Information Extraction of Filipino Disaster-Related Reports

A Thesis Proposal

Presented to

the Faculty of the College of Computer Studies

De La Salle University Manila

In Partial Fulfillment

of the Requirements for the Degree of

Bachelor of Science in Computer Science

by

DELA CRUZ, Kyle Mc Hale B.

GARCIA, John Paul F.

KALAW, Kristine Ma. Dominique F.

LU, Vilson E.

REGALADO, Ralph Vincent

Adviser

April 11, 2014

**Abstract**

The Philippines, being a disaster-prone country and the social media capital of the world, uses the social media to report the current status of their areas. Given these sources of information, relevant details could be extracted and utilized in order to provide more information for decision-makers. This could help the government and other institutions in deciding where to deploy relief resources. However, social media comes in different forms such as news feeds, blogs, and social networking sites. This means that it is harder to extract information due to the lack of structure. Also another problem lies with the nature of the Filipino language being morphologically rich and having other variations, thus, makes it difficult for the system to extract information. The information extraction system must be able to extract information from various sources like social networking sites, news feeds, and blogs, taking into consideration the different variations of the Filipino languages.

Keywords: Information extraction, disaster management, social media

Table of Contents

[1.0 Research Description 1-1](#_Toc384681906)

[1.1 Overview of the Current State of Technology 1-1](#_Toc384681907)

[1.2 Research Objectives 1-2](#_Toc384681908)

[1.2.1 General Objective 1-2](#_Toc384681909)

[1.2.2 Specific Objectives 1-2](#_Toc384681910)

[1.3 Scope and Limitations of the Research 1-3](#_Toc384681911)

[1.4 Significance of the Research 1-3](#_Toc384681912)

[2.0 Review of Related Works 2-5](#_Toc384681913)

[2.1 Machine Learning-Based Information Extraction Systems 2-5](#_Toc384681914)

[2.2 Rule-Based Information Extraction Systems 2-6](#_Toc384681915)

[2.3 Other Information Extraction Systems 2-8](#_Toc384681916)

[3.0 Research Methodology 3-12](#_Toc384681917)

[3.1 Investigation and Research Analysis 3-12](#_Toc384681918)

[3.2 System Design 3-12](#_Toc384681919)

[3.2.1 Sprints 3-13](#_Toc384681920)

[3.2.2 Sprint Planning Meetings 3-13](#_Toc384681921)

[3.2.3 Scrum Meetings 3-13](#_Toc384681922)

[3.3 System Development 3-13](#_Toc384681923)

[3.4 System Integration and Testing 3-13](#_Toc384681924)

[3.5 System Evaluation 3-13](#_Toc384681925)

[3.6 Documentation 3-14](#_Toc384681926)

[3.6.1 Calendar of Activities 3-15](#_Toc384681927)

[4.0 References 4-16](#_Toc384681928)

[5.0 Appendix 5-18](#_Toc384681929)

[5.1 Appendix A 5-18](#_Toc384681930)

[5.2 Appendix B 5-19](#_Toc384681931)

**List of Tables**

[Table 2‑1. Summary of reviewed information extraction systems. 2-11](#_Toc384681932)

[Table 3‑1. Timetable of Activities (April 2014 - April 2015). 3-15](#_Toc384681933)

**List of Figure**

[Figure 3‑1. Research Methodology Phases 3-12](file:///C:\Users\Vilson\SkyDrive\Gridlock\Thesis%20Proposal.docx#_Toc384681934)

# Research Description

This chapter introduces the research which will be undertaken in the field of Text Classification and Information Extraction (IE) in Natural Language Processing (NLP) for disaster management. This chapter is divided into four sections. The first section will talk about the motivations and the problem that needs to be addressed. The second section will discuss the objectives of the research. The third section will discuss the scope and limitations of the study. Lastly, the fourth section will tackle the significance of the research with regards to the Philippine society.

## Overview of the Current State of Technology

According to a report of the United Nations International Strategy for Disaster Reduction (UNISDR) Scientific and Technical Advisory Group, disasters have destroyed lives as well as livelihood across the world. Just between 2000 and 2012, about 2 million people died and an estimate of US$ 1.7 trillion of damage were sustained in disasters. In the same report, the UNISDR posits the use and research of new scientific and technological advancements in disaster management (Southgate et al., 2013).

Social media are online applications, platforms, and media which aim to facilitate interaction, collaboration and the sharing of content. In a study of Universal McCann and an analysis of 24/7 Wall St., LLC about social media, the Philippines got a high rank in most of the categories. This led to the country being dubbed as the “Social Media Capital of the World” (Universal McCann, 2008; Stockdale & McIntyre, 2011). Social media plays a vital role in disaster management. After the Haiti earthquake in 2010, numerous posts and photos were published in various social media sites. 48 hours later, the Red Cross has received a donation of US$8 million. Social media has enabled the generation of community crisis maps and interagency maps, a map that works as an intermediary between the public and relief organizations (Gao, Barbier & Goolsby, 2011). Patrick Meier, a crisis mapper, makes use of social media to improve the efficiency of relief efforts. He launched the website MicroMappers[[1]](#footnote-2), that quickly sort through online data, from tweets to uploaded photos, and then display the information on satellite maps, to assist in relief efforts during the disaster of Super Typhoon Haiyan (also called Yolanda) in the Philippines (Howard, 2013).

