experts. In this study, we aim to describe a simpler but self-contained approach to the identification

[2]For implementation one can use KNIME data mining platform (Berthold et al., 2007) with its default parameters and settings.

[3]This projection will subsequently be referred to as the *Cluster/PCA projection score*.

of ontological relationships from text, which would serve as supplement resources to further ontology creation practices.

3.3.1. Forming Taxonomical Structure

Once the significant concepts and multi-word terminologies are extracted, they can be arranged into a hierarchical structure which to some extent reflects the taxonomical relationships (i.e. *is-a*) between the domain concepts.

Multi-word terms extracted from Section 3.2.1. are first filtered, dropped if none of the keywords found in the concept identification phase is present in the term. Terms that passed the filter are then arranged to form a taxonomical hierarchy following a simple linguistic rule: if a multi-word term has a domain concept keyword as its headword, namely the multi-word term ends with the concept keyword, then the multi-word term is considered the subordinate of the concept.

In addition to the *is-a* relation acquired above, it is also possible to derive a *related-to* relation between terms and single word concepts. This can be achieved by linking a multi-word term with its highest ranking composing concept according to the Cluster/PCA projection score described in Section 3.2.2..

3.3.2. Causal Relationship Extraction and Local Grammars

In this paper, we took an alternative approach to identifying causal relationship for ontology learning than utilising dependency parsing techniques. The method, which exploits a linguistic phenomenon called *local grammars* (Gross, 1993), was inspired by Traboulsi et al. (2004)'s work, where one of the authors has expressed the view that local grammar patterns maybe be used in exploring causal relationships.

Local grammars "*are rules that govern the simultaneous choice of a set of words used in a specialist context*" (Traboulsi et al., 2004). Method based on generalisations of local grammars has been applied to explore syntactic structures for verbs such as "bend" and "link" (Mason, 2004).

Our local grammar extraction algorithm is implemented largely as a variant of the Apriori algorithm which mines frequent sequential patterns from large databases (Agrawal and Srikant, 1994). The local grammar extraction process starts with a list of candidate seed verbs; the algorithm then iteratively searches through the corpus' concordance space using CiCui's indexation facility and concordance engine, extending the seed words and forming longer candidate patterns while dropping the ones whose frequency fall below a certain threshold.

4. Case Study and Results

4.1. Data

The corpus analysed in the case study was assembled using articles collected from the ProQuest database, consisting of English text published on newspapers, magazines and wire feeds during the period of 1987 to 2014. The keywords used for the query of articles were a list of disaster-related terms selected using ProQuest's built-in thesaurus with the word "disaster" set as the "seed" term:

Avalanches	Floods
Backup systems	Humanitarian aid
Catastrophes	Hurricane
Contingency planning	Landslides & mudslides
Disaster recovery	Management of crises
Disaster relief	Storm damage
Disasters	Storms
Earthquake damage	Tidal waves
Earthquakes	Tides
Emergencies	Tsunamis

We instructed ProQuest to search the subject field of the articles by using the "SU" operator provided by ProQuest's searching interface. The search was performed only against the meta-data section of the articles (i.e. not on the full-text) to ensure high relevancy of the text retrieved.

The corpus constructed using the aforementioned criteria contains 18,046 articles with duplications removed, totalling 10,938,539 words. Most of our texts are from financial publications and more than half come from two major financial newspapers, *The Wall Street Journal* and *The Financial Times* (see Table 3). There have been suggestions how one should use the ontology of news to organise text corpora (Fernández et al., 2010).

Source	Percentage
Wall Street Journal	39%
Financial Times	14%
Asia Pulse	12%
The Economist	4%
The Financial Express	3%
Macleans	2%
The Economic Times	2%
The Financial Post	2%
Others	22%

Table 3: The distribution of texts in the corpus by publication source.

4.2. Terminology Collation and Analysis

Single Word Concepts Table 4 displays the top ten single word concepts ranked according to each of the four metrics described in Section 3.2.2. as well as a fifth metric (with the header of "Cluster/PCA Proj." in the table), which is the single word's projection of its word vector on the first principle dimension obtained from the PCA analysis of the word cluster centres[4]. Further analysis suggested that the weirdness scoring scheme[5] tends to assign higher weights to very specific domain concepts, especially disaster types – *quake*, *earthquake*, *tsunami*, *flood*, etc., while frequency based metrics promote more general domain concepts such

[4]Based on empirical experiments, we decided to set the number of clusters assumed by the fuzzy c-means algorithm to 16 and retained the top three clusters as per ranked by the method described in Section 3.2.2..

[5]Davies (2008)'s COCA was used as the reference corpus for the calculation of weirdness. We manually developed a mapping from Penn Treebank's POS tag set to CLAWS7 POS tag set so that the weirdness scores can be computed more accurately.

Rank	Cluster/PCA Proj.	Weirdness	Concept Frequency	Terminology Frequency	Document Frequency
1	government	quake	government	government	death
2	disaster	earthquake	disaster	disaster	toll
3	earthquake	tsunami	datum	datum	government
4	official	north	death	dozen	official
5	datum	datum	toll	group	earthquake
6	death	flood	price	project	product
7	toll	insurer	official	company	disaster
8	company	west	earthquake	development	relief
9	relief	neighborhood	company	system	warming
10	area	floodwater	area	price	agency

Table 4: Top 10 single word concepts ranked according to the projection of word vectors on the first principal component from the PCA analysis described in Section 3.2.2., weirdness, concept frequency, terminology frequency and document frequency.

as *disaster*, *government*, *business*, etc. The top ranking concepts according to the Cluster/PCA projection score, by contrast, contains a mixture of both specific and general domain concepts.

Organisation of Multi-word Terms Following the method detailed in Section 3.3.1., we took the 111 single word concepts from the top ranking cluster as per the Cluster/PCA projection score and form a taxonomical structure for the disaster management terminologies.

An small example ontology was produced in RDF format and imported into Protégé[6]. We manually adjusted the ontology to add taxonomical structure for the single word concepts since such relationships could not be derived automatically from the text without referring to external knowledge bases. A reproduction of a visualisation of the ontology in OntoGraf[7] is shown in Figure 5.

The ontology thus produced may later be enriched by domain experts and knowledge engineers, adding more relationships such as object properties, subclasses, equivalence, etc.

The computer-aided ontology creation process may help to achieve a continuous work flow where ontologies can be updated based on the output of the system described in this paper while more text is being acquired.

4.3. Relation Identification through Local Grammar Patterns

Verbs The distribution of verbs in the two corpora, ours and COCA, is not as different as is the case for nouns, especially when we look at the weirdness of key verbs. Note that verbs comprise 10.5% of the corpus, 5 times more than the first 100 nouns. Nevertheless, there are 17 verbs that weirdness more than 10, one, the word *according* has a weirdness of 200, and *rebuild* has a weirdness of 53. Reporting of an ongoing disaster usually involves statements by authority figures, and occasionally victims, that involve the phrase of *according to [Named Entity]*.

[6]The Protégé package: `http://protege.stanford.edu`

[7]`http://protegewiki.stanford.edu/wiki/OntoGraf`

Local Grammar and Ontological Relationship Looking at the weird keyword in our corpus, *hurricane*, we explored the collocation patterns between the keyword and verbs in our corpus, using the following local grammar patterns:

The result was that there were six key patterns that occurred in the corpus involving 21 different verb. The most frequent pattern was *caused by [Determiner] hurricane [Named Entity]* – in the network below we show the various relationships between verbs and *hurricane*: the frequency of the patterns is scaled within each of the six relationships, with minimum given the score of unity and others compared with the minimum frequency (see Figure 6).

5. Conclusion and Future Work

In the above we presented a corpus-based method of creating prototype partial ontology of a specialist domain. The identification of concepts through "dominant" keywords appears to work well and in earlier research on financial risk we had found that experts generally agreed with our findings. The work on causal relationships is critical for understanding the dynamics of a given domain – the named entities provide a static view of the domain. Our work reported in this paper is based exclusively on newspaper corpus, addition of texts from emergency management agencies will enrich the ontology presented above.

6. Acknowledgements

The authors would like to thank EU sponsored Slaindail Project (FP7 Security sponsored project #6076921).

7. References

Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco and CA and USA. Morgan Kaufmann Publishers Inc.

Ahmad, K. and Gillam, L. (2005). Automatic Ontology Extraction from Unstructured Texts. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F.,

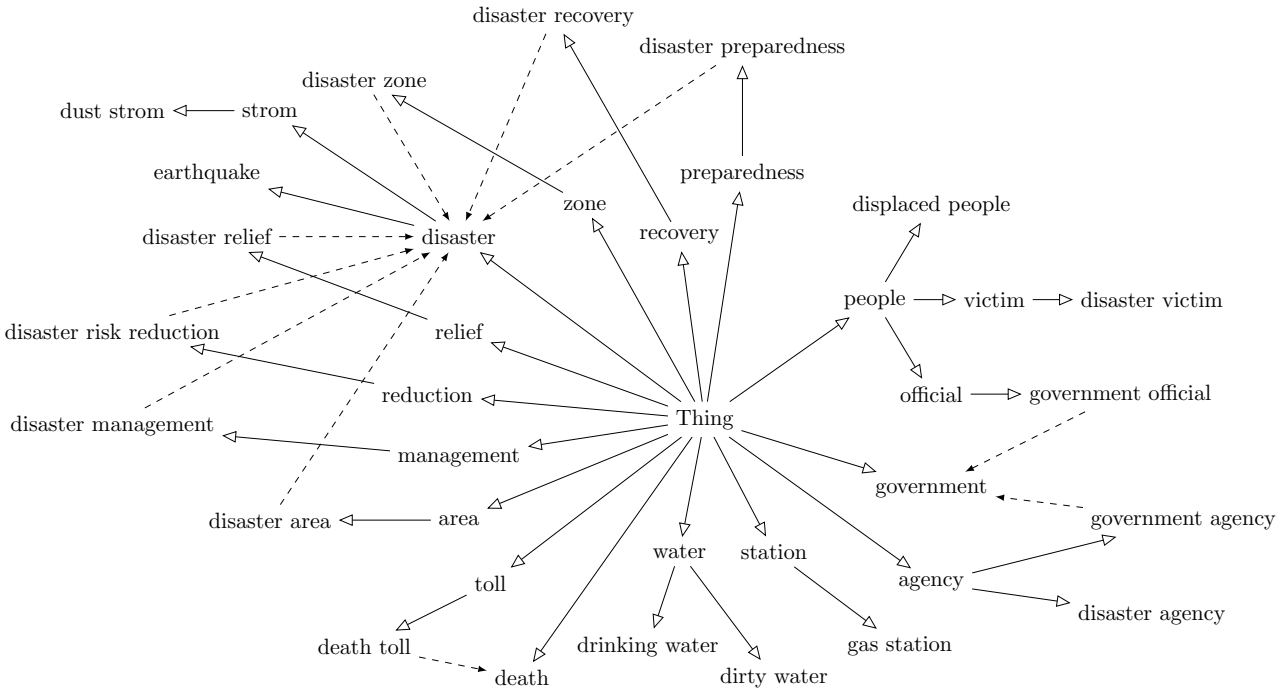


Figure 5: A example toy ontology for Disaster Management produced with the assistance of the method presented in this paper. The ontology was made from high frequency multi-word terms extracted from the corpus as well as their associated single word concepts as detailed in Section 3.3.1.. The solid arrows represent *is-a* relationships (e.g. an *earthquake* is a *disaster*) while the dashed arrows denote *related-to* relationships (e.g. *government official* and *government agency* are related to the concept of *government*)

.

Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Meersman, R., and Tari, Z., editors, *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, volume 3761 of *Lecture Notes in Computer Science*, pages 1330–1346. Springer Berlin Heidelberg, Berlin and Heidelberg.

Ahmad, K. (1995). Pragmatics of specialist terms: The acquisition and representation of terminology. In Carbonell, J. G., Siekmann, J., Goos, G., Hartmanis, J., Leeuwen, J., and Steffens, P., editors, *Machine Translation and the Lexicon*, volume 898 of *Lecture Notes in Computer Science*, pages 51–76. Springer Berlin Heidelberg, Berlin and Heidelberg.

Ahmad, K. (2007). Artificial Ontologies and Real Thoughts: Populating the Semantic Web? In Basili, R. and Pazienza, M. T., editors, *AI*IA 2007: Artificial Intelligence and Human-Oriented Computing*, volume 4733 of *Lecture Notes in Computer Science*, pages 3–23. Springer Berlin Heidelberg, Berlin and Heidelberg.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer.

Buitelaar, P., Cimiano, P., and Magnini, B. (2005). Ontology Learning from Text: An Overview. In Buitelaar, P., Cimiano, P., and Magnini, B., editors, *Ontology learning from text*, volume v. 123 of *Frontiers in artificial intelligence and applications*, pages 3–12. IOS Press, Amsterdam and Washington and DC.

Chou, C.-H., Zahedi, F. M., and Zhao, H. (2011). Ontology for Developing Web Sites for Natural Disaster Management: Methodology and Implementation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 41(1):50–62.

