# Information Extraction and Opinion Organization for an e-Legislation Framework for the Philippine Senate

**Allan Borra** **Charibeth Cheng** **Rachel E. O. Roxas** **Sherwin Ona**
Center for Language Technologies &
Center for ICT for Development
College of Computer Studies
De La Salle University, Manila, Philippines
{borgz.borra,chari.cheng,rachel.roxas,sherwin.ona}@delasalle.ph

## Abstract

This paper outlines the Language Technologies (LT) used for an e-Legislation Framework prototyped for the Philippine Senate's Committee on Accountability of Public Officials and Investigations (or Blue Ribbon Committee). The e-Legislation system uses an e-Participation framework of having both top-down (or government providing information to citizenry) and ground-up empowerment (or citizens participation). The Language Technologies employed manage the information obfuscated in unstructured text coming from both directions mainly for the purpose of policy-making. The top-down component utilizes a conventional Document Management System augmented with Information Extraction that allows for better and almost instantaneous management of information from uploaded documents. The ground-up component uses an online forum scheme and augmented with Automatic Opinion Classification and Clustering. Both e-Participation framework components (top-down and ground-up) are integrated in a single portal. This paper focuses on the technical issues of the language technologies used: information extraction and opinion classification with data clustering. Preliminary testing and results are discussed to which the information extraction performed 95.42% accuracy while the opinion organization consisting of the detection, classification and clustering modules have accuracy rates of 50%, 50.5% and 53.85%, respectively.

## 1 Introduction

The increase use of ICT in different sectors makes it a viable medium for e-Government and e-Participation. Macintosh (2007) outlined an e-Participation framework as shown in Figure 1. As shown, e-Participation has two aspects: top-down and ground-up. The interplay of the two components is vital in sustaining the whole e-Participation framework. Transparency and pushing of information of government empowers citizenry to participate. Empowered citizenry's active participation may lead to good government and governance, as well as towards crafting of more pertinent policies. The main medium between these two components is texts and language that resemble in conversations and documents. The main goal is to structure the information from unstructured text coming from both directions.

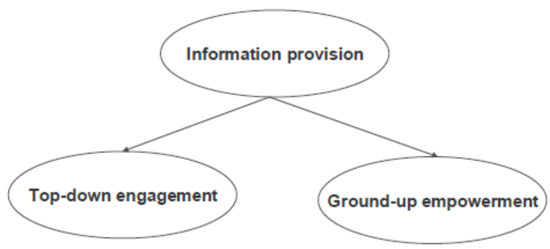


**Figure 1. e-Participation Framework by Macintosh (2007)**

An e-Legislation portal was, thus, developed to have both components of top-down engagement and ground-up empowerment involved. Describing Figure 2, an open-source Document Management System leverages the top-down (government to grassroots) pushing of information to citizenry, while an online forum scheme was implemented to leverage the ground-up empowerment by allowing for citizenry (or netizens) to actively participate, post opinions and comments, and interact with government and other citizenry. As documents and information being pushed, as well as netizens' opinions and comments, increase in size and magnitude, information get obfuscated more easily, especially since the main sources of information are in texts within documents, comments and opinions.



Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

These reiterate the need to structure the information found in these unstructured texts coming from both components. This is where the portal utilize Language Technology tools to augment the two components and structure the information and open possibility to facilitate policy-making and interaction, information retrieval, and may even open up creation of new information from the structured data.

