# A Bilingual Chatbot Using Support Vector Classifier On an Automatic Corpus Engine Dataset

Jasper Kyle Catapang[1], Geoffrey A. Solano[2], Nathaniel Oco[3]
[1,2]Department of Physical Sciences and Mathematics, College of Arts and Sciences
University of the Philippines Manila - Manila City, Philippines
[2]Department of Computer Science, College of Engineering
University of the Philippines Diliman - Quezon City, Philippines
[3]College of Liberal Arts
De La Salle University - Manila City, Philippines
Email: [1]jcatapang@up.edu.ph, [2]gasolano@upm.edu.ph, [3]nathanoco@yahoo.com

***Abstract*—Brands are shifting to digital services to cater to their customers who have been spending more time online. The technology that exists today enhances customer experience and actualizes customer expectations through virtual service agents or "e-service agents" during real-time interactions. Brands in most countries have to deal with bilingual customers as globalization occurs. Business process outsourcing is among the Philippines' top foreign exchange earner aside from overseas workers' remittances. These Philippine companies offer customer service but are mostly left manned—needing constant supervision. As a solution, the researchers present a bilingual retail chatbot that could handle the two official languages of the Philippines, Filipino—based on Tagalog—and English, and their code-switching variant Taglish. The proposed bilingual retail chatbot uses k-fold grid search cross-validation on a dataset constructed by a bilingual automatic corpus engine and a combination of both (1) support vector classifier—–for intent identification, and (2) hash set containment—–for attribute identification.**

***Index Terms*—chatbot, natural language processing, support vector machine, grid search, automatic corpus engine**

## I. Introduction

Eighty-seven percent of in-store purchase decisions are influenced by service agents and the customers that are most likely to purchase from trusted salespersons amount to seventy-seven percent. Service agents determine the success of service exchange [1]. Strengthening consumer preferences and awareness by providing deep experiences is one of the main objectives of brand managers over their competitors [2].

Meeting customer expectations, companies gain customer satisfaction, earn loyalty, generate positive feedback, obtain beneficial purchase intentions, and gain profit [3]. Service agents inform consumers on trends and help consumers save time and give credible advice [4]. The traditional service agent interactions occur face-to-face, but with the proliferation of social media, companies are now fulfilling customer needs outside the conventional service area for quick response. Online services are accessible, efficient, and cost and time-saving [5].

Consumers and companies communicating through mobile messenger apps via text messages is already a familiar interface. Consumers can effortlessly reach out to companies just by chatting at a convenient time for the consumers, instead of calling, e-mailing, or opening an application [6]. Chatbots can be used to automate company-customer interaction through natural language [7][8] and chatbots generate natural language as output to engage in conversation [9]. These chatbots require huge training data to perform considerably well. Brands in most countries have to deal with bilingual customers as globalization occurs. This brings about the demand for chatbots that understand more than one language—which needs the help of natural language processing and machine learning.

Algorithms that can learn from input information and make predictions based on the same data are what machine learning deals with. Performing machine learning tasks on enormous datasets is a computationally expensive job and would need a significantly longer amount of time to complete. Clustering and classification of data of this volume utilizing supervised and unsupervised techniques are the challenges machine learning face in the present [10], [11], [12], [13].

A set of findings [14] by 2017 A.T. Kearney Global Services Location Index states that the Philippines is rank seven in the world's top outsourcing destinations. The country ranking is based on business environment, financial attractiveness, people skills and availability. K. Lema [15] also points out that the United States is the Philippines' largest outsourcing client, comprising an estimate of 65% of the local business process outsourcing (BPO) market. BPO companies in the Philippines also serve Europe, New Zealand, and Australia. Philippine Statistics Authority (PSA) data [16] shows that the Philippines has 851 registered BPO companies, more than half of which are call centers (429). Almost four hundred (46.2%) firms provide computer or IT-related services. V. Zoleta [17] adds that by contributing more than 1 billion Philippine pesos total revenue, the BPO industry is the Philippines' top foreign exchange earner.