Social Media Monitoring for Disasters (SOMIDIA) (Cheng et al., 2011) is “a crisis mapping system that focuses on plotting authentic crisis events on an interactive map in near real time.” It uses data from news sites, blogs, web forms, SMS, and social networking sites (i.e. Facebook and Twitter) to know where the disasters are in near real-time. SOMIDIA extracts information like the type of disaster, date and time of the disaster, and the location of the disaster, so that the system can display the disasters in a map in near real-time. However, the data comes from different sources. Extracting information from different sources could be difficult for the computer as these data are unstructured. SOMIDIA needs an information extraction module to be able to collect the needed information.

Information extraction is defined as “extracting structured data from unstructured data as provided, for example, text document” (Feilmayr, 2011). Information extraction systems have been used in different fields such as clinical narratives – MedEx (Xu, & et. al., 2010), healthcare – VnHIES (Dung & Kamayama, 2007), and legal documents - Legal TRUTHS (Cheng et al., 2009). On the current implementation of SOMIDIA, it already has an information extraction system for English and Filipino texts. The problem is that SOMIDIA’s information extraction module for Filipino has difficulty in extracting the disaster and location from a given text. When compared to the information extracted from English text, the English dataset performed significantly better than the Filipino dataset (Cheng et al., 2011).

One of the main problem with information extraction when it comes to the Filipino or “Taglish” language is the characteristic of the two aforementioned languages. By nature, the Filipino language, in general, is a morphologically rich language. Morphologically rich languages (MRL) are languages that have words which are composed of the combination of the root word and a number of morpheme components epentheses (the addition of suffixes to add new meanings to the word), metatheses (the addition of suffixes that may change the spelling of the entire word), replacement (the addition of suffixes that may change the sound of certain parts of the root word), infixation (the addition of infixes splits the root word), and reduplication (repeating a combination of letters from the root word to give it a new meaning) (Stone, 2004). With these, systems that are being built for analyzing/translating the Filipino language are having problems when it comes to identifying the most concise translation of the Filipino word into other languages. But there have been researches that made use of Dependency Parsing and Part-Of-Speech (POS) Tagging to analyse a given Tagalog statement. In the Dependency Parsing approach, a dependency structure will be constructed based on the input statement and then, identify the syntactic head of the given statement from the constructed tree and lastly, analyze it by linking the head and its dependents. (Manguilimotan & Matsumoto, 2011). In the POS Tagging approach, the input statement is chopped into words and then these words are tagged or labeled according to their function in the statement and then given syntactical analysis (Manguilimotan & Matsumoto, 2009).

The existing information extraction system present in SOMIDIA has limitations when it comes to handling text because it can only handle English text. The system still has problems when handling Filipino text for the reason that the information extraction done on text written in Filipino is using keyword-based searching. Currently, there are no stable tools that can handle Filipino text (Cheng et. al., 2011). Social media data in the Philippines shows inconsistency because when Filipinos communicate or post in social networking sites they use different languages such as English, Filipino, Taglish and “TXTSPK”, especially today, when people with different language and culture communicate on the same medium (social networking sites) (Ghedin, 2011). Another problem with social media data is the existence of micro-posts, or short posts, such as tweets which do not have much contextual information and tend to be less grammatical. (Maynard et. al., 2012).

## Research Objectives

This section presents the general and specific objectives of the proposed research.

### General Objective

To develop an information extraction system that extracts relevant information from disaster-related texts from social media and takes into consideration the different available variations in the Filipino language.

### Specific Objectives

The following are the specific objectives of the research.

1. To review different information extraction systems;
2. To identify data source that will be used for the information extraction system;
3. To review different natural language processing techniques that will pre-process data for the information extraction system;
4. To review different information extraction techniques;
5. To evaluate existing tools and resources which could be incorporated in the information extraction components of the system;
6. To determine the metrics for evaluating the information extraction system;

## Scope and Limitations of the Research

The research aims to design an information extraction system for the Filipino language. It will also cover the review of various information extraction systems in order to know the different approaches on implementing them. Different existing domain-independent, domain-dependent information extraction systems will be reviewed in order to understand the architectures, implementation and components of an information extraction system. It will also review information extraction for MRL in order to understand the techniques used to extract from MRL since the Filipino language is considered to be an MRL.

The research will identify the source where the data will be collected. Example of data source would be Facebook and Twitter. Being the social media capital of the world, Filipinos post status of their lives in popular social networking sites like Facebook and microblogging sites like Twitter. There are also news from online newspaper. Identifying the data source that will be used in the information extraction will help in choosing appropriate preprocessing techniques and algorithms.

In order for the information extraction system to perform better, the research will review different natural language processing techniques that will preprocess the data before feeding it to the information extraction system. Examples of the NLP techniques that will be reviewed are text classification and text normalization. Text classification is the process of automatically assigning a text or document into a predefined category based on their content (Özsu & Liu, 2009). Texts may need to be classified according to categories so that the system can use appropriate algorithm to extract the information. Text normalization is the transforming of ill-formed words into their canonical forms (Han & Baldwin, 2011). The information extraction system will need a text normalizer as data coming from social media sites (Twitter and Facebook) are noisy. Most of the text has no structure, incorrectly spelled words, and invented terms.