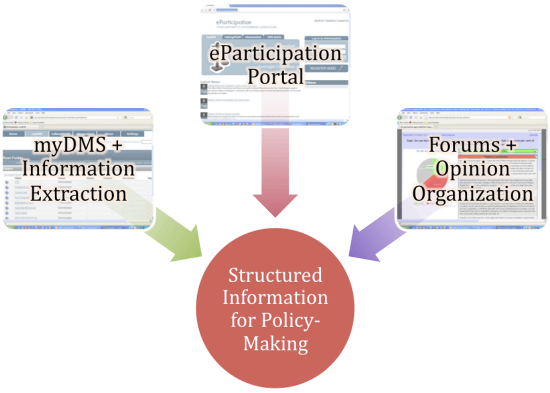


**Figure 2. ICT Tools for e-Participation Utilizing Language Technology**

## 2 Information Extraction + Document Management System = Knowledge Management

Currently, there is no electronic document management system infrastructure in the Blue Ribbon Committee (BRC), specifically in the Blue Ribbon Oversight Office Management (BROOM) of the Philippine Senate. In fact, only one person is tasked and knowledgeable of the agency's documents filing, cataloguing and retrieval procedures. The study, therefore, had a layer of modeling the business rules and process to implement the document management system (shown in Figure 3) before information extraction research were conducted and implemented. Although the whole experience of the business process modeling is a very related discourse, the focus of this section is on the information extraction and the technical aspect of the technology.

Information extraction is the process of transforming unstructured information of documents into a structured database of structured information. The underlying architecture is based on Hobb's (1993) Architecture: text zoning, pre-processing, filtering, pre-parsing, parsing, fragment combination, semantic interpretation, lexical disambiguation, co-reference resolution, and template generation. Modifications to the architecture, such as the sequence and functions of modules, were done to address the idiosyncrasies of the documents available in the government agency.

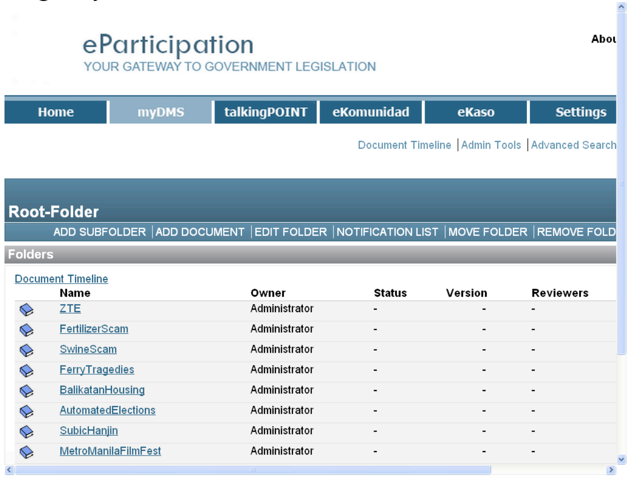


**Figure 3. Document Management System Screenshot**

### 2.1 System Architecture

The e-Legislation information extraction architecture can process different types of Blue Ribbon Committee documents. Although the Document Management System can handle any file types, only documents that manifest a regular format are processed. These documents are hearing highlights, hearing invitations, senate memorandums, documented evidences, requested administrative documents, complaints, endorsements, referrals, notification or notice of hearings and committee reports. As a result, the system handles different templates for each type of document. Considering this, the system's semantics would still be able to differentiate what template to use for a specific document.

Figure 4 shows the architecture of the system. The modules are the preprocessor, pre-parser, semantic tagger, co-reference resolution, template filler and evaluation. Under the preprocessor, there are 6 submodules: tokenizer, sentence splitter, cross-reference, part of speech tagger, unknown word and named entity recognition.

In a nutshell, a document undergoes initially the Pre-processing stages which undergoes:

1. Sentence Splitting, which removes headers and breaks down input document into series of sentences;
2. Tokenizing, which simply breaks down sentences and detects word boundaries;
3. Cross-referencing, which further sifts through the tokens and looks for entities (names) in the sentences following Schwartz and Hearst (2003) acronym detection;

4. Part-Of-Speech (POS) Tagging, which annotates the tokens within the sentence with appropriate Part of Speech Tags using LingPipe (2003);
5. Unknown-Word Detection, which classifies words that are unclassified or unknown from the POS Tagger process. It uses the ANNIE POS Tagger (1995) to represent the unknown words and classify them for post processing; and
6. Named Entity Recognition, which uses LingPipe's (2003) Dictionary Mapping named entity recognition or a look-up table dynamically added by the cross-reference phase;