As a case study, the researchers have selected Taglish as a code-switching topic with its application on local BPO customer service. Taglish is the alternation of English and Tagalog. Filipino is based on Tagalog—with the former being the national language of the Philippines. Data collected in [18] reveals that there are about twenty-two million native speakers of Tagalog.

This paper proposes the use of a support vector machine to train on the dataset generated by a bilingual automatic corpus engine. The bilingual automatic corpus engine is an engine that takes keywords—nouns, adjectives, and verbs—from a primary language, English, generates these keywords' translation on the secondary language, Tagalog, and combines them permutatively.

The following are the researchers' contributions:

1) Creation of a bilingual automatic corpus engine that generates grammar-free phrases in the two languages of choice
2) Construction of a high-accuracy support vector machine classifier using keywords used as input for the bilingual automatic corpus engine

The research paper is divided into six parts. In section II, the related works are discussed to show the process of customer service, to highlight the different chatbots, to explain different metrics of chatbot evaluation, and to lay out how the corpus engine can be implemented. In section III, the step-by-step process of creating the entire chatbot is discussed—from building the corpus engine to the response generation. In section IV, the results of the experimental study are shown and explained. Then, in section V, the conclusion of the study is given with the statistics and findings by the researchers. Lastly, recommendations for the improvement of the chatbot created for the experimental study are given in section VI.

## II. Related Work

### A. Customer service

An article [19] states that the companies that support products with an efficient client support and service are most likely to succeed. The underlying answer to enhancing client service is a systematized procedure. It can be achieved with elaborate method documentation. The result would be an improved division productivity and potency, and method controls that profit corporately in terms of issue resolution, outreach, and engagement. Technical support is a procedure focused around addressing technical issues brought in by clients. Furthermore, the article [19] enumerates that this process includes client identity verification, question clarification, issue escalation, problem resolution, issue monitoring, and any work needed to solve the customer's problem or provide satisfactory alternatives.

### B. Chatbot

Atwell and Shawar [7] explain that ELIZA is one of the first chatbots using a strategy of keyword coordination. It was created by Joseph Weizenbaum. The goal was to convince the client that he or she could inquire for information and then the chatbot would search for particular keywords. If relevant words were retrieved from the query, then the appropriate response was responded by ELIZA. Otherwise, to keep the discussion from stopping, ELIZA would continue to get more information from the client. ELIZA uses Artificial Intelligence Markup Language (AIML).

Retrieval-based methods obtain candidates for possible replies from a pre-constructed index. After which, these rank the candidates and choose a response from the top-ranked candidates. Generation-based methods, on the other hand, make use of natural language generation techniques to answer a query or reply to a message. W. Lu et al [20] note that they have opted to work on the response selection for retrieval-based chatbots in a single turn scenario, since retrieval-based methods can give fluent replies. Additionally, single turn is the foundation of conversation in any chatbot.

Segura et al [21] have another chatbot, named Chatbol. It is deployed as a Slack client for text-based input interaction with users in Spanish. One of Chatbol's main modules, an NLU block via RasaNLU, is trained to obtain the intents and associated entities related to user's questions about football players, teams, trainers, and fixtures. RasaNLU is built on an SVM architecture internally. The data for the entities is obtained by creating SPARQL queries to the Wikidata website. After which, the retrieved information is employed to update the precise chatbot responses.

### C. Chatbot Evaluation

Shawar and Atwell [22] adopted three evaluation metrics for their chatbot.