The research will review different information extraction techniques that will be used for the information extraction systems. Some of the techniques that will be reviewed are Named Entity Recognition (NER), lexical analysis, and conference analysis. Lexical analysis involves splitting up sentences into words and performing Part-Of-Speech tagging to each word (Grishman, 1997). NER is the classification of each word into a category (Zhou & Su, 2002). Conference analysis is the resolving of references for the pronouns (Grishman, 1997).

Existing tools that will be used in building the information extraction system will be reviewed and evaluated. Example of NLP tools are OpenNLP and Lingpipe. OpenNLP is a machine learning based toolkit for the processing of natural language text that can support a number of common NLP tasks like tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution (Apache Software Foundation, 2010). On the other hand, Lingpipe is a toolkit for processing text using computational linguistics that can perform certain tasks like finding names of people/organizations/event, classify Twitter data, and check spellings (Alias-I, 2011).

In order to evaluate the information extraction system, the research will determine the metrics to measure the system’s performance.

## Significance of the Research

Being the social media capital of the world, the Philippines generates a lot of diversified information that cannot be easily tapped because of the limited capabilities and tools that are available in processing the language unto which these information are written in, the Filipino language. With an information extraction system that is specifically built for the Filipino language, people can explore more possibilities and opportunities with regards to effectively utilizing this information from the web.

In the disaster management standpoint, respective stakeholders can collect disaster-related information in a way that is less strict because with an information extraction system specifically built for the Filipino language, these stakeholders can effortlessly accept and process information that are written in a much more natural and open way. With this, they can reach out to more people and to more places because they can have a system that understands how Filipinos speak and communicate through the different social media platforms available.

In the local business standpoint, respective stakeholders can gain valuable information from their customers. These stakeholders could use social media sites to collect information about their customers’ respective preferences about certain products, brands or services. With this, local businesses can effectively reach to their customers’ wants and need because they can have a system that understands how customers react to certain products, brands or services when they are in the different social media platforms available.

In the case of SOMIDIA, improving the on-board information extraction system can further enhance the usability and accuracy of the information presented in the existing SOMIDIA system. With an enhanced and improved information extraction system, SOMIDIA can accept and process information that are written in a more open and unstructured way, or simply, information that are written in the format of the different variations in the Filipino language like the ‘TXTSPK’, ‘Jejemon’ and the ‘Bekimon’. With an improved information extraction algorithm, the new information extraction system will be able to increase the probability of accurately and precisely understanding the normal Filipino language and at the same time, it will include support for the different variations like the ‘TXTSPK’, ‘Jejemon’ or ‘Bekimon’ since most of the information that will be used by the SOMIDIA system are taken from the different social media platforms and they are written in a very open and informal way.

# Review of Related Works

This chapter discusses the features capabilities, and limitations of existing research, algorithms, or software that are related or similar to the thesis.

## Machine Learning-Based Information Extraction Systems

This part discusses information extraction systems that use machine learning-based techniques.

#### Machine Learning for Information Extraction in Informal Domains (Freitag, 2000)

The researchers of the paper explored one of variation of the slot-filling problem and that is to find the best unbroken fragment of text to fill a given slot in the answer template. There is a definite template that is given to an IE task. The template consists of fields that need to be filled with instances from the text source. The researchers set two ways of simplifying how to study the behaviour of the algorithms to be developed: to isolate each field learning problem and focus on fields that is not instantiated or have a unique instance in a text source. With this, they found two primary aspects: multi-strategy learning and feature engineering. Multi-strategy learning because they believed that there is no single representation for all IE problems. Feature engineering because ML of a feature set is needed to help adapt to domains containing novel structures since they will target informal domains. The researchers used four ML components: rote learning, term-space learning, learning abstract structure with grammatical inference, and relational learning for information extraction. They did experiments to gauge the performance of the four learners.

To conclude, the researchers found out that it is possible to perform IE from informal domains found in the internet. Also, they stated that ML is a rich source of ideas for different algorithms that can be trained to perform IE. They have shown that with the right ML techniques it is possible to train effective extractors with very simple document representations.

#### TOPO - Information Extraction System for Natural Disaster Reports from Spanish Newspaper Article (Téllez-Valero, 2005)

This information extraction system extracts information related to natural disasters from newspaper articles written in Spanish. The system extracts the following information: (1) information related to the disaster itself (date, place and magnitude), (2) information related to buildings (number of destroyed buildings, affected houses), (3) information related to people (number of dead, missing or wounded), (4) information related to infrastructure (number of affected hectares, economic lost). It is able to extract information on natural disaster like hurricanes, forest fires, inundations, droughts and earthquake.

The system uses general information extraction system architecture. First, the document is turned into Boolean vectors representing the presence and absence of certain words. This is the document feature extraction stage. In order to limit the dimension, they used information gain technique. After it is turned into a Boolean vector, it will now be classified. They used Support Vector Machine (SVM), Naïve Bayes (NB), C4.5, k-Nearest Neighbours (kNN). After it has been classified, it needed to select text that might contain relevant information. This is the candidate text selection stage. They used grammar to select the text and a dictionary of names and number to treat grammar exceptions. Then the output will be candidates of relevant information. Then, the system will now select which of the information will be used. This uses the same algorithms in the text classification stage. They used different classifier for different output.