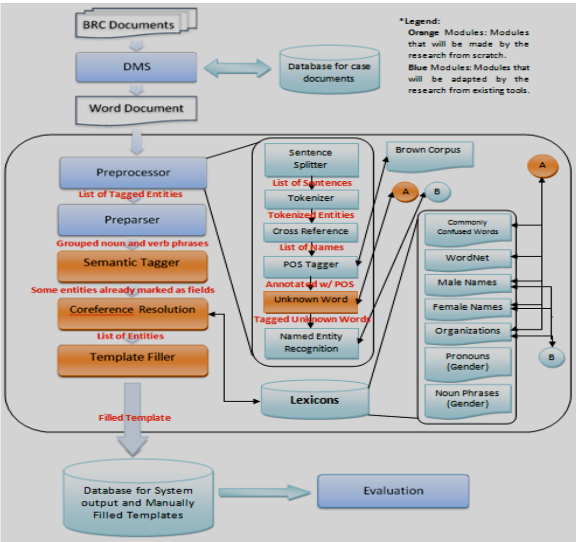


**Figure 4. Information Extraction System Architecture for e-Participation**

The Pre-Parser module follows pre-processing stage. This module establishes the phrases (noun and verb phrases) in the POS-tagged sentences using LingPipe (2003). Having established these phrases, the Semantic Tagger follows which is responsible for extracting candidate values for each field found in a certain template of a document. The Semantic Tagger makes use of the outputs provided by the previous modules to determine the correct candidate values. It basically goes over a certain document and finds the candidates by the use of semantic rules or patterns. The Co-Reference Resolution follows which uses the algorithm of Ruslan Mitkov (1998) for anaphora resolution. The algorithm was augmented to address cataphoric resolutions, which were present in the documents of the agency.

Finally, the Template filler follows which normalizes the Semantic Tagger entries such as dates, values and names to standardize these entries before adding to the database.

## 3 Forums + Opinion Classification and Clustering = Opinion Organization

For the ground-up (or bottom-up) participation component, a web-based opinion detection and classification system, aptly named Vox Pop, was developed. It allows for the public to voice out their opinions regarding topics of discussion (created by moderators) by posting on the e-Legislation system. Vox Pop is able to detect opinions based on the input text of the respondents, annotate opinions to separate them from non-opinions, classify opinions by polarity (as shown in Figure 5) and by topic, clustered together these opinions, and present them through graphical representations of the data. The system has three main modules, namely: the opinion detection module, the opinion classification module and the clustering module.

Again, a whole discourse on managing forums and netizens (or e-citizens) as well as the processes adopted for promoting, regulating and cultivating skills of netizenship are very related and in fact, determined the configuration and business processes of the forum. Nevertheless, the focus of this section is on the technical aspect or the Language Technology used in the forum.

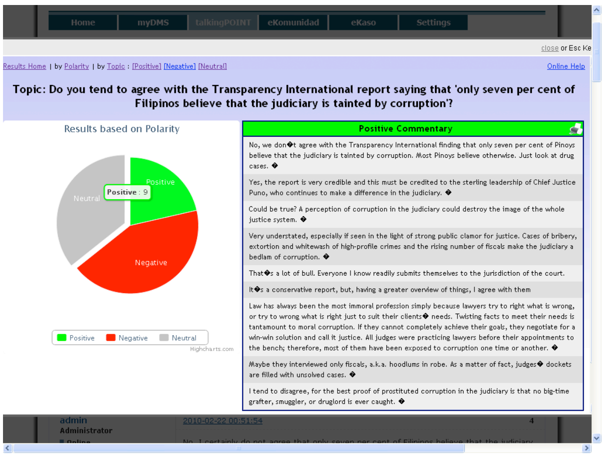


**Figure 5. Opinion Polarity Report Screenshot**

### 3.1 Detection

Commentaries are first gathered into a repository then go through the first module of the system, which is the Opinion Detection Module. The first module detects quotations and opinions present within the commentary. The heuristic goes by using signal phrases to detect quotations taken from previous posts or passages from other sources that are present within the commentary. The presence of quotations taken from previous posts, which can be considered as opinion spamming, could duplicate commentaries within the topic thus resulting in the presence of an additional positive or negative score. This feature prevents the occurrence of these duplicate commentaries or Opinion Spam for a more accurate positive or negative result.