1) Efficiency of the dialog via matching type
2) Quality of dialog criteria via reply type
3) Open-ended feedback as users' satisfaction assessment

They measured the effectiveness of four sample dialogs based on the criteria shown in Table 1. They also measured the effectiveness of the adopted learning mechanisms, and observed the increase of the ability of the chatbot to replies to general queries as shown in Table 1 as well.

| Matching Type | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| Atomic | 1 | 3 | 6 | 3 |
| First word | 9 | 15 | 23 | 4 |
| Most significant | 13 | 2 | 19 | 9 |
| No match | 0 | 1 | 3 | 1 |
| Number of turns | 23 | 21 | 51 | 17 |

TABLE I: Response type frequency

The frequency of each type in each dialogue generated between the user and the Afrikaans chatbot shown in Shawar and Atwell's research was calculated; these absolute frequencies are normalized to correlative probabilities. No test was applied, this approach to evaluation by dialog efficiency metrics shows that the first word and the most significant methods actually improve the ability to create appropriate replies to users and let the conversation continue.

To quantify the quality of each response, Shawar and Atwell [22] aimed to classify replies according to an independent human evaluation of reasonableness: the reply is reasonable, the reply is simply weird, or the reply made no sense. The transcript was given to an Afrikaans-speaking teacher and they requested the teacher to mark each reply according to these categories. The turn number per dialog and the frequencies of each response type were approximated by Shawar and Atwell.

The prototypes were based only on literal pattern matching against corpus utterances. Furthermore, Shawar and Atwell [22] discuss that their Afrikaans-speaking analyzers found these prototypes disheartening. Several of their conversation trials found exact matches in the training corpus, so the chatbot responded with the default reply most of the time. Yielding more positive feedback, the AIML pattern matching via least frequent-word and first-word techniques had visible success. Shawar and Atwell's analyzers found the conversations much more engaging and also less cyclic. They quantify user satisfaction by the same informal user feedback.

### D. Corpus building

An article [23] mention that pre-dating the technical feasibility of digital resources, there was already the concept of large-scale construction of bodies of text for linguistic analysis. Integral to various aspects of research into the nature and functioning of human language, corpus-based linguistics has numerous applications in different fields.

Additionally, [23] has pointed out that the volume of data, as well as its diversity, is to be included—indicating that the definition of an industrial-style production line environment must be introduced. The following is the process:

1) capturing of information from every commercial associate chosen and preparing of materials to another concretized format
2) principal checking and conversion wherein every piece of text must be checked against the format the information was stored in, automatically convert it to another project standard format, and make a record that it was actually accessed
3) automatic addition of linguistic segmentation and word-class tagging
4) text cataloging and final checking lexically annotated texts; this produced a detailed header from the database of the project and then the very text is appended to the dataset

Gabrielatos [24] proposes a paper that has an attainable measure of the additional terms' relevance to a user inquiry. Based on information retrieval processes, relative query term relevance (RQTR), can be integrated with another approached utilized in building different textual datasets from the internet, namely keyword analysis. RQTR is obtained through query term relevance (QTR) and a constant B. QTR is calculated as $\frac{CQ\&T}{T}$—where $CQ\&T$ are the core query and candidate term respectively. $T$ is the number of texts that is considered output by a query that only contains the candidate term—while RQTR as $\frac{100(QTR-B)}{B}$. It does not need the number of significant documents in the database, and also, it is independent of reference corpora. Although it does not make use of judgments from the users of document relevance, it does permit subjective decisions—but subjective decisions are triangulated against keyness and RQTR.

Overall, the related works have provided the foundations for the solution in developing a customer service agent framework that can cater the needs of Philippine brands by addressing their requirements in English, Tagalog, and even Taglish—needed by the researchers of this bilingual customer service agent. However, there is still a need to supply a proper corpus builder or corpus engine specifically for Taglish to give the appropriate solution. This research aims to bridge that gap to propose a feasible solution.

## III. Experimental Study

To answer the problem aforementioned, the researchers designed a bilingual retail chatbot that uses k-fold grid search cross-validation on a dataset constructed by a bilingual automatic corpus engine and a combination of both (1) support vector machine—–for intent identification, and (2) hash set containment—–for attribute identification. The two languages are English, as the primary language, and Tagalog, as the secondary language.