This architecture boasts its portability because it is language independent and domain adaptive. It is language independent because its training features and candidate text segments are based on simple lexical rules. It is domain adaptive because it only needed to change the training corpus.

The text filtering stage was evaluated on 134 news reports on the metrics of precision, recall and F-measure. The algorithm that produced the best result was the SVM. They got an F-measure from 72% to 88% on classification of news reports. The information extraction stage was evaluated on 1353 text segments that consist of names, dates and quantities randomly taken from 365 news reports. The best classifier for name and quantities was SVM, while kNN for dates. The overall system got an average of 72% on F-Measure.

## Rule-Based Information Extraction Systems

This part discusses information extraction systems that use rule-based techniques.

#### Vietnamese Real Estate (VRE) Information Extraction (Pham & Pham, 2012)

The Vietnamese Real Estate (VRE) Information Extraction system extracts information from Vietnamese Real Estate Advertisements. It collects information like the type of estate, category of the estate, area, zone, price, name of the author, and contact details. The system uses the GATE framework for its architecture.

For its data, it has to pass certain criteria before it is fed into the system. First, it must be news articles related to real estate advertisement. Second, only one advertisement from each input data file. Lastly, it must be strip off of all its HTML tags. After the data has met all the criteria, it will now go to data normalization first. The data normalization helps reduce ambiguity and helps the human in annotation. First, it must add the necessary punctuation at the end of the sentence. Second, it merges multiple paragraphs into one. Third, normalize the punctuations, remove redundant spaces and capitalizes the first character after each punctuation. Then lastly, normalize the telephone, price, area and zone to a common pattern. After the data is normalized, it will now be manually annotated using Callisto, an annotation software.

After it has been annotated, it is now ready to go to the information extraction system. It will go first through the tokenizer. The tokenizer will output two types of annotations, Word and Split. The Word annotations contains the part-of-speech, the word, checks if the first letter is capitalized, and other features (kind and nation). This will be used to create the Jolly and Pleasant Experience (JAPE) rules. The Split annotation contains the delimiter. After it goes through the tokenizer, it will now go through the Gazetteer. The gazetteers are dictionaries that are created during the system development. It contains dictionaries for potential named entities (person, location) or categories, phrases uses in contextual rules (name prefix or verbs that are likely to follow a person’s name), and potential ambiguous entities. The output of the gazetteer is a lookup annotation covering the specific semantics. After the gazetteer, it will now be passed to the JAPE transducer. The JAPE transducer is responsible for extracting the information. It uses JAPE rules to recognize the entities that will be needed to extract. The output is the annotated documents.

The system has been tested in a lenient and strict criterion. An entity that is recognized correctly when the type is correct but the span overlap in the annotated corpus is called the lenient criteria. On the other hand, an entity that is recognized correctly when the type and span are the same in the annotated corpus is called strict criteria. On the lenient criteria on test data, it measured 96% in F-measure. While on the strict criteria, it measured 91% in F-measure. The problem is on the data. The writing styles of the people are very diverse. The system has problem in recognizing some of the entities like the zone entity because some of the zone entity are very long and does not use capitalization.

#### Business Specific Online Information Extraction from German Websites (Lee & Geierhos, 2009)

The Business Specific Online Information Extraction System is a system that extracts information from the information pages of a German business website like its company profile, contact page, imprint and then identifies relevant business specific information. The system concentrates on the extraction of specific business information like company names, addresses, contact details, names of CEOs, etc. With regards to the way how the researchers pre-process their chosen input data, they interpret the HTML structure of documents and analyse some contextual facts to transform the unstructured web pages into structured forms. The approach applied by the researchers is quite robust in variability of the DOM (for the web pages), upgradeable and keeps data up-to-date. The evaluation metrics showed high efficiency of information access to the generated data. In their conclusion, they stated that the developed technique is also adaptive to non-German websites with slight language-specific modifications, and experimental results from real-life websites confirm the feasibility of their approach.

In their proposed system, the researchers had two main modules for processing and extracting information from the German Information Web Pages: one for establishing a relational database storing company information and the other is for providing a query module. Within these two modules are three sub process that are done to further process the input data: (A) Localization of the Information Pages on the Web; (B) Document Analysis and Information Extraction; lastly, (C) Query Processing. In sub process A (Localization of the Information Page), a web crawler is fed with the URL’s of the web pages that are stored in the specialized database and then it fetches them from the web. Afterwards, the proposed system will then retrieve the document by following the anchor tags that lead to the information pages. On the other hand, in sub process B (Document Analysis and Information Extraction), the fetched Information Pages are sent to an ‘info analyser’ module which examines the HTML content of the page and then extracts the needed information bits. Here, the system exploits the internal structure of the named entities and uses sublanguage-specific contexts or attribute classes to identify the attribute-value pairs. Lastly, in sub process C, the user of the system is given the right to query the database for the information bits that he/she needs and then add these bits to the index.