Sentences that are tagged as opinions are forwarded to the classification module while the sentences that are untagged or tagged as quotations are disregarded in the succeeding module.

The output of the opinion detection module is the combination of the detection of quotations and the detection of opinions. These two detection processes are needed in order for the classification process to determine which parts of the commentary are to be forwarded to the Opinion Classification Module. Lines of text that are tagged as opinions are selected to undergo the classification process while sentences that are not tagged or are tagged as quotations will not undergo the classification process.

### 3.2 Classification

After the Opinion Detection module tags commentaries as opinions, all opinions are then classified and tagged with their polarity. To determine the polarity of an opinion, it goes through four steps, namely: Part of Speech Tagging, Polarity Score Generation, Polarity Score Computation and Determining the Polarity of the Commentary. This module uses MontyTagger (Liu, 2003) for part-of-speech tagging and SentiWordNet (Esuli & Sebastiani, 2006) for polarity score generation.

In computing the Polarity score, there are three levels of computation, namely: Word-level, Sentence-level and Commentary-level. In the computation for the word level polarity, the Positivity and Negativity scores of all of the synsets of a particular adjective or adverb, depending on use, will be averaged in order to compute for the scores of that particular word.

After computing for the word level scores, the Positivity and Negativity scores of all adjectives and adverbs in a particular sentence will be added and then averaged in order to come up with the scores of that particular sentence. Finally, this process is repeated one more time, this time adding and averaging the scores of sentences, in order to come up with the commentary-level scores.

### 3.3 Clustering

After being classified by polarity, the commentaries would then be clustered by topic. Each commentary would first undergo two types of pre-processing, namely, stop words removal and stemming of words. After pre-processing the commentaries, the mean of each commentary would then be computed, and then the Euclidean distance between the commentaries and will finally be subjected to the K-Means Clustering proper.

The clustering algorithm used by the system is based on Collaborative Decision Support System (CoDeS) (Chiu et al., 2008). However, the implementation is slightly altered from CoDeS. While CoDeS accepts commentaries without any pre-processing for clustering, Vox Pop's clustering module accepts commentaries which are already classified by polarity by the classification module.

## 4 Prototype Results and Preliminary Findings

The e-Legislation portal integrates the Document Management System (DMS) and the Online Forums with the Information Extraction and Opinion Organization technologies, respectively (see Figure 6). Moreover, the portal provides for features that exploit the structured information coming from the two language technologies and allows users to access or view these structured data. For the Document Management System with Information Extraction, keyword searches are not limited to just the filenames since more complex database queries are available. Moreover, visual modes of reporting by exploiting the structured database from information extraction are made available. An example can be shown in Figure 7 where case activities of the Senate Committee can be visually reported thru a timeline by utilizing extracted date information from related documents in a case.

The scheme of the interaction of the DMS and the online forum, as well as the rules and regulations established in the study for governing the forums, hinges on principles of e-Democracy. These discussions, again, covers another whole set of discourse that will not be covered by this paper. Nevertheless, the same principles of e-Democracy lead to the scheme of having two forums that addresses issues of inclusivity and exclusivity (Islam, 2008) of netizens and having set-up the forums as a self-managed system of e-Participation as envisioned by Banathy (1996).

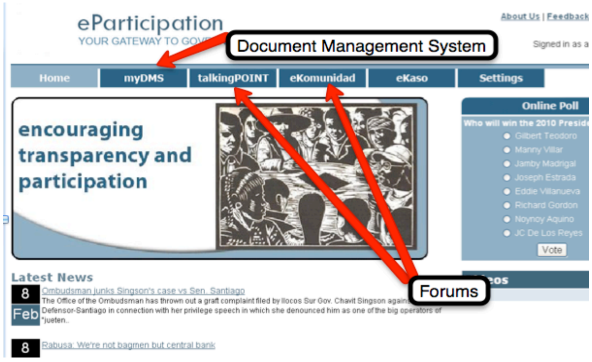


**Figure 6. e-Participation Portal Integrating DMS and Online Forums**


**Figure 7. Timeline Report for Senate Blue Ribbon Committee Cases**

The subsequent subsections will provide insights as to the evaluation and results of the performance of the two Language Technologies used for the e-Legislation Portal.