The third-party technologies the researchers have utilized for the experimental study are the following:

1) WordNet—for the retrieval of hyponyms and hypernyms based on synset relationships
2) Google Translate API—for translation of the primary language keywords
3) Heroku PostgreSQL—for storage of item inventory and branch directory

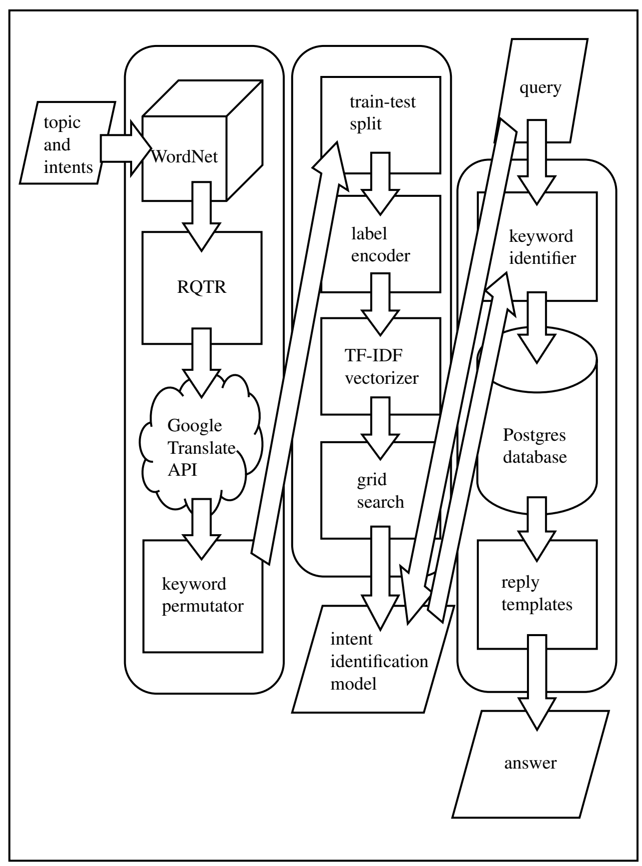


Fig. 1: Proposed bilingual chatbot system architecture

Figure 1 illustrates the system architecture of the experimental study that would be discussed in detail in the proceeding subsections.

### A. Keyword Selection for Corpus Seed

The topic and intents are used as inputs—called corpus seed—for the bilingual automatic corpus engine. The affiliation of the topic is extracted through a lexical database of the primary language, in this case: WordNet for English—through its hypernym (umbrella term). This affiliation would be used to identify the attributes associated with the extracted affiliation. The attributes are nouns, verbs, and adjectives. These are attributes would be considered as candidate terms and their relevance would be calculated as described in Gabrielatos' research [24] by calculating the RQTR. The attributes are then extracted of co-hyponyms—words belonging under the same hypernym. The words collected from the attributes and intents are translated to the secondary language—in this experiment, Tagalog— via Google Translate API. Then these words are used to construct the input of the bilingual automatic corpus engine. The bilingual automatic corpus engine would process the output dataset and the most influential words so that they are incorporated. The corpus generated would be labeled with individual intents, and would serve as input for the classifier.

Table 2 shows an example corpus seed for the automatic corpus engine. It shows how different keywords are sorted which would be permuted with each other to produce multiple phrases.

| header | verbs | objects | adjectives |
|---|---|---|---|
| like | buy | sneakers | red |
| want | purchase | sandals | blue |
| available | - | polo shirt | striped |
| stock | - | long boots | leather |

TABLE II: An example corpus seed in English for the automatic corpus engine for the purchase intent

Table 3 shows a sample output of the bilingual automatic corpus engine:

| text | intention |
|---|---|
| want buy dress pula | purchase |
| meron bang yellow bag | purchase |
| bilihan sa Philadelphia | location |
| gusto ko itim sapatos | purchase |
| branch near Arizona | location |