For the Information Page Analyser (info analyser) in sub process B, the input data has to further go through a number of processes to finally extract the information needed by the user. When given an Information Page, the analyser starts by pre-processing the frame structure and existing JavaScript of the page. And before creating the expressive DOM Tree, the HTML file of the page has to be validated and corrected, if needed, by using a special tool called ‘tidy’. After doing so, the system will now be able to locate the minimal data region (or the data region of the information bit searched for) surrounded by a number of HTML tags which contain the information record being searched for. By doing a depth- first traversal of the expressive DOM tree, the desired sub tree can be isolated based on the headings of the data record like the following: “Herausgeber” (publisher), “Betreiber” (operator), “Anbieter” (provider) and etc. The system was programmed to disregard domain name irrelevant information; thus, the analyser will work further on with a pruned DOM tree. After identifying the minimal data region, all information bits that are relevant to the domain name are extracted by using the Named-Entity Recognition technique and the attribute-value process (each attribute has a corresponding value that is indicated by the structure of the HTML file it is in) with respect to its external contexts and internal features. The system’s analyser module considers about 20 attribute classes and searches their corresponding values on the information page of business websites. The following are some of the attribute classes that are considered by the analyser: company name, address, phone and fax number, e-mail, CEO, management board, domain owner, contact person, register court, financial office, register number, value added tax number (VAT ID), and etc. After extracting the information bits needed from the pruned DOM trees, the information bits are then normalized to make sure that all information are consistent. The following are the classes that are affected by the normalization process: company names, legal form, register number, address (street, zip code, city), contact (phone and fax number, email), person name, and legal notification (tax number, VAT ID).

To conclude, the system performed surprisingly accurate with an average precision score of 99.1% and a recall score of 91.3% from a small test corpus that is composed of approximately 150 business web pages. The only encountered problem by the system is when value for certain attributes is erroneously represented like text in phone numbers and etc.

#### Ontology-Based Information Extraction (OBIE) System for French Newspaper Articles (Nebhi, 2012)

Since most of the information extraction systems are based on the English language, it poses a problem on other languages for there are not much tools available. In order to address this problem, the system maps the extracted entities to the ontology.

This system extracts person, location and organization on a French newspaper article. It collects data from LeMonde.fr. The system uses the GATE framework for annotation of entities in text and maps them to the ontology. It uses DBpedia databank that is based on Wikipedia projects. It contains 3,220,000 instances and is organized into a hierarchy of 320 classes and 1650 different properties. The system consists of 4 parts: pre-processing, gazetteer, rule-based semantic annotation, and the output. First, the system will pre-process the text. It will perform tokenizer, sentence splitter and POS tagger using the GATE application. After it is pre-processed, it will now go to the gazetteer. It will perform a lookup for the named entity recognition. After it passes through the gazetteer, grammar rules will be applied to create semantic annotation. The rules are written in JAPE which is part of the GATE framework. The system contains approximately 100 rules.

The system is evaluated using the Balance Distance Metrics (BDM) to take account the ontological similarity. It is also evaluated with the gold standards. They manually annotated the documents using concepts on DBpedia ontology, and then compared it with the gold standard. They only evaluated person, organization and location named entities. The system scored an average of 0.94 on the BDM and achieved a 91% F-Measure.

## Other Information Extraction Systems

This part discusses information extraction systems that use other techniques.

**SOMIDIA - Social Monitoring for Disaster Management (Cheng et al. 2011)**

SOMIDIA is a crisis mapping system that focuses on plotting disaster on an interactive map in near real time. SOMIDIA collects data from different sources like news feeds, posts, SMS, blogs and microblogs. One of the main component of SOMIDIA is its information extraction module. It extracts from both Filipino and English texts.

For the information extraction module, first, documents go through a tokenizer. They used OpenNLP to tokenize the document, then it will go to a sentence splitter. The sentence splitter accepts the list of tokens and annotation list. It has a list of abbreviations so that the system can distinguish abbreviation periods from a period. The goal of the sentence splitter is to separate sentences by adding appropriate ending markers (period). The system used OpenNLP’s sentence splitter for its sentence detection. After the document has been split into sentences, it will go through a language guesser. They needed to differentiate English text from Filipino text because the language has different extraction techniques. They used frequency distribution of the words to detect the language. The output of the language guesser is the document with added metadata of the language. If the text is in English, the language guesser will passed the document to the POS tagger. Else, it would be passed to a Filipino NER.

For the English information extraction module, first it will go through the POS tagger. It uses the OpenNLP’s POS tagger function. The output is a list of token with its corresponding POS tags. After the POS tagger, it will go through a chunker. The chunker groups the tokens to their corresponding part-of-speech tag. This will be used to determine noun and verb phrases. It uses OpenNLP’s noun and verb chunker. After chunking, it will passed through the English NER. The NER only focuses on proper nouns. It uses LingPipe because of flexibility. LingPipe’s NER uses three types of approaches, dictionary-based, rule-based and statistic based approaches. After the NER, it will go through coreference resolution. The coreference resolution will find the noun counterpart of the pronouns. It uses Russian Mitkov algorithm for the resolution and WordNet for the lexicon. The normalization (standardizing data, collapsing of same sentences) will be done in this phase. The last step is now the information extraction phase. It uses JAPE rules to extract the information. The rules are paired with the two-tiered bootstrapping algorithm. The first tier bootstrapping algorithm starts with a small seed of words or rules. Then from the seed, it will try to learn the extraction pattern. The learned pattern will be used to generate new extraction pattern. The process will then repeat. The second-tiered bootstrap is responsible for keeping the most relevant extraction pattern.

