**A Semi Supervised Approach to Word Sense Disambiguation Based on Label Propagation for Filipino Language**

A Thesis Proposal  
Presented to the Faculty of the  
College of Computer and Information Sciences  
Polytechnic University of the Philippines

In Partial Fulfilment  
of the Requirements for the Degree

Bachelor of Science in Computer Science

Fiestada Angelo F.

Lim John Kenno Mikko S.

October 2017

# TABLE OF CONTENTS

[TABLE OF CONTENTS ii](#_Toc495655485)

[CHAPTER 1 The Problem and Its Background 1](#_Toc495655486)

[1.1 BACKGROUND OF THE STUDY 1](#_Toc495655487)

[1.2 STATEMENT OF THE PROBLEM 2](#_Toc495655488)

[1.3 THEORETICAL/CONCEPTUAL FRAMEWORK 3](#_Toc495655489)

[1.3.1.1 THEORETICAL FRAMEWORK 3](#_Toc495655490)

[1.3.2 CONCEPTUAL FRAMEWORK OF THE STUDY 3](#_Toc495655491)

[1.4 SIGNIFICANCE OF THE STUDY 4](#_Toc495655492)

[1.5 SCOPE AND LIMITATION OF THE STUDY 4](#_Toc495655493)

[1.6 OPERATIONAL TERMS 4](#_Toc495655494)

[CHAPTER 2 Review of Related Literature 6](#_Toc495655495)

[2.1. Review of Related Literature 6](#_Toc495655498)

[2.1.1 Word Sense Disambiguation 6](#_Toc495655499)

[2.2. Related Studies 16](#_Toc495655500)

[CHAPTER 3 Research Methodology 24](#_Toc495655501)

[3.1. RESEARCH DESIGN 24](#_Toc495655505)

[3.2. SOURCES OF DATA 24](#_Toc495655506)

[3.3. INSTRUMENTATION 24](#_Toc495655507)

[3.3.1 SOFTWARE/HARDWARE TOOLS 24](#_Toc495655508)

[3.3.1.1 SYSTEM ARCHITECTURE 25](#_Toc495655509)

[3.3.2 DEVELOPMENT DETAILS 27](#_Toc495655510)

[3.3.3 RESEARCH INSTRUMENT 27](#_Toc495655511)

[3.4. DATA GENERATION/GATHERING PROCEDURE 28](#_Toc495655512)

[3.5. STATISTICAL TREATMENT OF DATA 29](#_Toc495655513)

[REFERENCES 31](#_Toc495655514)

[APPENDIX 34](#_Toc495655515)

[APPENDIX A: SAMPLE RESEARCH INSTRUMENT 34](#_Toc495655516)

[APPENDIX B: COMMUNICATIONS 34](#_Toc495655517)

[APPENDIX C: SCREENSHOTS 34](#_Toc495655518)

# CHAPTER 1 The Problem and Its Background

## BACKGROUND OF THE STUDY

In computational linguistic Word Sense Disambiguation is an open problem of natural language processing and ontology. Accuracy of current algorithms is difficult to state without a host of qualifications. In English, accuracy at the coarse-grained level is routinely above 90%, with some methods on particular homographs achieving over 96%. On finer-grained sense distinctions, top accuracy from 59.1% to 69% have been reported in recent evaluation exercises (SemEval 2007).

Word Sense Disambiguation (WSD) was first formulated into a distinct computational task during 1940s.During the 1970s WSD was a subtask of semantic interpretation systems developed within the field of artificial intelligence. A knowledge based or dictionary based disambiguation became available in the 1980s. In the 1990s the rising of statistical revolution in computational linguistics led to the use of supervised machine learning techniques in word disambiguation. The 2000s view supervised techniques to reach its peak and so make their attention shifter to coarser-grained senses, domain adaption, or combination of different methods.

There are two main approaches for Word Sense Disambiguation. The first approach is Supervised WSD wherein machine learning techniques is used to learn classifier from labeled training sets which are encoded in the form of features with their corresponding sense label or class. A supervised disambiguation requires a tagged corpus with semantic senses, which means each occurrence of ambiguous word is annotated with sense. These tagged with senses corpora are usually made by human expert with sense. The task of Naïve Bayes Classifier is to build a tool which correctly classifies a new word. The second approach is Unsupervised WSD which is based on unlabeled corpora, and does not make use of any manually sense-tagged corpus to give a sense for a word in context.