TABLE III: An example output dataset in English and Tagalog generated from the automatic corpus engine

### B. Training the SVM classifier

The bilingual corpus is split into 20% testing and 80% training data. Afterward, the labels are encoded. A term frequency-inverse document frequency vectorizer is applied to the training and testing sets, ready for grid search. The grid search is to be performed on the parameter grid C: [1, 10, 100] and kernels: linear, polynomial, and RBF, with 5-fold cross-validation. Gamma is set to be $\frac{1}{n}$, where $n$ is the number of features. The support vector machine architecture is used as supported by the framework of RasaNLU, the NLU unit used by Chatbol in Segura et al's paper [21].

J.P. Hwang et al [25] introduces support vector machine (SVM) as a supervised learning algorithm with use in classification and regression. Supervised learning algorithms are made up of dependent variables that are predicted from an independent variable set. There must exist a function that will map the input variables to the wanted output. This training goes on until it reaches a particular accuracy level. Examples of supervised learning are regression, k-nearest neighbors, etc. linear SVM is a binary classifier. Furthermore, [25] adds that given a training set $S = \{(x_1, y_1), (x_2, y_2), ...(x_n, y_n)\}$, where each point $x_i$ belongs to $R^m$, and $y_i$ is the classification label of $x_i$, a function $f(x) = w^T\phi(x_i) + b$ needs to be determined—where $w$ is the normal vector of the hyperplane, $b$ is a real number, and $\phi(x_i)$ is a mapping from $R^m$ to a higher dimensional space $R^{m+n}$ where $n$ is a positive integer. Instead of calculating the mapping $\phi$, SVMs use kernel functions. The three standard kernel functions are linear—$K(x_i, x_j) = x_i^T x_j$, polynomial—$K(x_i, x_j) = (x_i^T x_j + c)^d$, where $c$ and $d$ are parameters, and radial basis function (RBF)—$K(x_i, x_j) = exp(\frac{|x_i - x_j|^2}{\sigma^2})$, where $\sigma$ is another parameter. Afterward, cross-validation is utilized to safeguard the model from overfitting, particularly where the raw volume of data may be limited. In cross-validation, the model engineer makes a fixed number of partitions on the data, called folds, run the analysis on each fold, and then average the overall error estimate [26].

To address the task of searching for attribute words in the query, the following were considered. According to [27], linear search is used for unsorted arrays. For sorted arrays, binary search is used. This reduces the worst-case time complexity to *O(log n)*. Among other searching methods, a hash set gives a constant runtime *O(1)*. A hash set implements a hash function. A hash function operates on a key and then generates an address in the table.

### C. Response generation

The resulting row from the executed database query would provide the chatbot the words it requires to build an appropriate response to the query. The response is already hard-coded except the variables to be filled in by the database query result. Failure to obtain any result would prompt the chatbot to say that it does not understand the query.

## IV. Results and Discussion

The bilingual automatic corpus engine produced around 100,000 phrases in English and Tagalog. This corpus is the dataset set as input for the support vector machine. The support vector machine was trained with a gamma value of $\frac{1}{n}$ where $n$ is the number of features, a linear kernel, and a C value of 1 after these values were obtained from the grid-search with 5-fold cross-validation.

After training on the support vector classifier, it yielded an accuracy of 98% on the bilingual dataset. The chatbot is tested through three rounds of inquiry. The first round is for five English questions. The second round is for five Tagalog questions. Lastly, the third round, five Tagalog-English bilingual questions.

| inquiry | identification | actual label |
|---|---|---|
| Where is the nearest store? | location | location |
| Do you have a store in Miami? | location | location |
| I want to buy red sandals. | purchase | purchase |
| Do you have leather bags in New York City? | purchase/location | purchase/location |
| I would like to have an animal print dress. | purchase | purchase |