For the Filipino extraction module, the document will go through the Filipino NER. They created their own NER because there are no existing Filipino NER tool. It uses a dictionary-based and rule-based approaches for their NER. After it has been tagged, it will now go through the Filipino extractor, the Filipino extractor has pre-defined rules (e.g. <event> sa <location>) that will extract the needed information.

The system is evaluated using precision, recall and F-measure. They evaluated it on Tweets and news feeds. For English tweets, it scored a 75.17% F-measure on extracting disaster and 62.83% on extracting location. For Filipino Tweets, it scored 82.13% F-measure on disaster and 56.32% on extracting location. For news feeds, it scored 45.40% F-measure on English news feeds, while 38.82% on Filipino news feeds. The tweets scored higher because it is much easier to extract patterns on shorter text. The needed information will most likely be located near the text. On longer texts, the information needed might be located far away.

Table 2-1 shows a summary of all the reviewed information extraction system. The table indicates the system name, the language and type of data it can extract, the domain, NLP pre-processing techniques, information extraction techniques, and evaluation metrics used by the system.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **System** | **Language** | **Type of Data** | **Domain** | **Pre-processing Techniques** | **Information Extraction Techniques** | **Evaluation Metrics** |
| Machine Learning for Information Extraction in Informal Domains  (Freitag, 2000) | N/A | Documents (i.e. email) | Informal Domain | Not mentioned | Machine Learning-  Based | Precision, Recall |
| TOPO - Information Extraction System for Natural Disaster Reports From Spanish Newspaper Article (Téllez-Valero, 2005) | Spanish | Free-text | Natural Disasters | Text Classification, Document Feature Extraction | Machine Learning- Based | Precision,  Recall,  F-measure |
| VRE Information Extraction System (Pham & Pham, 2012) | Vietnamese | Free text | Real Estate Advertisement | Text Normalization | Rule-Based | Precision,  Recall,  F-measure |
| Business Specific Online Information Extraction from German Websites  (Lee & Geierhos, 2009) | German | Structured Text | Business Specific Information | Named Entity Recognition, Text Normalization, Attribute-Value Process | Rule-Based | Precision, Recall |
| Ontology-Based Information Extraction (OBIE) System  (Nebhi, 2012) | French | Free text | News article | Tokenization, POS Tagging, Sentence Splitter | Rule-Based, Ontology | Precision, Recall,  F-measure,  BDM |
| Social Monitoring for Disaster Management  (Cheng et al. 2011) | English, Filipino | Free text | News article, tweets | Tokenization, Sentence Splitter, Language Guesser | Machine-Learning Based | Precision, Recall, F-measure |

Table 2‑1. Summary of reviewed information extraction systems.

# Research Methodology

This section discusses the different activities that will be performed throughout the research. Scrum-based methodology, an iterative software development life cycle, will be applied in the course of this research in order to ensure that the research will be able to adapt to changes in requirements.

Figure 3-1 shows a diagram of the phases the research will undergo. The phases are as follows: investigation and research analysis, system design, system development, system integration and testing, system evaluation, and documentation. Regular consultation with the thesis adviser will also be conducted in order to keep the research on track for the whole duration of the thesis.

Documentation

Figure ‑. Research Methodology Phases

## Investigation and Research Analysis

This phase involves the study and understanding of the fundamental knowledge of the concepts, algorithms, techniques, and tools which can be used to implement the system as well as identifying the modules and requirements of the system to be developed. The main key activity involved in this phase is various literature reviews of related works. From those related works, the pre-processing techniques, information extraction techniques, tools, and evaluation metrics used are identified. The listed techniques, tools, and metrics are then compared and evaluated to see which ones can be adopted to the system.

## System Design

In this phase, the system will be designed according to the information gathered during the course of the Investigation and Research Analysis phase. It is in this phase where appropriate architectures, algorithms, information extraction techniques, and other necessary tools shall be identified so that they can be effectively utilized in the making of the system. Also, it is in this phase where necessary modules for the system will be identified based on the different processes and features that will be built into the system. This phase will cover the designs of the User Interfaces and the basic architecture for the databases that will store the data that will be gathered and used by the system. Finally, this phase will also cover the identification of the source of the data that will be used and processed by the system. And once the data sources have been identified, data collection will immediately commence.

### Sprints

A two-week timeframe for each sprint will be used. This is to ensure that there is progress in the research. Each member is expected to produce a working output based on the tasks assigned to him during the sprint planning meetings. The tasks may vary from developing a part of the system or to conduct further study regarding a certain concept.

### Sprint Planning Meetings

At the beginning of each sprint, a sprint planning meeting is conducted. Tasks that must be accomplished for the current sprint will be discussed here. Included in these meeting is the assignment and division of the tasks among the members of the team. Also, the evaluation of the tasks in the previous sprint are done here. If there are any unmet tasks, these will be carried over to the next sprint.