Unsupervised methods have the potential to overcome the knowledge acquisition bottleneck. These approaches to WSD are based on the idea that the same sense of a word will have similar neighboring words. They are able to induce word senses from input text by clustering word occurrences, and then classifying new occurrences into the induced clusters. They do not rely on labeled training text and, in their purest version, do not make use of any machine–readable resources like dictionaries, thesauri, ontologies. (Navigli, 2009)

Supervised methods undoubtedly perform better than other approaches. However, relying on the availability of large training corpora for different domains, languages, and tasks is not a realistic assumption. Due to the lack of large sense tagged corpora (as well as to the difficulty of manually creating them), the use of these kind of methods is still very limited. (Guzman-Cabrera et al. 2009)

The researcher proposed a Semi-supervised technique as an approach to WSD. This method begins by identification for a small number of training context. This could be accomplished by hand tagging with senses the context of *w* for which the sense of *w* is clear because some seed some collocations occur in these contexts. This annotation is made on the base of the dictionaries or by using WordNet (for English). The initial set of annotated contexts is used for learning a Naïve Bayesian Classifier. By repeating the process, the annotated part of corpus grows.

## STATEMENT OF THE PROBLEM

The study determined the potential of using semi-supervised learning based on Label Propagation. Specifically, it answered the following question:

1. What is the performance of Filipino Word Sense Disambiguation based on Label Propagation in terms of:

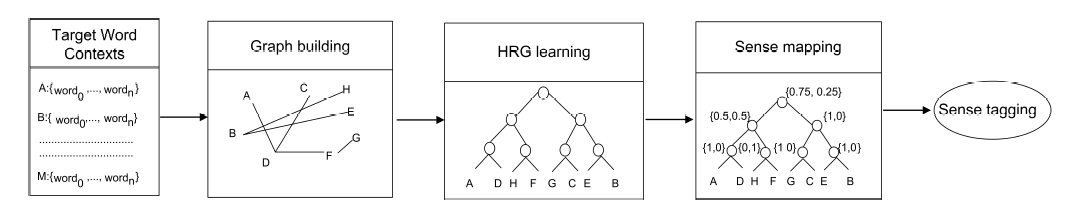
a. PRECSION

b. RECALL

c. F-MEASURE

## THEORETICAL/CONCEPTUAL FRAMEWORK

## THEORETICAL FRAMEWORK



**Figure 1.1 A method flow for Word Sense Disambiguation for Polysemous Word based on the study of** (Klapaftis & Manandar, 2010)

In the study of (Klapaftis & Manandar, 2010), “A Word Sense Disambiguation Using Hierarchical Graphs ” used an unsupervised method for inferring the hierarchical grouping of the senses of a polysemous word. The inferred hierarchical structures are applied to the problem of word sense disambiguation.

## CONCEPTUAL FRAMEWORK OF THE STUDY

Independent Variable

HRG Based Word Sense Disambiguation

Dependent Variable

Accuracy Word Sense Disambiguation

**Figure 1.2 Conceptual Framework of the Study**

Figure 1.2 shows that the independent variable will be Word Sense Disambiguation which will be chosen by the researchers because the aim of the researchers is to explore how the emoticon detection will influence sarcasm detection in terms of accuracy the dependent variable of the study.

## SIGNIFICANCE OF THE STUDY

The application, once completed will help the following beneficiaries:

Word Sense Disambiguation for Filipino Language Researcher – future researcher can further validate the results of this research. In addition to that researcher will be more knowledgeable on what will be the result with a hybrid approach such as knowledge based approach on word Sense Disambiguation.

Machine Translation for Tagalog Language – the result of this study could be used as a tool for the future researcher that is interested in Machine Translation or a construction of Filipino Language just like a study done by (Llanes & Ramos, FiCoBu: A Semi-Supervised Construction of Filipino WordNet Through Web Crawling Using Decision Tree Learning and Language Modelling for Synonyms and Part of Speech, 2016)

## SCOPE AND LIMITATION OF THE STUDY

This study is limited only in using Filipino language based in Tagalog for Word Sense Disambiguation. It is limited only to the use of different Filipino word Homonyms. It will utilize Naïve Bayesian Classifier as an approach for Semi-Supervised Learning and to increase the accuracy of the system. The scope of this approach is based on the study of (Klapaftis & Manandar, 2010) which uses a Hierarchical Random Graph. The researcher will use this approach for Filipino Language as a bases for Label Propagation for each polysemous word in Filipino Language.