Cimiano, P. and Völker, J. (2005). Text2Onto. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Montoyo, A., Muńoz, R., and Métais, E., editors, *Natural Language Processing and Information Systems*, volume 3513 of *Lecture Notes in Computer Science*, pages 227–238. Springer Berlin Heidelberg, Berlin and Heidelberg.

Cimiano, P., Pivk, A., Schmidt-Thieme, L., and Staab, S. (2005). Learning taxonomic relations from heterogeneous sources of evidence. In Buitelaar, P., Cimiano, P., and Magnini, B., editors, *Ontology learning from text*, volume v. 123 of *Frontiers in artificial intelligence and applications*, pages 59–73. IOS Press, Amsterdam and Washington and DC.

Cimiano, P. (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer US, Berlin, 1. ed edition.

Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Steven-

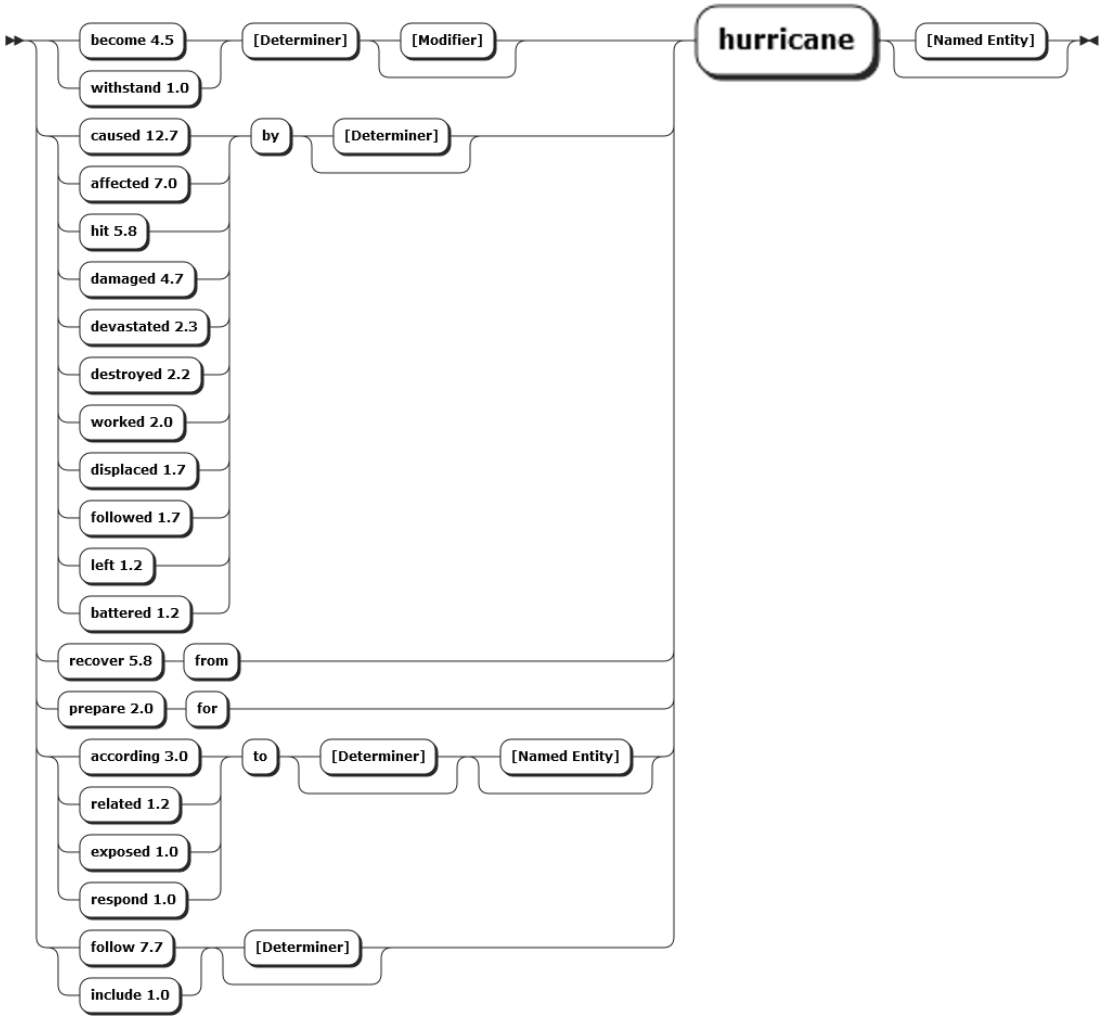


Figure 6: A relational network comprising a key concept and verbs that collocate with the concept. The scores of unity indicate that the part of the network has a minimum frequency and others within the network is measured w.r.t. to the minimum pattern.

son, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T. Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., and Jaiswal, P. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant & cell physiology*, 54(2):e1.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990-present.

Fernández, N., Fuentes, D., Sánchez, L., and Fisteus, J. A. (2010). The NEWS ontology: Design and applications. *Expert Systems with Applications*, 37(12):8694–8704.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Gelbukh, A., Sidorov, G., Lavin-Villa, E., and Chanona-Hernandez, L. (2010). Automatic Term Extraction Using Log-Likelihood Based Comparison with General Reference Corpus. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Hopfe, C. J., Rezgui, Y., Métais, E., Preece, A., and Li, H., editors, *Natural Language Processing and Information Systems*, volume 6177 of *Lecture Notes in Computer Science*, pages 248–255. Springer Berlin Heidelberg, Berlin and Heidelberg.

Gross, M. (1993). Local Grammars and Their Representation by Finite Automata. In Hoey, M., editor, *Data, Description, Discourse*, pages 26–38. Harper-Collins Publishers, London.

Jiang, Y., Lin, C. X., and Mei, Q. (2010). Context Comparison of Bursty Events in Web Search and Online Media. In Li, H. and M‘arquez, L., editors, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1077–1087, Cambridge and MA. Association for Computational Linguistics.

Klien, E. M., Lutz, M., and Kuhn, W. (2006). Ontology-based discovery of geographic information services—An application in disaster management. *Computers, Environment and Urban Systems*, 30(1):102–123.

Maedche, A. and Staab, S. (2001). Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2):72–79.

Martin, L. and Simon, S. (2008). A Formula for Disaster: The Department of Homeland Security's Virtual Ontology. *Space and Polity*, 12(3):281–296.

Mason, O. (2004). Automatic Processing of Local Grammar Patterns. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, University of Birmingham*, pages 166–171.

Navigli, R. and Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30(2):151–179.

Rapisardi, E. and Franco, S. D. (2004). Disaster Resilience and the Babel of Semantic.

Smith, B. (2004). Ontology. In Floridi, L., editor, *The Blackwell Guide to the Philosophy of Computing and Information*. Blackwell, Malden and Mass and Oxford.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Hearst, M. and Ostendorf, M., editors, *the 2003 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 173–180.

Traboulsi, H., Cheng, D., and Ahmad, K. (2004). Text Corpora, Local Grammars and Prediction. In *Proceedings of the Fourth International Conference on Language*. European Language Resources Association.

Wang, Z. and Lei, J. (2013). A Study of Spatial Information Sharing Model Based on Geological Hazard Domain Ontology. In *2013 Fifth International Conference on Computational and Information Sciences (ICCIS)*, pages 268–270.

Wong, W., Liu, W., and Bennamoun, M. (2012). Ontology learning from text. *ACM Computing Surveys*, 44(4):1–36.

Wu, K., Li, L., Li, J., and Li, T. (2013). Ontology-enriched multi-document summarization in disaster management using submodular function. *Information Sciences*, 224:118–129.

Zouaq, A. and Nkambou, R. (2009). Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project. *IEEE Transactions on Knowledge and Data Engineering*, 21(11):1559–1572.

Zouaq, A., Gasevic, D., and Hatala, M. (2011). Towards open ontology learning and filtering. *Information Systems*, 36(7):1064–1081.

A Study in Domain-Independent Information Extraction for Disaster Management

Lars Döhling, Jirka Lewandowski, Ulf Leser

Department of Computer Science, Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany
doehling, lewandow, leser@informatik.hu-berlin.de

Abstract

During and after natural disasters, detailed information about their impact is a key for successful relief operations. In the 21st century, such information can be found on the Web, traditionally provided by news agencies and recently through social media by affected people themselves. Manual information acquisition from such texts requires ongoing reading and analyzing, a costly process with very limited scalability. Automatic extraction offers fast information acquisition, but usually requires specifically trained extraction models based on annotated data. Due to changes in the language used, switching domains like from earthquake to flood requires training a new model in many approaches. Retraining in turn demands annotated data for the new domain. In this work, we study the cross-domain robustness of models for extracting casualty numbers from disaster reports. Our models are based on dictionaries, regular expressions, and patterns in dependency graphs. We provide an evaluation on extraction robustness across two disaster types – earthquakes and floods. It shows that applying extra-domain models without retraining gives a relative F1 decrease of solely 9 %. This is a fairly small drop compared to previous results for similar complex extraction tasks.

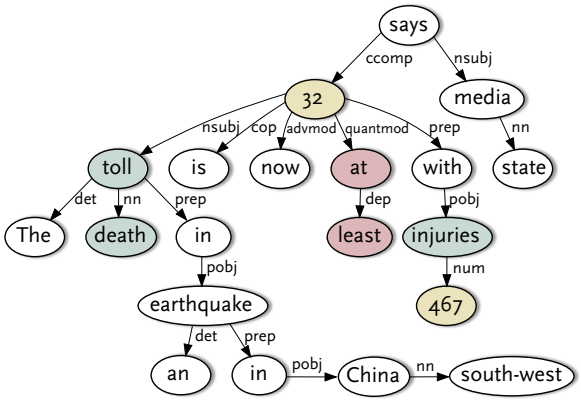
Keywords: natural language processing, cross-domain learning, n-ary relationship extraction, dependency patterns

1. Introduction

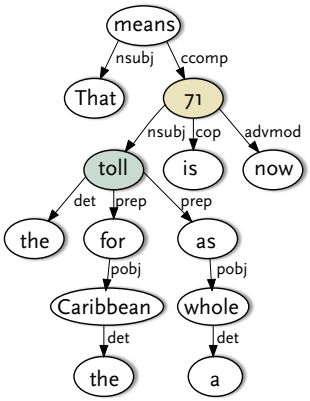
Crisis events like earthquakes or disease outbreaks are striking humankind regularly. In the aftermath, decision makers require precise and timely information to assess damages and to coordinate relief operations (Guha-Sapi and Lechat, 1986). Understanding "the big picture" in emergency situations is obviously essential for effective responses. Today, the Internet plays an important role for information acquisition, especially if no on-site contact is available. Supportive information are published both in conventional sources like newspapers (Döhling and Leser, 2011) as well as in social media, e.g. web forums (Qu et al., 2009) or microblogs (Vieweg et al., 2010). These sources offer the most details available, but searching and analyzing them manually is a time-consuming and therefore costly task.

Information Extraction (IE) studies the problem of automatically extracting structured information from given unstructured text (Sarawagi, 2008). By offering fast and on-line information acquisition, IE methods may help to mitigate disaster effects. Developing IE applications typically involves supervised learning, i.e. requires annotated training data to generate and test extraction models. Examples are CRFs for entity recognition (McCallum, 2003) or SVMs for classification tasks (Cortes and Vapnik, 1995). To achieve high quality results, such training data must be sourced from the same domain as the extraction will be applied to. Here, the term 'domain' refers to the texts used, especially their type (news article, tweet) and topic (earthquake, flood). As a consequence, generated models from these training data are domain-specific as well. While achieving optimal results in the original domain, they often perform poorly when applied to different, even closely related domains. For instance, (McClosky, 2010) evaluated the cross performance of PCFG-based parsing models for corpora from diverse domains. Syntactical sentence parsing is a prerequisite for many state-of-the-art IE methods, including the one presented in this paper. For closely related domains, the F1 score decreased relatively by 3 %, while for more distant domains it dropped by 10 %. (Jakob and Gurevych, 2010) studied recognizing opinion targets (what the opinion is about) in user-generated reviews. They observed a relative F1 decrease of 12 % on average when applying CRF-based models across four topics. (Tikk et al., 2010) analyzed an even more complex IE task, (binary) relationship extraction. They measured the cross-corpus performance of SVM-based models for extracting protein-protein interactions. Although all corpora consisted of biomedical texts, their experiments revealed a relative F1 decrease of 24 % on average. To prevent such performance losses, applying extraction methods in new domains mostly requires retraining appropriate models. Retraining in turn demands new annotated data and annotating is an expensive and cumbersome manual task. An alternative approach is to use extraction methods based on robust models performing well across domains.

In this paper, we present such models for extract casualty numbers from disaster reports in multiple domains. Casualty numbers are an indicator for the scale of damage, determining the required extend of relief operations. Our extraction models (Section 2) are based on dictionaries, regular expressions, and patterns in dependency graphs. Cross-domain evaluation results for earthquake and flood reports are given in Section 3 and discussed in Section 4.