### 4.1 Testing Setup for DMS with Information Extraction

The study used 50 documents from the agency (Blue Ribbon Committee) for evaluation. The system supports seven kinds of documents and of those 50 documents, 5 are notice of hearing documents, 5 are agenda documents, 5 are order documents, 12 are subpoena documents, 10 are scenario documents, 8 are hearing invitation documents, and 5 are hearing highlights documents. Each type of document has their own set of fields and one of every type of document was used as basis in obtaining the rules for the various fields for the document of its type. Each notice of hearing document has the same format across every notice of hearing documents and this also applies to the other type of documents. Establishing the fields for each document type involved a series of focus-group discussions with different stakeholders ranging from private citizens to non-government organizations (NGO's) and even to representatives from Commission (government) on Women.

In evaluating the extraction performance in terms of accuracy, the representatives from the government agency manually derived and provided the gold standard (or the answer key). The gold standard is matched with the output of the information extraction module to which a score of an absolute score of 1 or 0 is given if it's an exact match or not, respectively. A non-exact match also constitute added, removed or altered words from the correct answer.

The accuracy of the system from the training data given is 100 percent accurate, the reason for this output is because the system is constructed with the training data so it shows that the system is implemented and tested side by side with the training data. The output of the system is based on the pattern that was given to the research by the Blue Ribbon Committee. The resource person indicated that the pattern is being followed for all the documents, so as long as the pattern is being followed by the implementation of the system, then the output would be accurate.

### 4.2 Testing the Actual Data for Information Extraction

Unit-testing the information extraction from the training data showed a 100 percent accuracy performance. In other words, the algorithms of the sub-modules of information extraction were finely tuned for the doing things correctly. After testing the system with the training data, a new set of 50 documents was evaluated. The output of the system is mostly the same with only a small set of differences due to the mismatch of format of values from the training versus the actual data. In other words, there are certain parts in the new test data that are not found in the training or initial test data. One attribute to the disparity is on the irregularity of the document formatting. Another would be the addition of fields and entities manually extracted by the resource people creating the gold standards. ***Table 1*** shows the result of the actual data and only two types of documents have inaccuracies. Overall, averaging the performance of the system, it shows a 95.42% accuracy.

*Table 1. Actual Data Test Results*

| Number of Docs. | Type of Docs. | Accuracy |
|---|---|---|
| 4 | Notice of Hearing | 100% |
| 6 | Agenda | 100% |
| 5 | Order | 100% |
| 17 | Subpoena | 85% |
| 7 | Scenario | 100% |
| 6 | Hearing Invitation | 83% |
| 5 | Hearing Highlight | 100% |

### 4.3 Testing Setup for Forums and Opinion Organization

The corpus was built from commentaries obtained from the website of the Philippine Star Inbox World (www.philstar.com). It contains 1002 commentaries from 22 topics. A linguist evaluator was tapped and did six sets of evaluations, four of which were for the classification module; one was for the detection module, while the last set was for the clustering module.

### 4.4 Detection Results

In order to check whether detected opinions are correctly annotated by the system, the same set of commentaries used to evaluate the classification module was fed into the opinion detection module. All two hundred commentaries were split into sentences and were determined whether they are opinionated or not. One hundred opinionated and another one hundred non-opinionated sentences, as determined by the system, were then randomly selected from the output to be evaluated by the linguist whether they are opinionated or not. All in all, two hundred (200) sentences from one hundred and one (101) commentaries were selected to be evaluated by the linguist. Of the two hundred sentences evaluated by the linguist, one hundred commentaries or **50%** matched with the detection done with the system.

An observation regarding this matter is that even though there are an equal number of 'opinionated' and 'non-opinionated' commentaries as tagged by the system, the evaluators tagged a much greater number of commentaries as 'opinionated'. This may be due to the evaluators having a pre-conceived notation that most of the commentaries are opinionated because of the knowledge that the commentaries come from an opinionated source. However, the system does not know this and instead, bases judgment on the opinion markers present in a commentary.