## OPERATIONAL TERMS

* Semi-Supervised Learning-  a class of [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) tasks and techniques that also make use of unlabeled [data](https://en.wikipedia.org/wiki/Data) for training – typically a small amount of [labeled data](https://en.wikipedia.org/wiki/Labeled_data) with a large amount of unlabeled data.
* Filipino Language-), in this usage, refers to the [national language](https://en.wikipedia.org/wiki/National_language) (*Wikang pambansa*/*Pambansang wika*) of the [Philippines](https://en.wikipedia.org/wiki/Philippines).
* Word Sense Disambiguation- identifying which [sense](https://en.wikipedia.org/wiki/Word_sense) of a [word](https://en.wikipedia.org/wiki/Word) (i.e. [meaning](https://en.wikipedia.org/wiki/Meaning_(linguistics))) is used in a [sentence](https://en.wikipedia.org/wiki/Sentence_(linguistics)), when the word [has multiple meanings](https://en.wikipedia.org/wiki/Polysemy).
* Hierarchical Random Graph – random graph whose definition encodes some information about the expected community structure of the graph. The definition is given in terms of a dendrogram which gives information about the probability that any two nodes in the graph are connected.
* Label Propagation – is an algorithm for describing the structure of the network in terms of community. It has advantages in terms of running time and amount of prior information needed about the network structure

# CHAPTER 2 Review of Related Literature



## Review of Related Literature

WSD is a hard task as it deals with the full complexities of language and aims at identifying a semantic structure from apparently unstructured sources of text. Research in the field of WSD has been conducted since the early 1950s. Throughout the decade different approach was used by different research to cope with AI-complete problem of Word Sense Disambiguation.

## 2.1.1 Word Sense Disambiguation

Word sense disambiguation (WSD) is the ability to identify the meaning of words in context in a computational manner. WSD is considered an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence.

The difficulty of WSD does not originate from single cause but from a variety of factors. One factor is that there are many approach in the representation of word sense like enumeration of set of senses, and rule based generation of a new word senses. The second factor is that WSD heavily relies on a bunch source of knowledge which can vary from corpora of texts which can be either annotated with word senses, to more structured resources such as machine-readable dictionaries.

Word sense disambiguation is the ability to computationally determine which sense of a word is activated by its use in a particular context. WSD is usually performed on one or more texts (although in principle bags of words, i.e., collections of naturally occurring words, might be employed.

There are two main techniques used in WSD: the supervised learning approach and the supervised learning approach. Supervised learning approaches use machine learning techniques to learn a classifier from a labeled training sets of examples encoded in terms of a number of features together with their appropriate sense label. Unsupervised learning approaches are mostly based on unlabeled corpora, and do not exploit any manually sense-tagged corpus to provide a sense choice for a word context.

According to Navigli, 2009 supervised methods undoubtedly perform better than other approaches. However, relying on the availability of large training corpora for different domains languages and task is not a realistic assumption. Ng, 1997 stated that to obtain high accuracy wide coverage disambiguation system, probably need a corpus of about 3.2 million sense tagged words. The human effort for constructing such training corpus can be estimated to be 27 person-years, at a throughput of one word per minute.

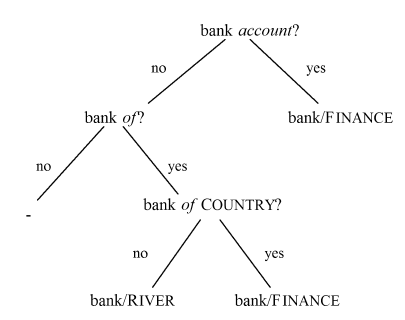
**2.1.1.1 Supervised Word Sense Disambiguation**

Supervised WSD uses machine-learning techniques for inducing a classifier from manually sense-annotated data sets. Usually, the classifier (often called *word expert*) is concerned with a single word and performs a classification task in order to assign the appropriate sense to each instance of that word. The training set used to learn the classifier typically contains a set of examples in which a given target word is manually tagged with a sense from the sense inventory of a reference dictionary. Supervised approaches to WSD have obtained better results than unsupervised methods.