(a) 'The death toll in an earthquake in south-west China is now at least 32, with 467 injuries, state media says.'[1]



(b) 'That means the toll for the Caribbean as a whole is now 71.'[2]

Figure 1: Dependency graphs in the (a) earthquake and (b) flood domain. All relevant entities are colored.

2. Information Extraction

The information extraction procedure is based on our method presented in (Döhling and Leser, 2011). It allows to extract arbitrary facts from texts, formalized as n-ary relationships. Here, n denotes the number of entities – single words or word groups – used to express the fact. For instance, the sentence 'The death toll [...] is now at least 32, with 467 injuries [...]'[1] contains two facts: $\geq$ 32 killed and 467 injured. We formalize these facts as 4-tuples [modifier, quantity, subject, type], resulting in [at least, 32, –, death toll] and [–, 467, –, injuries]. Each 4-tuple is defined by:

- Modifier: modifies quantity values, e.g. 'at least', 'about', or 'more than'
- Quantity: numbers casualties and consists of two subtypes: cardinal ('12', 'ten', 'no', 'a') and vague ('many', 'hundreds', 'some')
- Subject: characterizes casualties explicitly, e.g. 'people', 'villagers', or 'students'
- Type: describes the type of damage and consists of multiple subtypes, e.g. injured or trapped

2.1. Extraction Pipeline and Model

The automatic fact extraction consists of three steps. First, we recognize all entities (relevant words or word groups) of the targeted relationship, e.g. '32', 'at least', or 'injuries'. Cardinal quantities are recognized by a domain-independent regular expression, all other entities by a dictionary derived from training data. In contrast to more sophisticated extraction models, e.g. CRFs, dictionaries carry very little contextual information. Consequently, they are potentially more robust when applied across domains. In addition to (Döhling and Leser, 2011), we enhance this step by two optional post filters for cardinal quantities: M-Filter and A-Filter. As the regular expression identifying cardinals does not encode the context, the M-Filter removes potentially false positives by revoking those surrounded by units of measurement, e.g. 'ft', 'km', '$', or '%'. The A-Filter withdraws all 'a'/'an' annotations, as the majority of these terms refer to the indefinite article and not to the cardinal 1, resulting in many false positives. Next, we infer semantic relationships between pairs of entities by matching patterns in dependency graphs. Dependency graphs (Figure 1) model the syntactical relationships between the words of a sentence as typed, directed edges between them. By offering direct access to sentence structures, they often reveal relations between words far apart more easily than if modeled as lists of words (Fundel et al., 2007). Given the example in Figure 1a, the distance on the surface level between the related entities 'death toll' and '32' is ten words, whereas they are directly connected in the corresponding dependency representation. We use the shortest paths between two entities as patterns, collected from training data. (Bunescu and Mooney, 2005) showed that these paths are well suited to capture relationships between entities within sentences. Similar to dictionary entries, shortest paths carry minimal contextual information, again supporting cross-domain robustness. For instance, although Figure 1a's sentence contains 'earthquake', this domain-specific keyword is not part of the shortest paths (Figure 2). In addition, the shortest path in Figure 1b is the same as in Figure 1a, emphasizing potential domain-independence of patterns. The pattern matching is adjustable by optionally ignoring the dependency type or direction, or the entity subtype. Ignoring the entity subtype originates from the observation that subtypes are often interchangeable within sentences, e.g. 'injured 13 people' vs. 'buried many people'.

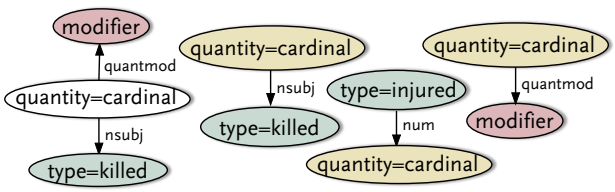


Figure 2: Shortest paths patterns, derived from Figure 1's dependency graphs.

[1] news.bbc.co.uk/2/hi/asia-pacific/7591152.stm
[2] au.news.yahoo.com/world/a/15270957/haiti-storm/

The results of the pattern matching step are entity graphs with edges between all pairs of related entities (Figure 3). By finding maximal cliques (McDonald et al., 2005) in these entity graphs, the binary relationships are finally merged into tuples of the desired n-ary relationship. In this paper, that is the 4-ary relationship modeling reported casualties. Compared to (Döhling and Leser, 2011), we enhance the last step by an optional, domain-independent post filter for handling enumerations of facts within one sentence. Given '[...] killed X and injured Y [...]', our method also extracts false tuples like [–, X, –, injured] due to similar dependency structures compared to true tuples. As each quantity belongs to only one type, the Enum-Filter investigates all tuples sharing the same quantity and keeps only the most probable one. Its decision is based on the sentence's token sequence level. It considers distances between entities as well as linguistic hints, such as 'and'.

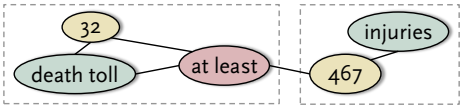


Figure 3: An entity graph, derived from Figure 1a's dependency graph. The rectangles mark all contained maximal and valid cliques, i.e. having one type entity.

3. Evaluation and Results

We evaluated the domain independence of acquired extraction models by comparing their intra-domain performance against cross-domain results. Each model consists of (1) the entity dictionary, (2) the dependency patterns, and (3) the pipeline configuration, i.e. matching and filter switches.

3.1. Data sets

We used two data sets, consisting of news articles collected from the web reporting on earthquakes and floods, respectively (Table 1). The earthquake articles were sourced from BBC and Yahoo! News in 2009/10. The flood articles were selected from various search engine results in 2012. Each article was manually annotated with the 4-ary relationship, covering six casualty types: injured, killed, homeless, affected, missing, and trapped. Both corpora are available on request. We also examined the inter-annotator agreement on corpus samples, which was 82 % on average. We partitioned each corpus into a training (2/3) and an evaluation set (1/3) by stratified random sampling on the sentence level.

	Earthquake	Flood
Documents	245	412
Sentences	4795	8616
Tokens	100 303	187 894
Relationship tuples	1277	1860
size=2	483 (38 %)	570 (31 %)
=3	507 (40 %)	900 (48 %)
=4	287 (22 %)	390 (21 %)
type=killed	825 (65 %)	1362 (73 %)
=injured	224 (17 %)	63 (4 %)
=trapped	74 (6 %)	7 (0 %)
=missing	81 (6 %)	166 (9 %)
=homeless	49 (4 %)	88 (5 %)
=affected	24 (2 %)	174 (9 %)

Table 1: Corpus statistics; size refers the number of set entities within tuples.

Parameter	Earthquake	Flood
M-Filter	enabled	
A-Filter	enabled	disabled
Ignore dependency type	true	
Ignore dependency direction	false	
Ignore entity subtype	true	
Enum-Filter	enabled	

Table 2: Best extraction pipeline configurations per corpus, F1-optimized at the final relationship level.

3.2. Experiments

For the intra-domain experiments (source=target), the models were trained on the training set and evaluated on the evaluation set within the same domain. For the cross-domain experiments (source≠target), the models were trained on the extra-domain training set and evaluated on the evaluation set. permitting fair comparisons. The extraction configuration was derived by 5-fold cross-validation on the respective training set, maximizing the average F1 score (Table 2). Our experiments showed a relative F1 decrease of 9.0 % on average (geometric mean) if applying models across domains (Table 3, top). Both recall (−12.6 %) as well as precision (−4.5 %) declined.

	Target	
Source	Earthquake	Flood
Earthquake	78/317/114	153/437/202
	.803/.735/**.768**	.741/.684/**.711**
Flood	94/285/146	159/518/121
	.752/.661/**.704**	.765/.811/**.787**
Enhanced		
Earthquake	117/333/98	_
	.740/.773/**.756**	
Flood	_	221/526/113
		.704/.823/**.759**

Table 3: Intra-/cross-domain evaluation results (top) and enhanced intra-domain results (bottom); Numbers are false positives/true positives/false negatives and precision/recall/F1 at the final relationship level.

Encouraged by the low decrease in performance, we further analyzed the potential benefit of adding extra-domain data in the intra-domain setting. For each domain, we enhanced the intra-domain training set by in-

cluding the complete extra-domain data. We kept the extraction configuration. The resulting mixed-domain models were evaluated on the intra-domain evaluation set as before. Both domains showed a small relative increase in recall (+2.8 %), but a significant decrease in precision (−7.9 %) (Table 3, bottom). The resulting F1 scores were slightly lower than those without extra-domain data (−2.3 %).

4. Discussion and Conclusion

We studied the cross-domain robustness of models for extracting casualty numbers from disaster reports. Our evaluation showed that applying models across disaster types results in only 9 % F1 relative performance decrease. This is a small drop compared to, for instance, the 24 % observed for extracting protein-protein interactions (Tikk et al., 2010), a similar complex extraction task.

By checking the trained models and comparing the underlying data sets, we identified two main reasons for the observed domain independence. Both are connected to each other and equally important. (1) Sentences reporting on casualties use similar structures and wordings to express facts, independent of the domain. (2) Most entries of the acquired entity dictionaries and pattern catalogues are comprised of domain-unspecific words. The dictionaries overlap by approximately 44 % of their entries. Of all entries, only about 3 % are domain-specific, e.g. 'quake toll' or 'drownings'. These figures can be observed in the patterns as well, having an overlap of around 32 % and a domain specificity of roughly 4 %.

We further investigated the potential positive effect of additional extra-domain data on the intra-domain performance. Due to larger dictionaries and pattern catalogues, we observed a slight increase in recall at the cost of a clear decrease in precision. So extra-domain data introduced more false than true positives, favoring single-domain over mixed-domain models.

5. Acknowledgements

We kindly thank Christoph Fischer for contributing to the annotations. We also thank Mariana Neves for providing valuable feedback on a draft of this article.

6. References

R.C. Bunescu and R.J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *HLT/EMNLP 2005*.

C. Cortes and V. Vapnik. 1995. Support-Vector Networks. *Machine Learning*, 20(3).

L. Döhling and U. Leser. 2011. EquatorNLP: Pattern-based Information Extraction for Disaster Response. In *Terra Cognita 2011*.

K. Fundel, R. Küffner, and R. Zimmer. 2007. RelEx - Relation extraction using dependency parse trees. *Bioinformatics*, 23(3).

D. Guha-Sapi and M.F. Lechat. 1986. Information systems and needs assessment in natural disasters: An approach for better disaster relief management. *Disasters*, 10(3).

N. Jakob and I. Gurevych. 2010. Extracting Opinion Targets in a Single- and Cross-domain Setting with Conditional Random Fields. In *EMNLP '10*.

A. McCallum. 2003. Efficiently Inducing Features of Conditional Random Fields. In *UAI '03*.

D. McClosky. 2010. *Any Domain Parsing: Automatic Domain Adapatation for Parsing*. Ph.D. thesis, Brown.

R. McDonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, and P. White. 2005. Simple algorithms for complex relation extraction with applications to biomedical IE. In *ACL '05*.

Y. Qu, P.F. Wu, and X. Wang. 2009. Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthq. In *HICSS'09*.

S. Sarawagi. 2008. Information Extraction. *Foundations and trends in databases*, 1(3).

D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. 2010. A Comprehensive Benchmark of Kernel Methods to Extract Protein-Protein Interactions from Literature. *PLoS Comput Biol*, 6(7).

S. Vieweg, A.L. Hughes, K. Starbird, and L. Palen. 2010. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *CHI 2010*.

OCR of Legacy Documents as a Building Block in Industrial Disaster Prevention

Daniel Isemann[1], Andreas Niekler[1], Benedict Preßler, Frank Viereck, Gerhard Heyer

Universität Leipzig
Department of Computer Science
{isemann, aniekler, heyerasv}@informatik.uni-leipzig.de, {mam10fdj, mam10hry}@studserv.uni-leipzig.de

Abstract

Legacy text documenting or recording operational details from industrial or other human-made facilities which constitute potential hazards may contain important safety critical information. In these cases it is desirable to make such safety critical information as is implicitly present in older records readily available and accessible to modern day authorities which have to deal with the facility from an industrial disaster prevention point of view. An important first step in such an analysis, we argue, is a robust optical character recognition (OCR) stage, for digitising often decade old records containing valuable information for industrial disaster prevention or management which may otherwise be lost or unavailable at the right time. In this paper we present an overview of a project concerned with the study and analysis of legacy records from the operational history of a deep geological repository for nuclear waste and present a preliminary study on segmenting letterhead information in legacy correspondence concerning this particular facility.

Keywords: legacy documents, OCR, layout analysis

1. Introduction

Industrial hazards as a special type of human-made hazards are often associated with facilities consuming or producing large amounts of energy. Such facilities can present a risk for human-made disaster or may, as was recently seen in the case of the Fukushima power plant, aggravate the impact of a natural disaster. Some of these facilities and industrial compounds are decades old and associated with a complex and varied operational history. In the case of deep geological repositories for nuclear waste the longevity and durability of the facility as well as certain safeguards for its continued and uninterrupted operation are, or at any rate should be, an integral part of the facility's design. It has been noted for some time that ensuring such longevity presets not only a formidable engineering challenge but also a problem of understanding and interpreting relevant information correctly. One aspect that has received a considerable amount of attention in this context is the semiotics of communicating the dangers of radiological waste to future generations (cf. (Trauth et al., 1993) for instance).