Another observation in the opinion detectionmodule is that the signal phrases that are used to compare can occur in a given sentence even without a following passage that is enclosed in quotation marks. For example, the signal phrase 'goes' can be used in a sentence as a signal phrase in (*There's a saying that goes, "The road to hell is paved with good intentions.")* and as a verb in *(There goes the alarm*). Majority of the signal phrases are verbs and these verbs may be confused as being verbs without the support of a passage enclosed in quotation marks.

Another observation to the system is that there are other signal phrases that are not included in the overall list of signal phrases used by the system. These signal phrases follow the same structure in quoting passages and sayings but if the signal phrase is not present in the overall list of signal phrases used by the system, the system may not consider it as a quotation. There may also be opinion markers that are not included in the overall list of opinion markers used by the system. These opinionated sentences with markers not found in the list may not be considered by the system as an opinion.

Finally, it was observed that the comparison of opinion markers and signal phrases cannot be accomplished when comparing them with the words in the sentence. This is because of the fact that the lists do not only contain a single word but there are also phrases contained within the list. Examples would be "points out" and "according to" for opinion markers and signal phrases respectively. These observations found in the system may have contributed to the accuracy rate of the opinion detection module.

### 4.5 Classification Results

In order to check whether classified opinions are correctly annotated by the system, two hundred commentaries were chosen from the corpus and were subjected to polarity annotation by the system. The linguist was then requested to identify the polarity of each commentary. Of the two hundred commentaries evaluated by the linguist, one hundred one commentaries or **50.5%** matched with the classification done with the system.

Three sources of errors were found. The first error is the failure to process double-negative phrases. The system gets the individual scores of the words in a commentary then adds them afterwards. An example would be the statement "*Another Aquino to dominate the Philippine as a leader has my yes. I hope Noynoy does not fail* *us.*" This statement is evaluated by the linguist as positive, as the commentary is a statement of support for Noynoy Aquino. The word '*Yes*' alone tells the human brain to evaluate the statement as positive. However, the formula used fails when faced with double negatives, or negative words placed next to the word '*not*'. The example sentence was marked as negative by the system because the words '*not*' and '*fail*', which are both negative in nature, were evaluated separately from each other by the system. This is why the negativity score of the statement increased, instead of the positivity score increasing if it were processed as one statement '*not fail*'.

The second error is the presence of high polarity words. Since the range of the scores of the words is normalized from 0 to 1, it is possible for several words of a particular polarity to overpower a word of the opposite polarity. An example of this would be the statement "*I believe he would make a good president of our country if given a chance, but he should be a good senator first*." This statement is evaluated by the linguist as negative, as the 'but' part of the commentary is not in support of *Noynoy Aquino*. However, it was marked as positive by the system. Although the word '*but*' has a high negativity score, the positivity score of the words '*believe*' and 'good', which appeared twice, overpowered the score of the word '*but*' because there are more words present which have high positivity scores.

The third error occurs when adjectives and adverbs are absent in a sentence. Adjectives and adverbs contain the sentiment of a statement. That is why in SentiWordNet, adjectives and adverbs have non-zero positivity and negativity scores. However, if a statement does not contain adjectives and adverbs, the positivity and negativity scores of these statements are both zero, leading the system to classify them as neutral. An example would be the statement "*A hundred percent yes*." This statement is evaluated by the linguist as positive, as the word "*Yes*" is enough to tell that the sentiment is positive. However, it was marked by the system as neutral because the words '*a*', '*percent*' and '*yes*' are nouns, while the word 'hundred' is an adjective which has both zero positivity and negativity scores.

### 4.6 Clustering

In order to check whether the clusters produced by the system are correct, two topics containing eighty one (81) commentaries were chosen to be clustered by the system and evaluated by the linguist afterwards. In generating the clusters, the commentaries in each topic were first segregated into three clusters, produced by the opinion classification module of the system. Afterwards, each polarity-segregated cluster was further segregated into three smaller clusters. Thus, all in all, eighteen clusters were produced by the system for the evaluation of the linguist. The clusters generated by the system were analyzed by the linguist whether 1) the commentaries in each cluster are related with each other, and 2) why some commentaries are singled out into single clusters. Of these eighteen clusters, thirteen contain multiple commentaries, while the remaining five contain only single commentaries.