TABLE IV: First round results on English queries

| inquiry | identification | actual label |
|---|---|---|
| May branch ba kayo sa Manila? | location | location |
| Meron ba kayong malapit sa amin? | location | location |
| Gusto ko bilhin yung puting bag. | purchase | purchase |
| Makakabili ba ako ng dilaw na tsinelas sa may Cebu? | purchase | purchase/location |
| May sinturong balat pa rin ba kayo, yung parang itim? | purchase | purchase |

TABLE V: Second round results on Tagalog queries

| inquiry | identification | actual label |
|---|---|---|
| Do you have like parang store sa may Makati? | location | location |
| Meron ba kayong location near Pasay? | location | location |
| I want that bag na puro stripes! | purchase | purchase |
| May available ba kayong orange duffel bag sa Quezon | purchase/location | purchase/location |
| I'd like to make bili that fancy na damit na pink. | purchase | purchase |

TABLE VI: Third round results on mixed English and Tagalog queries

For the first round, the validation inquiries were identified 100% accurately as shown in Table 4. In the second round, the validation inquiries were identified correctly but one query was unidentified completely as it was supposed to be a combined intent, giving it a 90% accuracy (4.5/5)—as evident in Table 5. Lastly, in Table 6, the mixed English and Tagalog inquiries yielded a 100% accuracy.

To evaluate the chatbot, the researchers used Shawar and Atwell's user satisfaction assessment on 10 test users. Table 7 shows the test users' assessment.

| Dialogue number | reasonable | weird | nonsensical |
|---|---|---|---|
| Number 1 | 10 | 0 | 0 |
| Number 2 | 8 | 2 | 0 |
| Number 3 | 10 | 0 | 0 |
| Number 4 | 10 | 0 | 0 |
| Number 5 | 9 | 1 | 0 |

TABLE VII: User assessment of the chatbot dialogue

Dialogue number 1 was assessed 100% reasonable; number 2: 80% reasonable, 20% weird; number 3: 100% reasonable; number 4: 100% reasonable; and number 5: 90% reasonable, 10%weird. The dialogues performed exceptionally well—94% reasonable in total.

## V. Conclusion

The accuracy of the model is 98 percent given that it is a narrow, closed domain chatbot. It is not expected to encounter data that has not yet been processed before. The specialization the automatic corpus engine provides ensure an efficient and smooth integration of the machine learning used and the database querying that is to be done.

On actual standardized queries, the classifier performs exceptionally on all three cases. It has identified all five intentions accurately for English queries. It has also identified all the intentions for the Tagalog queries—with one intention technically containing either purchase or location intent. Lastly, it has identified the intentions of the bilingual queries as well. The researchers believe this is again due to the narrowness of the topic that the SVM learned the data almost perfectly.

On chatbot evaluation using Shawar and Atwell's metric [22], the chatbot is 94% reasonable according to the test users.

## VI. Recommendations

The chatbot produced by the research can be further improved by using a named entity recognition algorithm rather than using hash set containment. The named entity recognition (NER) algorithm would eliminate the usage of multiple loops and make the process of attribute identification fully-automated and more systematized. The NER can be implemented with a long short-term memory, artificial recurrent neural network architecture.

## Acknowledgements

The authors would like to thank Payless Shoes Inc. for the use case they provided for an earlier version of the system. The authors would also like to thank Ms. Perl Gasmen for her invaluable insights and suggestions.

## References

[1] Insider-Trends, "Retail trends factfile 2017: Physical retail". [Online]. Available: http://www.insidertrends.com/retail-trends-factfile-2017-physical-retail. [Accessed 11-28-2018].

[2] G. Atwal, and A. Williams, "Luxury brand marketing—The experience is everything!". Journal of Brand Management, 16(5-6), 338-346. 2009.

[3] K.E. Reynolds, and S.E. Beatty, "Customer benefits and company consequences of customer-salesperson relationships in retailing". Journal of Retailing, 75(1), 11-32. 1999.