A different but related topic and one that has not been addressed in depth so far, is concerned with preserving access to the documentation of the operational history of safety critical facilities such as nuclear waste deposition compounds. 'Access' here refers not only to physical access, but additionally to terminological and semantic access to the institutional knowledge represented by and preserved in legacy documents.

In this paper we present an overview and preliminary results from the project "Wissensmanagement von Altdokumenten aus Forschung, Verwaltung und Betrieb"[2] funded by the German Federal Ministry of Education and Research, led by the German Research Centre for Environmental Health (Helmholtz Zentrum) in Munich and carried out partly at the University of Leipzig. The project is concerned with the study and analysis of legacy records from the operational history of a deep geological repository for nuclear waste in Lower Saxony, Germany. An important first step in such an analysis is a robust optical character recognition (OCR) stage, for digitising the often decade old legacy records which are at various stages of physical decay and contain information on, among other things, safety features of the facility or the delivery and placement of radioactive material within it.

We first give some background on the project at hand and, more generally, discuss the requirements of an OCR-based analysis of legacy records (Section 2). Following this, we present preliminary results from an experiment of segmenting letterheads into their constituent parts (Section 3). Taking this experiment as a starting point we briefly sketch possible lines of future work (Section 4) and finally conclude our discussion (Section 5).

2. Background

The facility in question was operated as a salt mine from 1909 until 1964 when it was sold to the Federal Republic of Germany. Subsequently the former salt mine became a deep geological repository for weakly and medium radioactive nuclear waste. Contrary to original plans strongly radioactive material was not deposited in the facility. Although the site has stopped accepting nuclear waste decades ago, it continues to be a stumbling block in contemporary political debate. In recent years a discussion has developed around the question whether due to certain structural shortcomings of the pit the nuclear waste should be recovered and relocated. In this context, the question which substances are kept where and in what quantities has to be addressed by recourse to archival material.

Extensive records survive from the operation of the deep geological repository. Most of these are kept on paper of

[1]These authors contributed equally as first authors.

[2]German for "Knowledge management of legacy records in science, administration and industry".

varying degrees of quality and some are in fairly poor condition. The physical files comprise a diverse range of documents of different types and vastly different sizes, ranging from individual paper snippets to specialist dossiers of hundreds of pages with multiple appendices. Examples of types of documents (with no claim to completeness) are:

- Geological and mining surveys
- Studies on radioactive waste management and reactive transport modelling
- Correspondence of scientific and administrative nature
- Telephone protocols and meeting minutes
- Delivery notes and receipts for radioactive substances
- Reports on and assessments of extraordinary events (e.g. earthquakes, operational incidents, etc.)
- Guidelines and written exhortations making reference to legal and regulatory frameworks governing nuclear waste disposal

These documents are set in many different typefaces (serif, sans serif, typewriter font, etc.) and come from a wide variety of reproduction methods (printing press, typewriter, gelatin hectograph). Moreover, many of these documents include handwritten notes, receipt stamps, date stamps or other administrative notes (see Appendix A, Figure 6 for an example of a document). Although the period where radioactive waste was accepted and actively managed in the facility only lasted from 1965 to 1978, the documents at hand span almost a hundred years if one includes historic mining surveys from the archives. Only the latest documents relevant to the facility's operation, such as reports of the 21st Parliamentary Inquiry Committee of the State Parliament of Lower Saxony are already available in digitised form.

2.1. Requirements for an OCR system for legacy records.

The aim of the project "Wissensmanagement von Altdokumenten aus Forschung, Verwaltung und Betrieb" is twofold: first, to preserve the surviving records and make them more readily accessible, e.g. through keyword search (*archiving*); second to achieve a more in-depth content analysis of documents and to put domain experts in a position to filter and evaluate these documents based on their specific information needs, e.g. by using sophisticated text mining approaches (*repurposing*). We will address these two requirements in turn, with a particular view to OCR technology.

Archiving. This includes digitising documents and linking digitised documents to correct and meaningful keywords. The general challenge presented by optical character recognition, is expected to be considerably more difficult in the particular case of legacy documents because of poor paper and print quality, contamination and damage. Over and beyond this subject-specific terminology used in a particular set of documents may necessitate special dictionaries or glossaries for OCR correction, which may need to be closely integrated with the overall OCR process.

Repurposing. The output of an OCR process may contain metadata pertaining to the process itself, such as which OCR system was used and which capabilities this system had, but it is devoid of metadata pertaining to the domain or the particular task under consideration. Depending on the requirements of a specific analysis or the nature of the questions put to a repository of documents appropriate metadata has to be defined and the documents have to be enriched with metadata accordingly. In order to make digitally preserved records accessible to the scrutiny of domain experts in the sense of a repurposing or rediscovering of information, it is crucial to attempt to generate as much relevant metadata as is feasible. In the case of legacy correspondence, information such as the date, sender or institutional affiliation of the recipient of a message may be important filters for a faceted expert analysis. Such information will also enrich possible interpretations of furthergoing analyses of the material, such as thematic clustering or topic analysis.

In the case of our project one of the desiderata is to subject available correspondence to a network analysis and for this to detect automatically when a document or letter was sent, who the sender and receiver were, which institutions they are or were associated with, whether the correspondence has been answered and so forth. It is envisaged that this will help in understanding institutional roles and expertise and allow to integrate critical knowledge extracted from the documents better into a coherent whole.

To facilitate the generation of relevant metadata it is necessary from an OCR point of view to achieve a robust layout analysis (i.e. identify blocks of text which belong together). A more comprehensive solution may include handwritten margin notes and stamps in the analysis, which are not covered by conventional OCR software.

2.2. Existing OCR solutions

Different techniques have been employed for OCR (Charles et al., 2012) and a number of OCR solutions for personal and business applications exist on the market, including commercial packages (ABBYY FineReader 11, OmniPage Ultimate, Prizmo), open source projects (OCRFeeder, OCRopus) as well as online and cloud services (Free OCR,[3] OCR in Google Drive, Online OCR[4]) (cf. Table 1). ABBYY FineReader has a strong OCR engine and is considered by some the market leader among commercial providers.[5] However, the commercial nature of the software may present an obstacle to a tailored solution which suits the particular needs of our project. Open source platforms such as OCRFeeder or OCRopus may be more accessible to customising a solution. A detailed evaluation of the advantages and disadvantages of existing OCR packages and a recommendation which one can best serve as the basis for a project specific OCR solution given the objectives and requirements of our project has to wait until more scanned material from the archives of the deep geo-

[3] `http://www.free-ocr.com` (last accessed 30/3/2014).

[4] `http://www.onlineocr.net` (last accessed 30/3/2014).

[5] Compare e.g. `http://www.pcmag.com/article2/0,2817,2403191,00.asp` (last accessed 30/3/2014).

logical repository becomes available. Partly due to logistic and technical reasons connected to the scanning process, partly due to the politically sensitive nature of the material only a small number of documents from the facility's archive has been made available to project partners as of this writing. We will present preliminary analyses in this paper performed on samples taken from a corpus of 75 scanned documents ranging from 1911 to 2005. As the majority of the documents are not yet digitised, an evaluation of different digitisation approaches may encompass the entire workflow (scanning, document segmentation, OCR, text analysis, etc.) and a tailored OCR solution may interact with lexical and semantic processing as well as the level of scanner specifications and hardware drivers.

Scanned documents from three different decades were processed with the available services Free OCR, Online OCR and the OCR service provided by Google Drive. The initial sentences of some meeting minutes from 1965 are shown in digitised form in Figure 1. Applying off the shelf OCR solutions to such a specialised task clearly has its limitations. The Google Drive OCR, for instance, recognises less than half of the words in the short segment correctly (Figure 2). In principle though OCR holds great promise for digitising legacy records. The text produced by the Online OCR service is almost entirely legible. However, the OCR engine failed to recognise key terms such as "Versuchsarbeiten" (exploratory work), "Unterbringung" (depositing) and "radioaktive[n] Abfälle[n]" (radioactive waste) correctly (Figure 3).

3. Case Study: Segmentation of Letterheads

As a first foray into possible approaches for a network analysis on legacy correspondence, letterheads from correspondence concerning the operation of the facility were subjected to analysis with the goal of segmenting the text into blocks belonging together. OCR was performed on available scans of 24 letters using the open source OCRopus package which was especially adapted for the case of legacy files.[6] It is worth pointing out, that as OCRopus is supported by Google[7] it is likely the engine behind the OCR service offered as part of Google Drive. The relatively poor performance of Google Drive OCR on our data (cf. Figures 2 and 3) may perhaps be seen as an indication that legacy records necessitate a highly customised OCR solution. The output of the OCR process is in hOCR format, an XHTML based format in which recognised text can be tagged and absolutely positioned and unrecognised elements, such as stamps or marks on the paper can be embedded as images (cf. (Breuel, 2010)). The analysis was performed purely on the letterheads, with the body of the letters and embedded images stripped away first (see Appendix A for an illustration of the preprocessing steps).

The chosen approach for grouping the text into blocks which belong together was that of an agglomerative (i.e. bottom-up) hierarchical clustering with subsequent cluster flattening. The clustering made use of the absolute position information of text elements (Figure 4). In a preprocessing step the horizontal and vertical position coordinates were extracted from the hOCR file.

```
<span class='ocr_line' title='bbox
    261 2648 1064 2719' style=
        'position:absolute;
        left:93.2142857143px;
        top:313.928571429px'>
    Forschung und</span><br />
```

Figure 4: A text element from from the hOCR output of a letter. The position information that was extracted for the cluster analysis is highlighted in bold.

The clustering method chosen was that of complete linkage clustering, where the distance between two clusters is determined by the maximum of the distances between individual cluster elements. Although not universally recommended (cf. e.g. (Romesburg, 2004, p. 126f.)) it seemed appropriate for the case at hand as it tends to lead to compact clusters of approximately equal diameter and avoids certain artefacts of other agglomerative schemes, such as the chaining phenomenon sometimes found in single-linkage clustering (Everitt et al., 2001). Euclidian distance was employed as distance metric.

In order to determine the number and identity of the suggested text blocks the hierarchically clustered data was "flattened" by observing the variance (i.e. the average distance between data points in clusters) for five different scenarios, ranging from two to six clusters. We flattened the clustering on different hierarchy levels and calculated the inner cluster average distance for each number of clusters. We selected that number of clusters for which the next smaller number exhibited a significant increase in variance compared to previous variance gaps (Table 2). This method allowed us to determine at which hierarchy level the clustering merges clusters which are not representing nearby content on the letterheads. In this fashion the elements in each letterhead were grouped into two to five distinct blocks.[8]

Thus while the overall clustering approach is bottom-up it is guided by the general, "top down" expectation that clusters will be compact, of (roughly) uniform size and usually range from two to five in number. This approach identifies geometrical layout units in a letterhead which will ideally correspond to logical units (e.g. address, sender, date etc.). The intuition behind the variance gap method is that a particularly big decrease in variance is likely to signify the merging of two clusters which are somewhat distant from each other and hence potentially the merging of two logically separate text blocks (Figure 5).

[6] Thanks are due to Mathias Bank and Behrang Shafei of the CID Group for supplying their current best effort OCR results for the letters in question.

[7] https://code.google.com/p/ocropus (last accessed 30/3/2014).

[8] The cluster analysis and cluster flattening was implemented using a free version of the data analytics toolkit RapidMiner Studio (http://rapidminer.com/products/rapidminer-studio, last accessed 30/3/2014).

Doc. Nr.	Variance for n clusters 2	3	4	5	6	Estim. # of clusters
3	0.52	0.42	0.36	0.24	0.23	5
7	0.80	0.43	0.28	0.23	0.19	4
11	0.50	0.38	0.33	0.22	0.16	5
15	0.59	0.55	0.30	0.22	0.16	4
17	0.55	0.33	-	-	-	3
21	0.71	0.40	0.38	0.24	0.20	5
34	0.86	0.62	0.32	0.28	0.21	4
36	0.68	0.36	0.30	0.26	0.20	4
42	0.70	0.40	0.37	0.24	0.22	5
44	0.70	0.40	0.37	0.24	0.22	5
46	0.65	0.41	0.38	0.23	0.21	5
48	0.62	0.37	0.25	-	-	3
50	0.51	0.36	0.28	0.23	0.21	5
52	0.55	0.38	0.28	0.23	0.17	4
54	0.55	0.34	0.31	0.25	0.22	3
56	0.63	0.38	0.25	-	-	3
58	0.53	0.48	0.44	0.30	0.24	3
66	0.58	0.54	0.30	0.28	0.24	4
137	0.77	0.41	0.36	0.30	0.19	6
150	0.80	0.41	0.33	0.28	0.21	4
154	0.50	0.41	0.27	0.23	0.21	5
186	0.55	0.33	-	-	-	2
200	0.51	0.46	0.34	0.33	0.29	4
206	0.78	0.59	0.31	0.25	0.19	4

Table 2: The average within cluster variance (average cluster distortion) for different clusterings, i.e. different numbers of clusters, for the 24 letterheads in our experiment. The estimated optimal number of clusters in each case corresponds to the point where decreasing the number of clusters would result in a significant increase in cluster distortion. Some of the input files contained less than six text fragments. In these cases a dash is displayed in place of the distortion value.