### Scrum Meetings

Scrum meetings of 10-15 minutes in duration will be conducted daily. The purpose of this is to update each member what has or has not been accomplished yet in the assigned task. This ensures that there is daily progress and if there are issues that hinders a member from accomplishing his assigned task.

## System Development

In this phase, actual developmentof the system will be done. It will follow the design made during the System Design phase. Data collection will also be done in this phase. Each member of the team will be assigned to modules. The development of the system will follow a scrum-based methodology wherein the system is developed in an iterative manner. Daily and weekly meetings, as well as regular consultations with the adviser, are conducted in order to assess the progress of the thesis and to plan the succeeding tasks.

## System Integration and Testing

In this phase, all the modules that have been developed during the System Development phase will be integrated into one system. This phase will cover unit testing processes for each module to ensure that there will be no significant bugs that can be found after integration processes are completed. After finishing the integration process, the system will be subjected to another round of tests to check for any faulty integration and bugs that may have arose during the integration process.

## System Evaluation

In this phase, the system's performance will be evaluated based on the metrics that were chosen. As of the moment, the metrics that will be used in this phase will be the Precision, Recall and F-measure results of the information extracted by the system. The information that were extracted by the system will be subjected to a number of tests that will test its Precision, Recall and F-measure when compared to the information that were extracted manually and to those that are extracted from the training set. Although, the set of metrics that will be used might change during the course of the research as these metrics will be modified to fit the needs in accurately measuring the performance of the system to be developed.

## Documentation

Every activity or methodology that is performed will be fully documented so that they can be monitored when it comes to the modifications and progress that are made in accomplishing the documents and the system proposed in this research. Also, the documentation will be used for further references, in case there is a need to validate or cross-reference any future work that is in mind.

### Calendar of Activities

Table 3-1 shows a Gantt chart of the activities for the thesis period. Each bullet represents one week worth of activities

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Activities** | **Apr**  **(2014)** | **May** | **Jun** | **Jul** | **Aug** | **Sept** | **Oct** | **Nov** | **Dec** | **Jan** | **Feb** | **Mar** | **Apr**  **(2015)** |
| Investigation and Research | \_ \* \_ \* | \_ \_ \* \* | \_ \* \* \* | \* \* \* \* | \* \* \* \* |  |  |  |  |  |  |  |  |
| System Design |  |  |  |  | \_ \_ \* \* | \_ \* \* \* | \* \* \* \* | \* \* \* \* | \* \* \_ \_ | \_ \* \* \* | \* \* \_ \_ |  |  |
| System Development |  |  |  |  | \_ \_ \* \* | \_ \* \* \* | \* \* \* \* | \* \* \* \* | \* \* \_ \_ | \_ \* \* \* | \* \* \_ \_ |  |  |
| System Integration and Testing |  |  |  |  | \_ \_ \* \* | \_ \* \* \* | \* \* \* \* | \* \* \* \* | \* \* \_ \_ | \_ \* \* \* | \* \* \_ \_ |  |  |
| System Evaluation |  |  |  |  | \_ \_ \* \* | \_ \* \* \* | \* \* \* \* | \* \* \* \* | \* \* \_ \_ | \_ \* \* \* | \* \* \_ \_ | \* \* \* \* | \* \_ \_ \_ |
| Documentation | \_ \* \_ \* | \_ \_ \* \* | \_ \* \* \* | \* \* \* \* | \* \* \* \* | \* \* \* \* | \* \* \* \* | \* \* \* \* | \* \* \_ \_ | \_ \* \* \* | \* \* \* \* | \* \* \* \* | \* \_ \_ \_ |

Table 3‑1. Timetable of Activities (April 2014 - April 2015).

# References

alias-i. (2011). *What is lingpipe?*. Retrieved from *http://alias-i.com/lingpipe/*

Apache Software Foundation. (2010). *Welcome to apache opennlp*. Retrieved from https://opennlp.apache.org/

Asahara, M., & Matsumoto, Y. (2003, May). Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 8-15). Association for Computational Linguistics.

Cheng, H., Chua, J., Co, J., & Magpantay, A. B. (2011). *Social media monitoring for disasters*. Unpublished undergraduate thesis, De La Salle University, Manila, Philippines.

Cheng, T. T., Cua, J. L., Tan, M. D., Yao, K. G., & Roxas, R. E. (2009, October). Information

extraction from legal documents. In *Natural Language Processing, 2009. NLP'09. Eighth*

*International Symposium on* (pp. 157-162). IEEE.

Dung, T. Q., & Kameyama, W. (2007, March). A proposal of ontology-based health care information extraction system: Vnhies. In *Research, Innovation and Vision for the Future, 2007 IEEE International Conference on* (pp. 1-7). IEEE.

Feilmayr, C. (2011, August). Text Mining-Supported Information Extraction: An Extended Methodology for Developing Information Extraction Systems. In*Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on* (pp. 217-221). IEEE.

Freitag, D. (2000). Machine learning for information extraction in informal domains. Machine learning, 39(2-3), 169-202.

Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *Intelligent Systems, IEEE*, *26*(3), 10-14. doi: 10.1109/MIS.2011.52

Ghedin, G. (2011, November 16). *A social media lesson. from the philippines.* Retrieved from <http://www.youngdigitallab.com/en/social-media/a-social-media-lesson-from-the-philippines>

Grishman, R. (1997). Information extraction: Techniques and challenges. In*Information Extraction A Multidisciplinary Approach to an Emerging Information Technology* (pp. 10-27). Springer Berlin Heidelberg.

Han, B., & Baldwin, T. (2011, June). Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 368-378). Association for Computational Linguistics.

Intarapaiboon, P., Nantajeewarawat, E., & Theeramunkong, T. (2009). Information extraction from Thai text with unknown phrase boundaries. In*Advances in Knowledge Discovery and Data Mining* (pp. 525-532). Springer Berlin Heidelberg.

Lee, Y. S., & Geierhos, M. (2009). *Business specific online information extraction from german websites*. In Gelbukh, A. (Eds.), *CICLing* (pp. 369-381). Germany: Springer-Verlag Berlin Heidelberg.

Manguilimotan, E., & Matsumoto, Y. (2009). *Factors affecting part-of-speech tagging for tagalog*. In *PACLIC* (pp. 763-770).

Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Proceedings of@ NLP can u tag# user\_generated\_content*.

McCallum, A. (2005). Information extraction: Distilling structured data from unstructured text. *Queue*, *3*(9), 48-57.

Meier, P. (2013, September 18). [Web log message]. Retrieved from http://irevolution.net/2013/09/18/micromappers/

N´edellec C., Nazarenko A., & Bossy R. (2009). Information extraction. In S. Staab & R. Studer (Eds), *Handbook on ontologies* (pp 683-685). Dordecht: Springer.

Nebhi, K. (2012). Ontology-Based information extraction for french newspaper articles. In *KI 2012: Advances in Artificial Intelligence* (pp. 237-240). Springer Berlin Heidelberg.

[Özsu, M. T., & Liu, L. (2009). Text Categorization. *Encyclopedia of database systems* (p. 3044). New York: Springer.](http://www.bibme.org/)

Pham, L. V., & Pham, S. B. (2012, August). Information Extraction for Vietnamese Real Estate Advertisements. In *Knowledge and Systems Engineering (KSE), 2012 Fourth International Conference on* (pp. 181-186). IEEE.

Southgate, R., Roth, C., Schneider, J., Shi, P., Onishi, T., Wengner, D., Amman, W., Ogallo, L., Beddington J., & Murray, V. (2013). *Using science for disaster risk reduction*. Retrieved from www.preventionweb.net/go/scitech

Stockdale, C. & McIntyre, D.A. (2011, May 09). *The ten nations where facebook rules the internet*. Retrieved from <http://247wallst.com/technology-3/2011/05/09/the-ten-nations-where-facebook-rules-the-internet/>

Stone, R. (2004). Natural language processing challenges and advantages for philippine languages*. Proceedings from 1st Natural Language Processing Research Symposium (pp.81-86).*

Téllez-Valero, A., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2005). A machine learning

approach to information extraction. In *Computational Linguistics and Intelligent Text*

*Processing* (pp. 539-547). Springer Berlin Heidelberg.

Universal McCann. (2008). *Power to the people: Social media tracker wave 3*. Retrieved from <http://web.archive.org/web/20080921002044/http://www.universalmccann.com/Assets/wave_3_20080403093750.pdf>

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, *17*(1), 19-24.

Zhou, G., & Su, J. (2002, July). Named entity recognition using an HMM-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 473-480). Association for Computational Linguistics.

Zhou, S., Ling, T. W., Guan, J., Hu, J., & Zhou, A. (2003, March). Fast text classification: a training-corpus pruning based approach. In *Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings. Eighth International Conference on* (pp. 127-136). IEEE.

# Appendix

## Appendix A

|  |  |  |
| --- | --- | --- |
| **Category** | **Rank** | **Margin from Rank 1** |
| Blog Readership | 2 (90.3%) | 1.8% (South Korea) |
| Starting a Blog | 4 (65.8%) | 5.9% (South Korea) |
| Social Networks | 1 (83.1%) | -- |
| Photo Sharing | 1 (86.4%) | -- |
| Uploading Videos | 2 (60.5%) | 7.8% (China) |
| Watching Videos | 1 (98.6%) | -- |
| Podcasts | 5 (61.8%) | 12.5% (China) |
| RSS | 6 (45.2%) | 11.4% (Russia) |

*Table 5-1. Results of the study conducted by Universal McCann.*

## Appendix B

|  |  |  |  |
| --- | --- | --- | --- |
| **Morpheme Element** | **Root Word** | **Suffix** | **Filipino Word** |
| Elision | bigay | na- ; -an | nabigyan |
| Epenthesis | patay | -an | patayan |
| Metathesis | peteh (cebuano) | -en | pehten |
| Replacement | utos | -an | utusan |
| Nasal Assimilation | bigay | paN- | pamigay |
| Infixation | kain | -um- | kumain |
| Reduplication | matamis | - | matamistamis |

*Table 5-2. Examples of Filipino Morphemes.*

1. MicroMappers digital disaster response system. http://micromappers.com/ [↑](#footnote-ref-2)