[4] M. Holzwarth, C. Janiszewski, and M.M. Neumann, "The influence of avatars on online consumer shopping behavior". Journal of Marketing, 70(4), 19-36. 2006.

[5] A. Escobar, "The impact of the digital revolution in the development of market and communication strategies for the luxury sector (fashion luxury)". Central European Business Review, 5(2), 17. 2016.

[6] M. van Eeuwen, "Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers". 2017.

[7] E. Atwell, and B.A. Shawar, "Chatbots: are they really useful?." LDV forum. Vol. 22. No. 1. 2007.

[8] A. Kerly, P. Hall, and S. Bull, "Bringing chatbots into education: Towards natural language negotiation of open learner models". Knowledge-Based Systems, 20(2), 177-185. 2007.

[9] D. Griol, J. Carbo, and J.M. Molina, "An automatic dialog simulation technique to develop and evaluate interactive conversational agents." Applied Artificial Intelligence, 27(9), 759-780. doi:10.1080/08839514.2013.835230. 2013.

[10] D. J. C. MacKay, Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge, England. 2003.

[11] E. Alpaydin, Introduction to Machine Learning (Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA. 2004.

[12] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, New York. 2007.

[13] K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA. 2012.

[14] A.T. Kearney Global Services, "The 2017 A.T. Kearney Global Services Location Index". [Online]. Available: https://www.atkearney.com/digital-transformation/gsli/full-report. [Accessed 3-28-2018].

[15] K. Lema, "Rise of the machines: Philippine outsourcing industry braces for AI". [Online]. Available: https://www.reuters.com/article/us-philippines-economy-outsourcing/rise-of-the-machines-philippine-outsourcing-industry-braces-for-ai-idUSKBN1D90BH. [Accessed 3-28-2018].

[16] Philippine Statistics Authority, LABSTAT Updates Vol 22 No 13 on Industry Profile - BPO 2018. 2018.

[17] V. Zoleta, "Business Process Outsourcing to the Philippines [Complete Guide]". [Online]. Available: https://grit.ph/bpo. [Accessed 3-28-2018].

[18] C.K. Cheng, N.R.T. Lim, and R.E.O Roxas, "Philippine Language Resources: Trends and Directions". Proceedings of the 7th Workshop on Asian Langauge Resource (ALR7), Singapore. 2009.

[19] Opsdog, "What are Customer Service Flow Charts?". [Online]. Available: https://opsdog.com/categories/workflows/customer-service. [Accessed 3-28-2018].

[20] Y. Wu, W. Wu, C. Xing, C. Xu, Z. Li, and M. Zu, "A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots". Computational Linguistics, Volume 45, Issue 1. March 2019.

[21] C. Segura, J. Luque, R.E. Banchs, M.R. Costa-jussa, "Chatbol, a chatbot for the Spanish "La Liga"". International Workshop on Spoken Dialog System Technology 2018 (IWSDS), Singapore. 2018.

[22] E. Atwell, and B.A. Shawar, "Different measurements metrics to evaluate a chatbot system". NAACL-HLT-Dialog '07 Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies, pp. 89-96. 2007.

[23] "How to build a corpus", [Online]. Available: http://users.ox.ac.uk/lou/wip/ Boston/howto.htm. [Accessed 3-28-2018].

[24] C. Gabrielatos, "Selecting query terms to build a specialised corpus from a restricted-access database". ICAME Journal 31:5-43. January 2007.

[25] J.P. Hwang, S. Park, and E. Kim, "A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function". Expert Systems with Applications, 38(7), 8580–8585. 2011.

[26] "Cross-Validation in plain English?" [Online]. Available: https://stats.stackexchange.com/questions/ 1826/cross-validation-in-plain-english. [Accessed: 2019-05-03].

[27] "Concept of Hashing". [Online]. Available: https://www.cs.cmu.edu/adamchik/15-121 /lectures/Hashing/hashing.html. [Accessed: 2019-05-05].