4. Future Work

A number of lines of future work arise from our preliminary study. While the variance gap method described above is designed to capture an observers intuition as to how many text clusters can be found in a letterhead it is necessary to evaluate this approach rigorously and on a large document sample, e.g. by showing pictures similar to Figure 5 to multiple raters to determine which segmentation they favour. Such an evaluation framework would also provide an invaluable guide for expanding our agglomerative clustering approach, e.g. by experimenting with different clustering schemes and distance metrics. A weighted Euclidian metric in which vertical distance is privileged over horizontal distance might give more robust results in the context of letterheads.

A possible next step in the context of legacy correspondence is to employ methods of machine learning to classify letterhead blocks according to their semantics (i.e. according to what they signify, e.g. address, sender, date).

More generally, we will carry out quantitative evaluations to determine the most promising segmentation approaches for various types of text and documents found in the archives of the radioactive waste repository, as soon as more data becomes available. In light of our discussion in Section 2, a prerequisite for this will be to conduct a thorough requirements analysis among the domain experts concerned with processing and interpreting the documents in question.

5. Conclusion

We have introduced the project "Wissensmanagement von Altdokumenten aus Forschung, Verwaltung und Betrieb" and more generally have attempted to motivate the need for improved semantic access to legacy text on paper and some of the preconditions and requirements for turning repositories of paper documents into valuable practical resources for domain experts and researchers. This is an open problem which requires both adjustments to and interaction between current state of the art OCR technology as well as sophisticated text mining techniques. The archive of the deep geological repository for radioactive waste which we have studied is a point in case to illustrate how successfully providing such access may play an important role in the long term prevention and disaster mitigation planning for industrial hazards.

According to a well-known adage nothing is older than yesterday's newspaper. However, when things in a complex and potentially hazardous facility containing nuclear waste come to a head, nothing might be more topical than that delivery note for radioactive substances from 1965 which you happened to have overlooked when you so desperately needed to find it.

6. References

Thomas Breuel. 2010. The hOCR Embedded OCR Workflow and Output Format. `https://docs.google.com/document/d/1QQnIQtvdAC_8n92-LhwPcjtAUFwBlzE8EWnKAxlgVf0/preview`, March.

Pranob K. Charles, V. Harish, M. Swathi, and C. H. Deepthi. 2012. A review on the various techniques used for optical character recognition. *International Journal of Engineering Research and Applications*, 2(1):659–662.

Brian S. Everitt, Sabine Landau, and Morven Leese. 2001. *Cluster Analysis*. Arnold, London, fourth edition.

Charles Romesburg. 2004. *Cluster analysis for researchers*. Lulu Press, North Carolina.

Kathleen M. Trauth, Stephen C. Hora, and Robert V. Guzowski. 1993. Expert judgement on markers to deter inadvertent human intrusion into the waste isolation pilot plant. Sandia Report SAND92—1382, UC—721, Sandia National Laboratories, Albuquerque, New Mexico 87185 and Livermore, California 94550, November.

	special fonts supported	workflow integration	targeted at business applications	open source	free
ABBYY FineReader 11	✓	✓	✓	✗	✗
Free OCR	?	✗	✗	✗	✓
OCRFeeder	?	✗	✗	✓	✓
OCR in Google Drive	?	✗	✗	✗	✓
OCRopus	✓	✗	✗	✓	✓
OmniPage Ultimate	?	✓	✓	✗	✗
Online OCR	?	✗	✗	✗	✓
Prizmo	?	✓	✓	✗	✗

Table 1: Comparative overview over some available OCR packages and services.

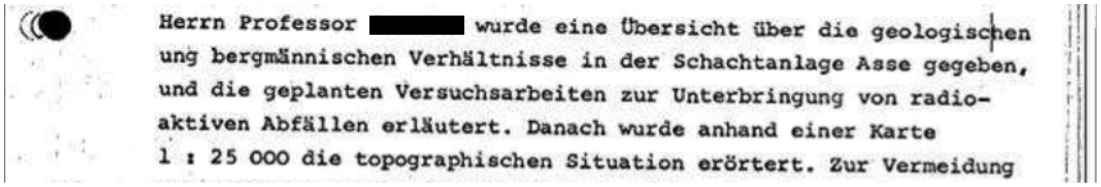


Figure 1: The two initial sentences of a 1965 meeting protocol from the archives. A personal name is blacked out for the purpose of protecting privacy.

. — Herrn Professor XXXXX wurde eine: bar:=icht ber? du: genlug:.scfuan um; hergmŁnnischen UerhŁltnisse in der: Schachtanlage um:
und die qaplanten versuchsarbaitan zu: Unterbringung man radiaMEŁILEn erlŁutert. Danach wurde anhand einer Harta
1 : 25 dia top-ugraphischen Eituatinn arŁ:terr:.. Zur vermeiflung

Figure 2: Output of the Google Drive OCR module (language setting German) for the segment depicted in Figure 1. The equivalent of a personal name, although not recognised correctly, has been replaced by ‘X’ marks.

Herrn Professor XXXXXX wurde ein* bercicht ber dio geologischen 14.11/4.9111111 une bergmŁnnischen VerhŁltnisse in der Schachtanlage Asse 9.agebenf und die geplanten Verauchsarbeiten zur Unterbrin9ung von radio-aktiven AbfŁllen erlŁutert. Danach wurde anhand einer Karte 1 : 25 000 die topographischen Situation errtert. Zur Vermeidung

Figure 3: Out of the *Online OCR* service (language setting German) for the segment depicted in Figure 1. The equivalent of a personal name, although not recognised correctly, has been replaced by ‘X’ marks.

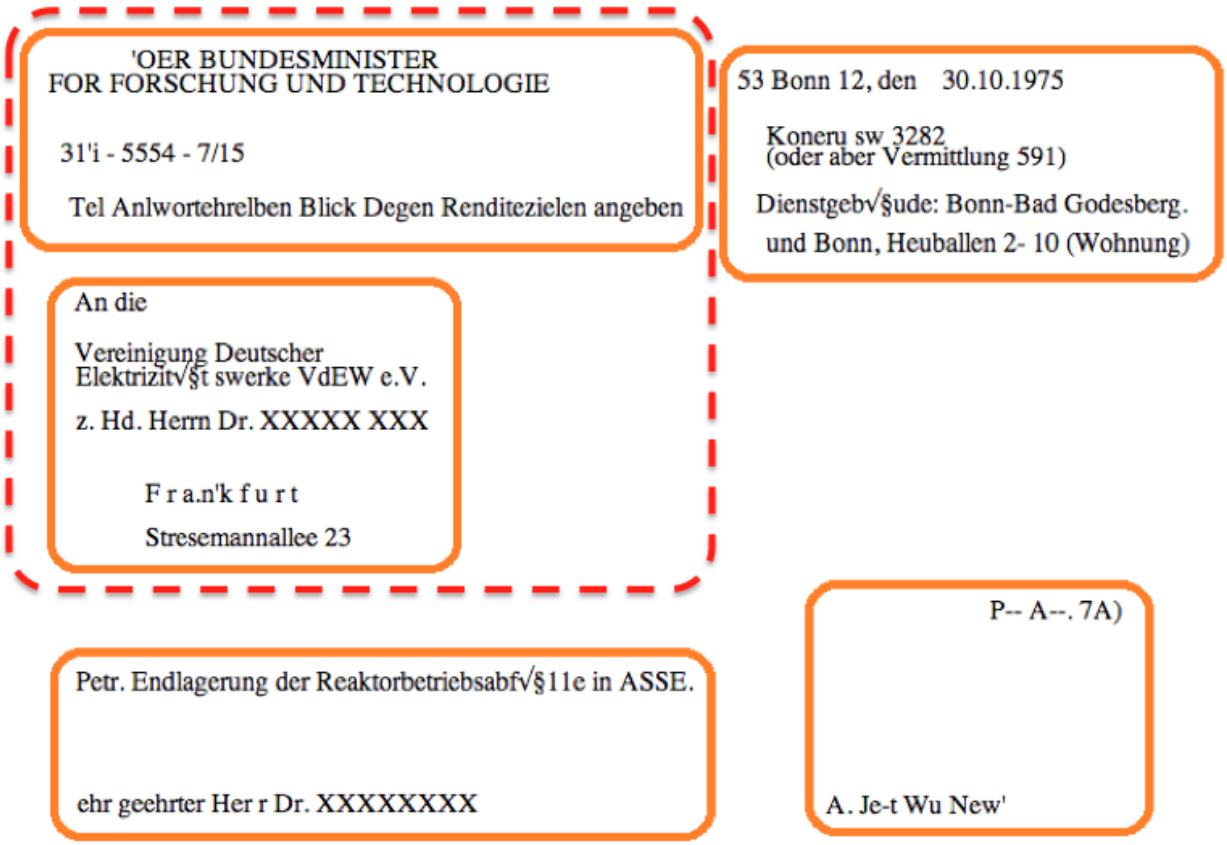
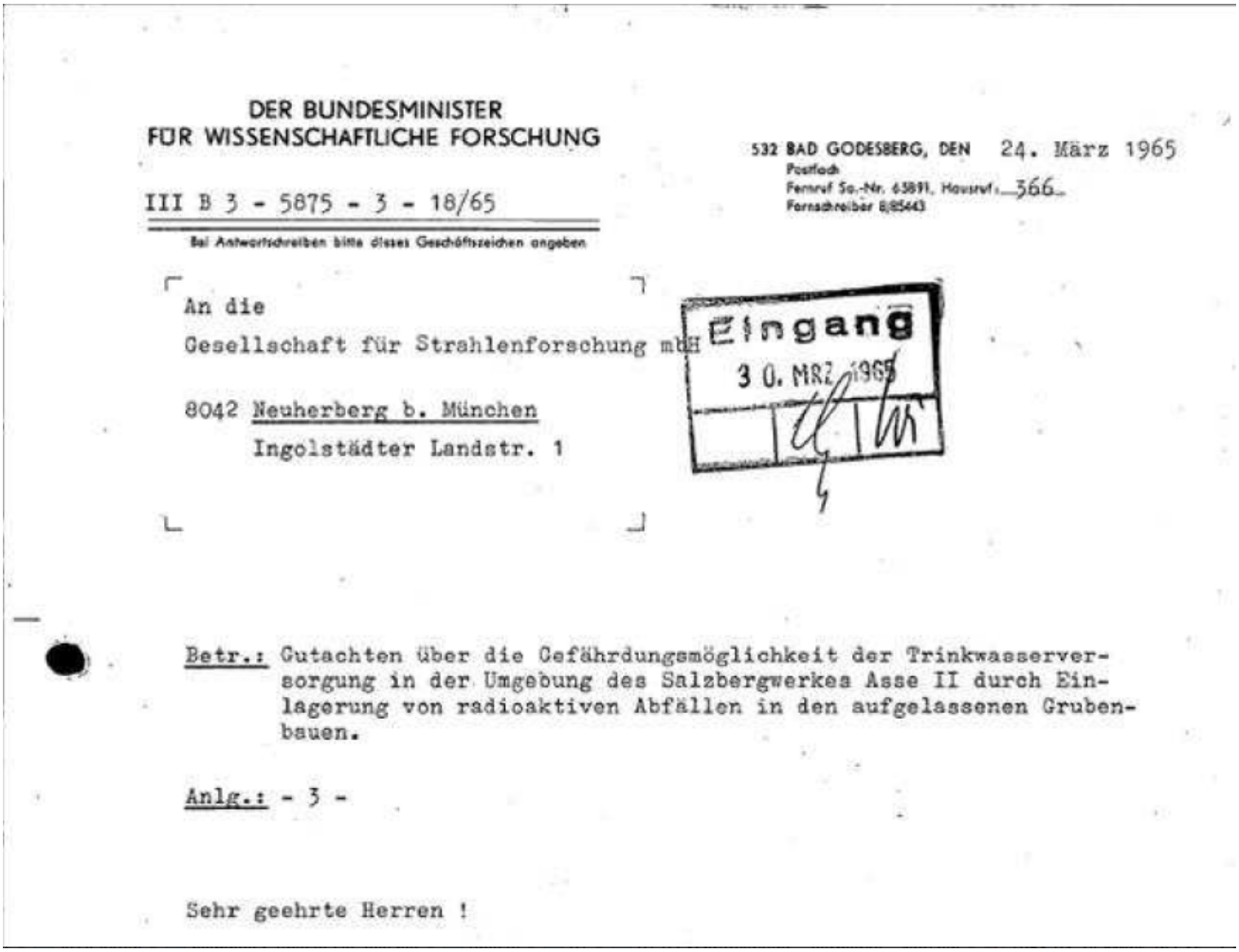


Figure 5: The best estimate produced by our method for the clustering of document no. 3 (cf. Appendix A, Figure 7) is shown by the solid line boxes. In this case the transition from four to five clusters showed the biggest decrease in variance which corresponds to the merging of the two clusters on the top left resulting in the cluster circled by a dashed line.