In the first topic, three unrelated clusters were deemed as such because the views in them are complex, vary with each other and some go beyond the intended topic. In the second topic, the three unrelated clusters were deemed as such because their views vary from each other. Another unrelated cluster contained an opinion and a declaration of support. These commentaries are grouped together because of the similar words that are found within them, such as 'Filipino' and 'honesty' in the first topic. However, the clustering algorithm does not take into account synonyms or words that are similar in context, resulting to some clusters being mismatched. All in all, the linguist evaluation shows a **53.85%** accuracy of the clustering module.

In the process of clustering, five commentaries were not clustered and instead, were isolated from the other commentaries. Of the five clusters containing only one commentary each, two of them were evaluated by the linguist being isolated because they sound like factual statements. On the other hand, the other two were evaluated by the linguist as being isolated because they contain alternate reasons on why they agree or disagree with the topic. Finally, the last cluster containing only one commentary was probably isolated because "*it cites very specific instances*", as the linguist points out.

These commentaries were isolated probably because the default number of clusters (three) is not the optimum number of clusters for the set of commentaries. An example would be the positive clusters under the second topic. When the number of clusters was set to three, one commentary was isolated while the other two clusters contained three and thirteen commentaries respectively. However, when the number of clusters was set to two, no more commentaries were isolated and two clusters containing multiple commentaries are formed.

## 5 Conclusions and Recommendations

Overall, the document management system was designed for the Blue Ribbon Oversight Office Management (BROOM) after constant requirements gathering, consultation and collaboration with BROOM. The Information Extraction module was also designed implemented and evaluated in the same manner and thus garnered very high extraction accuracy.

The system currently isn't robust enough to handle noise brought about by poor conversion of hardcopies to electronic versions (e.g. Optical Character Recognition). Moreover, the algorithms are highly dependent on the regularity of the documents and would perform differently if documents don't conform to regular formats.

It is recommended that improving the Information Extraction module entail the following:

1. Image recognition algorithms that allows for capturing and extracting signatures from senators and annotation information as extracting these data allows for interesting research and data mining value;
2. Improvement of semantic tagger to handle new templates without changing the code but instead process new documents from templates based on formal specifications; and
3. Include Filipino processing in extracting texts as transcript-type of documents would normally contain a mixture and code switching of English and Filipino languages.

For the opinion organization, the study focused on developing a system that uses text processing techniques in organizing the sentiments of public commentary.

The opinion detection module includes the detection of quotations and opinions given input commentaries to which opinion classification is affected. Quotation detection prevents quotations from being redundantly classified, thus providing more accurate results for classification.

The opinion classification module included part-of speech tagging, polarity score generation via SentiWordNet and word, sentence and commentary-level score computations. It was uncovered that part of speech tagging is important as adjectives and adverbs really do have the linguistic basis in classifying commentaries by sentiment. However, it was also shown that SentiWordNet should not be the sole tool used in dealing with polarity, as it only outputs the score of each word, and it does not consider more complex factors such as double negatives and idioms.

The clustering module includes stop words removal, stemming and the use of the K-Means clustering algorithm. Stop words removal and stemming are necessary in clustering as they filter commentaries, preventing non-relevant words such as prepositions, articles and pronouns from being used as the basis for clustering. However, having a fixed number of clusters, generated by the K-Means clustering algorithm, which is three in this case, is not the most optimal solution for all cases. If there are only few commentaries to be clustered, setting the number of clusters to a smaller number such as two might be more optimal. Conversely, three clusters might not be sufficient for a larger dataset, such as the ones containing thousands of commentaries in them.

All of these issues attributed to the dismal 50% overall accuracy performance of the opinion organization and classification and data clustering. Nevertheless, the different presentations and structured reporting of commentaries and opinion facilitated by the automated detection, classification and clustering still provide a way for structuring information that can facilitate policy making or legislation.