Appendix A

The following figures illustrate documents at different stages of our analysis: as input to the OCR process (Figure 6), output from the OCR process (Figure 7) and input to the cluster analysis (Figure 8). Figure 6 shows part of a letter from 1965, Figures 7 and 8 show hOCR output with and without embedded images for a different letter from 1975.



DER BUNDESMINISTER
FÜR WISSENSCHAFTLICHE FORSCHUNG

532 BAD GODESBERG, DEN 24. März 1965
Postfach
Fernruf Sa.-Nr. 63891, Hausruf 366
Fernschreiber 8/85443

III B 3 - 5875 - 3 - 18/65

Bei Antwortschreiben bitte dieses Geschäftszeichen angeben

An die
Gesellschaft für Strahlenforschung mbH

8042 Neuherberg b. München
Ingolstädter Landstr. 1

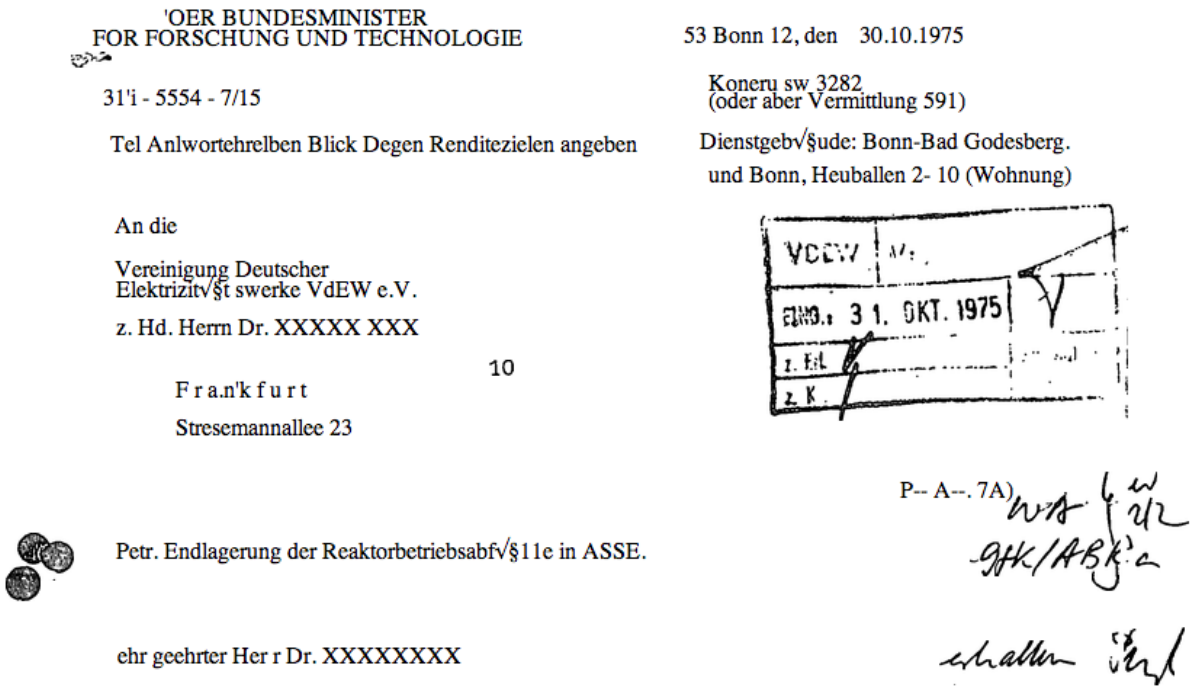
Eingang
30. MRZ 1965

Betr.: Gutachten über die Gefährdungsmöglichkeit der Trinkwasserversorgung in der Umgebung des Salzbergwerkes Asse II durch Einlagerung von radioaktiven Abfällen in den aufgelassenen Grubenbauen.

Anlg.: - 3 -

Sehr geehrte Herren !

Figure 6: Scan of a letterhead from one of the documents in the archive we studied.



'OER BUNDESMINISTER
FOR FORSCHUNG UND TECHNOLOGIE

31'i - 5554 - 7/15

Tel Anlwortehrelben Blick Degen Renditezielen angeben

An die

Vereinigung Deutscher
Elektrizitv§t swerke VdEW e.V.

z. Hd. Herrn Dr. XXXXX XXX

10

F r a.n'k f u r t

Stresemannallee 23

53 Bonn 12, den 30.10.1975

Koneru sw 3282
(oder aber Vermittlung 591)

Dienstgebv§ude: Bonn-Bad Godesberg.
und Bonn, Heuballen 2- 10 (Wohnung)

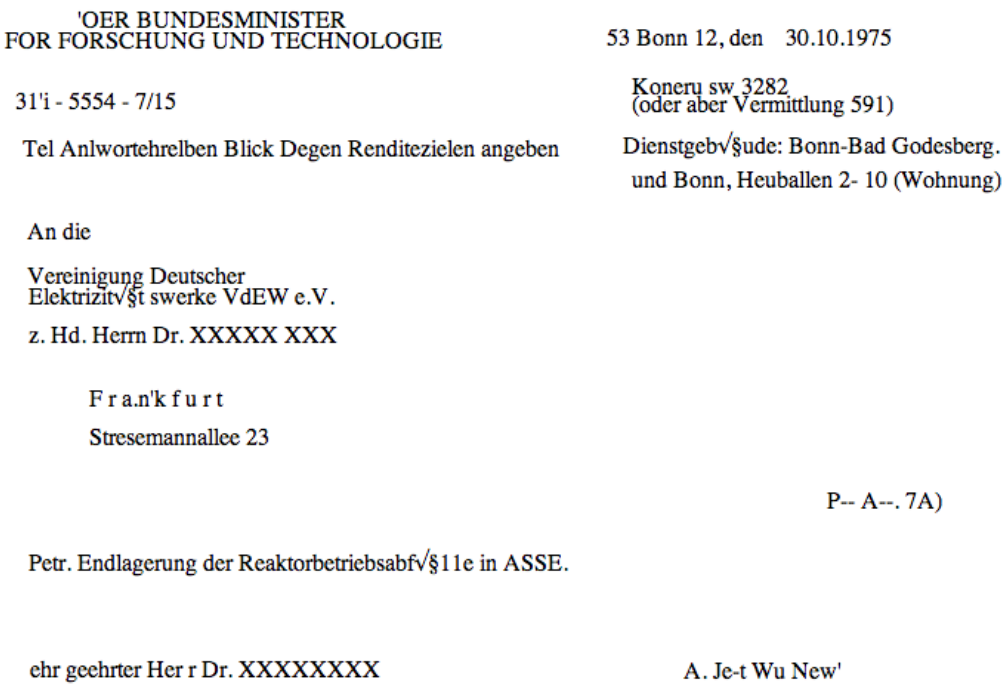
VDEW
EING.: 31. OKT. 1975
z. Kt.
z. K.

P-- A--. 7A)

Petr. Endlagerung der Reaktorbetriebsabfv§11e in ASSE.

ehr geehrter Her r Dr. XXXXXXXX

Figure 7: The original hOCR output for document no. 3 in our clustering experiment as displayed by a browser. For the purposes of this figure the body of the message has been removed and two names have been replaced by ‘X’ marks.



'OER BUNDESMINISTER
FOR FORSCHUNG UND TECHNOLOGIE

53 Bonn 12, den 30.10.1975

31'i - 5554 - 7/15

Koneru sw 3282
(oder aber Vermittlung 591)

Tel Anlwortehrelben Blick Degen Renditezielen angeben

Dienstgebv√§ude: Bonn-Bad Godesberg.
und Bonn, Heuballen 2- 10 (Wohnung)

An die

Vereinigung Deutscher
Elektrizitv√§t swerke VdEW e.V.

z. Hd. Herrn Dr. XXXXX XXX

F r a.n'k f u r t

Stresemannallee 23

P-- A--. 7A)

Petr. Endlagerung der Reaktorbetriebsabfv√§11e in ASSE.

ehr geehrter Her r Dr. XXXXXXXX

A. Je-t Wu New'

Figure 8: The modified hOCR output for document no. 3 (cf. Figure 7 above). Additionally to the body of the message images were stripped from the document and position information extracted from it served as input to the cluster analysis. For the purposes of this figure two names have been replaced by ‘X’ marks.

Determining levels of urgency and anxiety during a natural disaster: Noise, affect, and news in social media

Stephen Kelly, Khurshid Ahmad

Trinity College Dublin
Ireland
kellys25@scss.tcd.ie, kahmad@scss.tcd.ie

Abstract

Since 2010, and perhaps before that as well, news and views of and about citizens caught up in a natural disaster, like floods and hurricanes, are increasingly available through digital media channels. In social media via Twitter for instance- and in formal media, especially in the blogs accompanying news compiled by various public and private sector agencies, one can get information about events as they unfold. Monitoring this stream of digital information provides valuable information for rescue agencies. However, caution has to be exercised in that this stream of information can be ostensibly stored for future analysis, by say resilience planners, without due care for the privacy of named entities, including individuals, places and institutions. In this paper we present a scalable bag-of-words method for analysing social media and crowd-sourced documents to visualise the evolving signature of a disaster event comprising disaster and affect terms. We illustrate our method by using a hurricane and an earthquake case study and two systems developed at Trinity College Dublin an ontology-based, scale-oriented system called Rocksteady and a terminology and ontology extraction system called CiCui. Ethical questions raised by automatic collection and analysis of social media data, especially the collation and storage of named entities is discussed.

Keywords: keyword A, keyword B, keyword C

1. Introduction

The 2010 Haitian Earthquake can be regarded as a coming of age for the use of social media in disaster alleviation in one of the poorest countries in the world which has concomitant problems with infrastructure and civil order (Lewis et al., 2011). The authors established a text-message-based emergency reporting system called Mission 4636 . The SMS messages received by the system helped in tracking disaster victims and emergency responders. The Mission 4636 had facilities to translate between English and Haitian Creole being the language of mainly foreign aid workers and the local community (including the victims and the indispensable local volunteer workers) respectively (Lewis, 2010). Subsequently, digital message transmission and reception systems, some Internet-based, were used in earthquakes in Chile and in Pakistan (Munro and Manning, 2012). That the language used in digitally transmitted messages, including Twitter, SMS and other variants, is not quite up to normal language, and indeed is 'bad' language, has been looked at by (Eisenstein, 2013).

First responders are now using social media in dealing with floods and other civil emergencies and some volunteer groups have referred to the social media traffic as 'psychological first aid' (Taylor et al., 2012). During an emergency, the civil population has to place its trust in identifiable public bodies like police, civil guards, fire and rescue services. There is a need for data to be collected from and about the victims, and an equally important need is for the data to be protected. In Hurricane Sandy, it was noted that the potential victims could communicate the impact of the moving hurricane and seek help at the same time through language comprising terms related to an on-going emergency situation and also by using affect words to indicate the extent of the actual, potential or perceived damage[1]. In the well documented use of SMS during the Haitian Earthquake of 2010, volunteers were able to look at the telephone numbers of the victims, victims could be geo-located by virtue of the addresses they were giving, and an analysis of chat logs could perhaps also help in identifying the victims or their caregivers (Munro, 2013).

The analysis of the data gathered in a number of emergency situations can, perhaps, be used to design new social media systems for crisis management. McClendon and Robinson (McClendon and Robinson, 2013) have compared their system with a system developed in a US National Science Foundation Project (EPIC) to look at the relative merits of the two systems: the authors have emphasised the use of the the frequency, content, and location components of contributed information from a social media based crisis management system in the design of the system.

There are two key issues here that may relate to the use of digital media in general and social media in particular. First, how to analyse the contents of what are essentially text fragments regarding the possible impact of the disaster, especially as the disaster evolves. Second, what is the level of coverage named entities that can help in the identification of the person impacted and the objects associated with the person that were damaged. These two issues are discussed in this paper with the aid of an open-source corpus of Twitter and Facebook messages. These messages were analysed for the coverage of crisis words, affect levels and named entities. Our study is a precursor for work in the EU sponsored Slaindail Project (FP7 Security sponsored project #6076921), where we are exploring the notion

[1] http://www.cabume.co.uk/blog/the-positive-and-negative-sides-of-social-media-tech-and-superstorm-sandy.html

of *intrusion index* for a social media crisis management system proposed by one of the authors of this paper (KA).

2. Methods

Typically, systems like Facebook and Twitter, and systems used by researchers in disaster management, encode each message using a number of meta data elements. In large publicly available text repositories, like Lexis Nexis for example, each text message contains a time stamp, source, title and message text, an example of how each message is structured before being processed and analysed is shown in the following xml mark-up in figure 1.

```
<?xml version="1.0" encoding="UTF-8"?>
- <article>
    <source>FEMA on Facebook</source>
    <article-datetime>04 Nov 2012 00:00:00 +0000</article-datetime>
    <title>To help Hurricane Sandy disaster survivors avoid being misled
      by misinformation...</title>
    <text> Nov 04, 2012 (FEMA on
      Facebook:http://www.facebook.com/fema Delivered by Newstex)
      To help Hurricane Sandy disaster survivors avoid being misled by
      misinformation circulating on social networks, we've created a
      "Rumor Control" section on www.fema.gov/sandy. Check here for
      an on-going list of rumors and their true or false status. Hurricane
      Sandy | FEMA.gov www.fema.gov National Response Coordination
      CenterPhotosVolunteer Agency Coordination Preparing for
      Hurricane Sandy. The entire federal family continues to lean
      forward to support the states, tribes and communities in their
      ongoing response efforts as they work to save lives. </text>
</article>
```

Figure 1: A message from our corpus with relevant meta data. Content has originally been aggregated by Newstex[4]and stored by Lexis Nexis[5].

There is a rich variety of information available in the above text message: Some of the information that relates to a disaster event is available in the tagged title and text element of the message. Other information relates to the origination of the message source, time and date. An automatic analysis of such messages can suggest levels of anxiety, distress and hopefully ultimately relief. Equally important information is the location, duration, and impact of a disaster event contained within these messages. An analysis of a time-ordered sequence of such messages may throw some real-time or historical light on how the disaster unfolded. These messages, especially those sent by disaster management agencies comprise technical terms that have to be analysed. The messages sent by the potential and actual victims, and those related to the victims by proximity and familial/friendship ties, may contain affect terms related to the impact of, and recovery from, a disaster.

We have used the bag-of-words technique complemented by a part-of-speech tagger to identify the grammatical categories of the words analysed. The basic premise of this approach is that the frequency of a words and words related to it is in itself an indication of the meaning implicit in the message, and the acceptance of the meaning within the community which is transmitting and receiving the message. At the level of meaning, we have used a domain specific ontologically organised dictionary of terms to identify and quantify disaster related information at different levels of granularity. The term *disaster* is, for instance, a superordinate term for both *storm* and *flood*, and a *hurricane* is a kind of a *storm*, and a *tsunami* is caused by a *hurricane* and can cause *flooding*. The detection of a disaster term like an earthquake together with a proper noun (Haitian earthquake for example) will help in locating the earthquake.

The text of each message can be analysed, firstly using a dictionary of affect terms from the Harvard IV dictionary used for content analysis in the General Inquirer system(Stone et al., 1966), and secondly by an ontologically organised set of terms relating to the categories of Disaster, Hurricane, Fire, Emergency ,Victim and Floods (Figure 2). The ontology was built and exported using the Protege ontology editor and a knowledge acquisition system[6].

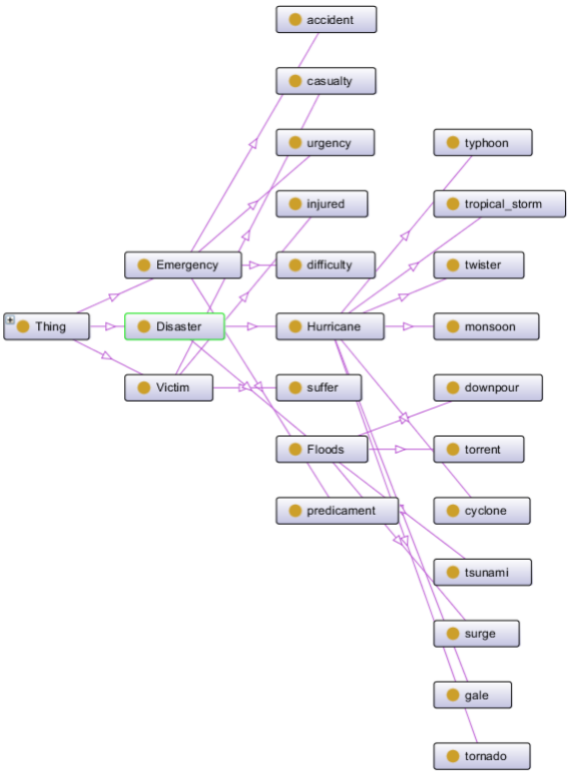


Figure 2: A sample of the ontologically organised disaster lexicon used. Categories are shown with a subset of terms for each node also shown. For clarity we show only a sample of term instances for each individual category. The output was generated using the ontograph plugin in Protege.

The number of messages related to a disaster increase from small numbers just before the disaster, to reach a maximum when the disaster impacts, subsequently, when the disaster is no longer news, then the frequency of the document drops to zero eventually. This requires a time ordering of messages. The frequency of the domain specific terms and affect terms has to be aggregated at each unit of the order: so, for example, if we wish to look at the hourly frequency count of a disaster term in the messages sent during the hour have to be aggregated over the whole hour and so on. The information in the messages can be scaled at lower time frequency minute-by-minute changes in frequency to hourly scale, hourly to daily scale and so forth.

The processing and quantification of text data is done under the aegis of the *Rocksteady Text and Affect Analysis system*

[6]http://protege.stanford.edu/

which makes it possible to track the time evolution of disaster terms and associated named entities. We look at text analysis at the word level, counting the occurrence of terms from each category and aggregating the count across the day for all messages posted for that day. We then generate a daily time series of named entities and affect using this approach. In our system architecture (see Figure 3) we use Rocksteady as our text engine combined with a wrapper to build a corpus of social media text, and to perform statistical analysis. Detailed grammatical analysis is carried out with a terminology and ontology analysis system CiCui[7].
The output of the system is generally in the form of a time series of disaster terms supplemented, if requested, by affect terms.

3. Results

We look at the signature of a disaster comprising affect levels and the frequency of named entities. We look at two case studies of disaster events using two different approaches. First, that of a hurricane comprising of messages from social mediate sites namely Facebook and Twitter, and second that of an earthquake using information that was crowd-sourced by the European-Mediterranean Seismological Center.

3.1. Case Study: The Impact of Hurricane Sandy

3.1.1. Data

We obtained historical data for Twitter and Facebook messages by searching the Lexis Nexis database by source for Twitter and Facebook. The Lexis Nexis database contains many social media blogs that often republish, link or summarise articles via Twitter and Facebook. In this way social media sites become a kind of echo chamber or informal location of aggregated news from social blogs, news sites and institutions. The reach and ease of dissemination of news through social media sites make it desirable for vendors to publish and inform users of articles using summarised or concise posts. This is particularly true for disaster and emergency services, companies and similar institutions who wish to quickly inform the general public of the latest or breaking news. The disadvantage of this method of information broadcasting, in particular for machine readability and interpretation, is in the noise induced through cross chatter and conversations, illegitimate sources, or invalidated information which has quickly become shared or viral (Kwak et al., 2010) (Lerman and Ghosh, 2010).
The selection of social media messages gathered from Lexis Nexis was collected initially by Newstex[8] who aggregate news and full-text feeds from authoritative, corporate, and independent publishers. The Twitter and Facebook content in our corpus consist of sources and publishers considered authoritative by Newstex and have been curated accordingly. Sources tend to be news sites, institutional and governmental bodies, companies, disaster services, emergency services, and influential social media bodies. In this way, potentially unreliable sources are omitted easing the burden of having to cope with noise information in social media text. We refined our search to only include messages that contained the term Hurricane Sandy and included all sources classified as being DISASTER & EMERGENCY AGENCIES publishing on Twitter and Facebook.
We also limit our study to the most relevant social media messages surrounding the landfall and subsequent aftermath of hurricane Sandy. The messages in our corpus are collected for the period between 22/10/2012 to 19/11/2012, the period of Hurricane Sandys formation, landfall on the 24th of October and dissipation of the storm on the 31st of October. We also include the period after the storms dissipation to take into account any reports on the aftermath of the storm and its impact. The final corpus used in much of the analysis consists of 29 days between (22/10/2012 to 19/11/2012) with roughly 1,440 items of text and 2,035,819 words.

3.1.2. A time serial reporting of the disaster

Messages containing terms that would be considered relevant to disaster terminology, mentions of the superordinate category *disaster*, reaches a peak during Hurricane Sandy's landfall on the 25th of October. Disaster related messages remained constant during the storms presence until it began to dissipate, falling dramatically on the 5th of November and subsequently returned to very few mentions. Most mentions of the storm and subsequent disaster stayed at low levels or zero by the 19th of November(Figure 4).
Shortly after the increase in chatter about the Hurricane and its landfall on the 25th of October, attention quickly changed to mentions of emergency and emergency services, reaching its peak on the 7th of November. This chatter quickly dissipates as news about causalities, victims and storm damage comes to light and takes hold of media attention. News and emergency service announcements decline dramatically leading to a surge in the mentions of casualties, victims and storm damage, reaching a peak as the impact of the storm has been assessed(Figure 5).
To see the interaction between the ontological categories in our corpus we look to correlate each time series. Each time series consists of mentions of terms for each category and aggregate for all messages in a day giving a single observation for each day. We aggregate the frequency of terms in each category, and then we make a comparison between each categorys occurrence. In this way it is possible to see if mentions of hurricanes or storms occur when negative sentiment is expressed. Correlation between categories is shown in the occurrence or absence of topics and how they interrelate (Table 1). Care is taken in the correlation analysis to ensure terms in each category are unique so no double counting occurs.
By observing the cross correlations, hurricane is more highly correlated with negative sentiment. This may suggest people express negativity frequently when mentioning the topic of the storm, this makes intuitive sense. Mentions hurricane and the storm are anti-correlated in the case of victim and result with a stronger than average correlation coefficient for this table. Occurrences and mentions of floods are also related to news about victims and casualties. Relations to negative sentiment are low however, by

[7]Both Rocksteady and Cicui were developed at Trinity College and patents have been filed for both the systems.

[8]http://newstex.com/about/what-is-authoritative-content/

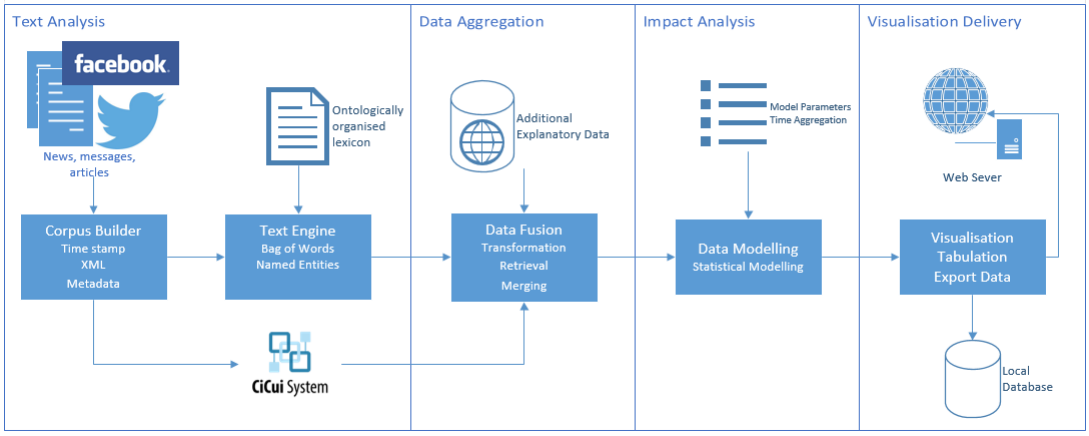


Figure 3: The general system architecture used to generate a time series of quantified text data.