It is recommended that improving the automated opinion organization entail the following:

1. As with the Information Extraction Module, include Filipino processing as texts in comments and opinions also include a mixture and code switching of English and Filipino languages;
2. Utilize SentiWordnet version 3.0 which increased by 20% in accuracy versus Version 1.0 (Baccianella, et al. 2010), as the current implementation involves version 1.0; and
3. Investigate machine learning on top of relying on lexical and rule-based resource such as SentiWordnet to allow for flexibility and robustness of system;

For both major modules addressing the top-down and bottom-up information, linguistic resources and algorithms are still currently being improved. But more importantly, the Blue Ribbon Oversight Office Management (BROOM) of the Philippine Senate is now steadily migrating to a new business process, creating the possibility of allowing the office to open its documents to the public (towards transparency and good governance) more expeditiously, and allowing feedback from citizenry (towards participation) as the office is currently moving to actual adoption of the eLegislation Portal.

## Acknowledgments

The e-Legislation Portal was greatly facilitated by the Blue Ribbon Oversight Office Management (BROOM) of the Philippine Senate headed by Atty. Rodolfo Quimbo, Director General who served as the "champion" in government for this endeavour. Successful business process modeling of the said office due to solid support and constant interaction led to a very functional and seamless integration of office automation and portal for pushing information to general public. Moreover, constant interaction and support for quality evaluation led to a robust and accurate information extraction performance. Dr. Shirley Lua, linguist and faculty member of the Literature Department of De La Salle University, did the daunting task of evaluating the opinion organization's performance.

The whole project is under the PanEGov project managed by IdeaCorp headed by the executive director, Dr. Emmanuel C. Lallana and funded by IDRC, Canada.

## References

Ann Macintosh. 2007. e-Democracy and e-Participation Research in Europe. In Chen, et al (eds.) *Digital Government: E-Government Research, Case Studies, and Implementation*. Springer.

Hobbs, J. R. 1993. The Generic Information Extraction System. In MUC5 '93: Proceedings of the *5th Conference on Message Understanding* (pp. 87-91). Morristown, NJ, USA: Association for Computational Linguistics. Available from http://dx.doi.org/10.3115/1072017.1072029

Schwartz, A. and Hearst, M. 2003. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. *Pacific Symposium on Biocomputing* 8:451-462. Available: http://helix-web.stanford.edu/psb03/schwartz.pdf

LingPipe: a suite of Java libraries for the linguistic analysis of human language. 2003-2007. Accessed March 2010. http://ir.exp.sis.pitt.edu/ne/lingpipe-2.4.0/index.html

ANNIE: POS Tagger (Plugins for GATE). 1995-2011. Accessed March 2010. http://gate.ac.uk/gate/doc/plugins.html

Mitkov, R. 1998. Robust Pronoun Resolution with Limited Knowledge. In *Acl-36: Proceedings of the 36th Annual Meeting of the Association for Computational Llinguistics and 17th International Conference on Computational Linguistics* (pp. 869–875). Morristown, NJ, USA: Association for Computational Linguistics. Available from http://dx.doi.org/10.3115/980691.980712

Esuli, A., Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the *5th Conference on Language Resources and Evaluation* (pp. 417-422).

Liu, Hugo. 2003. *Monty Tagger: Commonsense-Informed Part-of-Speech Tagging*. Available online: http://web.media.mit.edu/~hugo/montytagger/

Chiu, C., Dy, J., Lee, P., & Liu, M. 2008. *Collaborative Decision Support System* [Thesis]. Manila, Philippines: College of Computer Studies, De La Salle University.

Islam, M.S. 2008. *Towards a Sustainable e-Participation Implementation Model*. European Journal of e-Practice. Available online: http://www.epractice.eu/files/5.3.pdf

Banathy, Bela H. 1996. *Designing Social Systems in a Changing World*. New York, Plenum.

Baccianella S., Esuli A. & Sebastiani F. 2010. *SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining*. In: LREC 2010 - Seventh conference on International Language Resources and Evaluation (Valletta, Malta, 18-22 maggio 2010). Proceedings, pp. 2200 - 2204. ELRA, 2010.