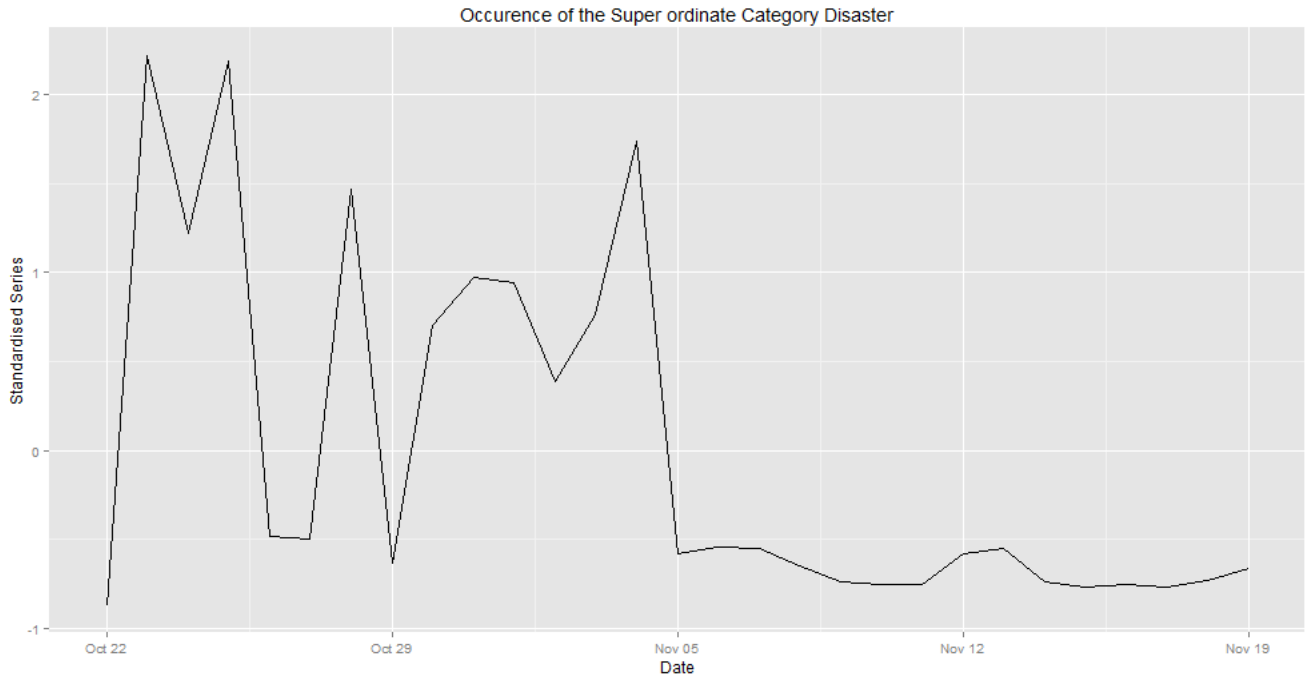


Figure 4: The relative occurrence of terms for the superordinate category Disaster.

observing our corpus, messages regarding victims and casualties are often reported as facts as opposed to sentiment.

3.2. Case Study: The Unfolding of an Earthquake

The impact of a disaster event will vary, not just on the intensity of the event, but on the location of the events occurrence and distance to urban or populated areas. People who are closer to the event will be greatly affected by its impact while the further the distance from the disaster, the less implications and impact will be experienced. This is certainly true for natural disasters such as earthquakes for instance. One can examine the ethics of this hypothesis; noting the names of locations and people may reveal their identity, we have taken care to make this information anonymous where possible.

3.2.1. Data

To examine this hypothesis we look to analyse public opinion or testimonies recorded about a 4.8 magnitude earthquake that occurred in England in 2008. The European-Mediterranean Seismological Center (EMSC) has recorded crowd sourced information about the earthquake and its impact, in particular to the specific area the individual is reporting. The testimony consists of a series of questions which determine the address or location of the observation, the effect of the earthquake on surrounding objects, buildings, animals, and people and what they were engaged in at the time of the observation. The eye witness reports consist of text describing the effect and intensity of the earthquake along with the distance the observation was made at from the epicenter. We analyse the text of the testimonials,

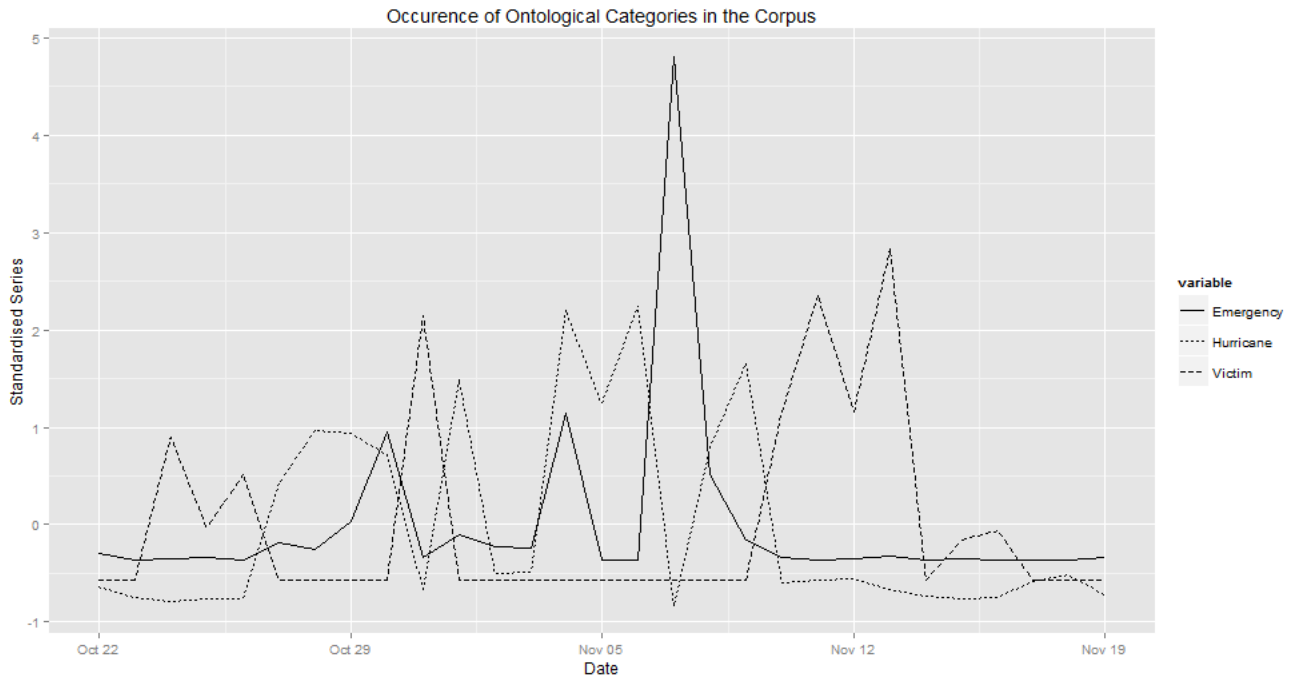


Figure 5: The relative occurrence of terms each category Emergency, Hurricane and Victim.

	Messages	Emergency	Fire	Hurricane	Flood	Victim	Negative	Positive	Weak
Messages	1	-0.12	0.01	-0.2	0.28	**0.34**	0.03	-0.06	-0.03
Emergency	-0.12	1	-0.1	0.06	-0.12	-0.2	-0.13	0.17	-0.09
Fire	0.01	-0.1	1	-0.2	-0.06	-0.01	-0.1	-0.26	-0.11
Hurricane	-0.2	0.06	-0.2	1	-0.08	**-0.38**	**0.37**	0.08	**0.39**
Negative	0.03	-0.13	-0.1	**0.37**	0.24	-0.03	1	**0.56**	0.9
Positive	-0.06	0.17	-0.26	0.08	0.18	-0.11	**0.56**	1	**0.52**
Victim	**0.34**	-0.2	-0.01	**-0.38**	0.28	1	-0.03	-0.11	-0.13
Weak	-0.03	-0.09	-0.11	**0.39**	-0.01	-0.13	**0.9**	**0.52**	1
Flood	0.28	-0.12	-0.06	-0.08	1	0.28	0.24	0.18	-0.01

Table 1: The correlation between categories in our corpus. The occurrence of certain categories with others has differing levels of correlation.

specifically calculating the negative affect of each testimonial, and record this result with the observation's distance from the epicenter.

3.2.2. The varying impact of the earthquake

We have calculated the average negative sentiment for testimonials by location in each 50 mile radius from the epicenter. The highest percentages of affect terms are present closer to the earthquakes epicenter (Table 2).

Reports of the earthquakes effect are more negative the closer the reports are to the epicenter. As the distance to the earthquakes source is increased the average negative sentiment contained in the reports decreases. The language used by observers within 50 miles of the earthquake center describes more a more violent impact[9]:

> *The noise was incredible, a deep thunderous groaning. The pipes in the house creaked under the strain.*
>
> Witness location: Lincoln (United Kingdom) (10km SW from epicenter)

Distance from epicenter	Negative affect
50 miles	7%
100 miles	6%
150 miles	5%
200 miles	3%
400 miles	3%

Table 2: The average negative sentiment of reports by members of the public describing the effect and intensity of an earthquake, organised by distance from the epicenter.

As distance increases from the epicenter, observers still experienced the earthquake but to a lesser degree:

[9] http://www.emsc-csem.org/Earthquake/Testimonies

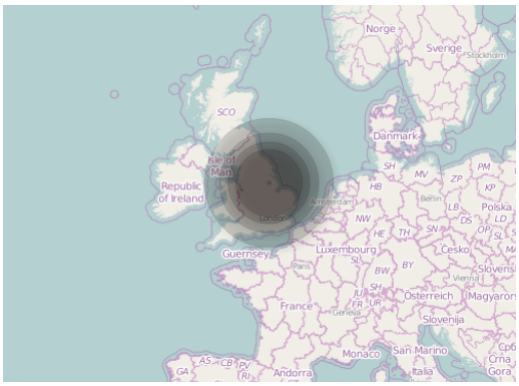


Figure 6: Earthquake epicenter (53.32N, 0.45 W) in England, reports collected of the earth quakes effect highlighted from the heatmap.

> *Woke me from sleep, could hear creaking noises from TV and other objects.*
>
> Witness location: Hoylake-West Kirby (United Kingdom) (179km W from epicenter)

The impact of the earthquake is much higher the closer to the epicenter where the influence is strongest. This is reflected in the language that observers use when describing their experience of the impact on their immediate surroundings and seen in the decreasing level of average negative sentiment used.

3.3. Information on Named Entities and an *Intrusion* Index

Within the corpus we consider the source of information and entities we are tracking and any implications that may occur from this kind of monitoring. We look at the distribution of different named entities, particularly those phrases or terms that are classified as proper nouns.

Using the Cicui system, we extracted over 27,000 proper nouns and hand labeled a sample of 1200. We classify them as either being an institution, an event, a source, an object, a legal term, a person, or a place (Table 3). The names of events, institutions, legal terminology, and object names comprise 31.3% of the hand labeled sample a total of 375 out of 1200 proper nouns. The proper nouns comprising person names and place names comprise 11.5% of the total an overwhelming number of the person names are that of people in authority generally well-known government officials for instance, and the same is true of place names. The overwhelming number of tokens labeled proper noun by the part-of-speech-tagger we used, Stanford POS Tagger, are the so-called *source* tags. These are the tags associated with Twitter messages; an in-depth analysis of these tags, with the active assistance of Twitter Inc., may help in getting closer to the identity of the person or persons who may have sent the message.

The person and place names not associated with authoritative names and locations of (disaster management) authorities will help in identifying the level of intrusion by a social media analysis system used in disaster management. There are important ethical consequences of unrestricted access, accidentally or deliberately, to a collection of named entities related to the victim-citizenry. We will be pursuing the establishment of an intrusion *index* of such an analysis system as a part of our investigation in the EU-sponsored Slaindail Project.

Category	Relative Frequency
Institution	8.4%
Event	7.9%
Source	57.1%
Object	7.1%
Legal Term	7.9%
Person	7.5%
Place	4%

Table 3: The percentage occurrence of proper nouns in our test sample. The category tags were used to label and track the occurrence of people, events and place mentions or citations.

4. Afterword

In this paper we looked at the distress content of social media messages and testimonials related to a disaster event made available to us via publicly-available collection of Twitter messages (via Lexis-Nexis News Service, available to all bona fide users of this publicly available system). We had created an ontologically organised lexicon of disaster management for broadening or narrowing our analysis of the duration and impact of a disaster event named entities and words related to action, help us to create a *signature of disaster* (see also Zhang and Ahmad in these workshop Proceedings). The ontologically organised disaster lexicon when suitably interfaced with the classic *affect lexicon*, the General Inquirer developed at Harvard University (Stone et al., 1966), can help in estimating affect related to distress and/or relief in the social media messages and testimonials. We have described an initial attempt to look at the level of personal intrusion associated with a social media analysis of digital media messages.

5. Acknowledgements

The authors would like to thank the EU sponsored Slaindail Project (FP7 Security sponsored project #6076921)

6. References

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Kristina Lerman and Rumi Ghosh. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 10:90–97.

William D Lewis, Robert Munro, and Stephan Vogel. 2011. Crisis mt: Developing a cookbook for mt in crisis situations. In *Proceedings of the Sixth Workshop on Statis-*

tical Machine Translation, pages 501–511. Association for Computational Linguistics.

Will Lewis. 2010. Haitian creole: how to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *EAMT 2010: Proceedings of the 14th Annual conference of the European Association for Machine Translation, Saint-Raphaël, France. 8pp*.

Susan McClendon and Anthony C Robinson. 2013. Leveraging geospatially-oriented social media communications in disaster response. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 5(1):22–40.

Robert Munro and Christopher D Manning. 2012. Short message communications: users, topics, and in-language processing. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, page 4. ACM.

Robert Munro. 2013. Crowdsourcing and the crisis-affected community. *Information retrieval*, 16(2):210–266.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Mel Taylor, Garrett Wells, Gwyneth Howell, and Beverley Raphael. 2012. The role of social media as psychological first aid as a support to community resilience building. *Australian Journal of Emergency Management, The*, 27(1):20.