The oldest approach in the field of Supervised learning is the Decision Tree Algorithm. It is an ordered set of rules for categorizing test instances (in the case of WSD, for assigning the appropriate sense to a target word). It can be seen as a list of weighted “if-then-else” rules. A training set is used for inducing a set of features. As a result, rules of the kind (*feature-value*, *sense*, *score*) are created. The ordering of these rules, based on their decreasing score, constitutes the decision list. Given a word occurrence *w* and its representation as a feature vector, the decision list is checked, and the feature with highest score that matches the input vector selects the word sense to be assigned:

ˆ*S* = argmax*Si*∈*SensesD*(*w*) *score*(*Si*)*.*

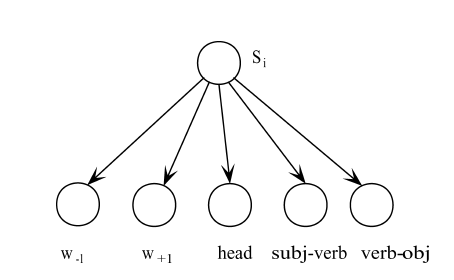
Another approach under supervised learning is the *decision tree* is a predictive model used to represent classification rules with a tree structure that recursively partitions the training data set. Each internal node of a decision tree represents a test on a feature value, and each branch represents an outcome of the test. A prediction is made when a terminal node (i.e., a leaf) is reached. Even though they represent the predictive model in a compact and human-readable way, they suffer from several issues, such as data sparseness due to features with a large number of values, unreliability of the predictions due to small training sets.

****

**Figure 1.4 An example of Decision tree for a word “*bank*”**

Naïve Bayes Classifier is a simple probabilistic classifier based on Bayes’ Theorem. It relies on the calculation of the conditional probability of each class of a vector given the features in the instances.

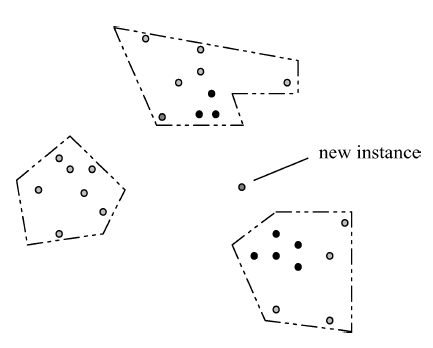
(the denominator is also discarded as it does not influence the calculations). The probabilities and are estimated, respectively, as the relative



**Figure 1.5 An example of Bayesian Network**

occurrence frequencies in the training set of sense *Si* and feature *f j* in the presence of sense. Zero counts need to be smoothed: for instance, they can be replaced with where *N* is the size of the training set (Ng 1997; Escudero et al. 2000).

Exemplar-based (or instance-based, or memory-based) learningis a supervised algorithm in which the classification model is built from examples. The model retains examples in memory as points in the feature space and, as new examples are subjected to classification, they are progressively added to the model.



**Figure 1.6 An example of kNN classification on a bidimensional plane.**

In kNN the classification of a new example represented in terms of its *feature* values, is based on the senses of the *k* most similar previously stored examples. The distance between and every stored exampleis calculated, for example, with the Hamming distance

,

where the weight of the feature. The set of theclosest instances is selected and the new instance is predicted to belong to the class assigned to the most numerous instances within the set.

**2.1.1.2 Unsupervised Word Sense Disambiguation**

Unsupervised methods have the potential to overcome the knowledge acquisition bottleneck [Gale et al. 1992b], that is, the lack of large-scale resources manually annotated with word senses. These approaches to WSD are based on the idea that the same sense of a word will have similar neighboring words. They are able to induce word senses from input text by clustering word occurrences, and then classifying new occurrences into the induced clusters. They do not rely on labeled training text and, in their purest version, do not make use of any machine–readable resources like dictionaries, thesauri, ontologies, etc. However, the main disadvantage of fully unsupervised systems is that, as they do not exploit any dictionary, they cannot rely on a shared reference inventory of senses.

A first set of unsupervised approaches is based on the notion of *context clustering*. Each occurrence of a target word in a corpus is represented as a *context vector*. The vectors are then clustered into groups, each identifying a sense of the target word. A word *w* in a corpus can be represented as a vector whose *j* th component counts the number of times that word *wj* cooccurs with *w* within a fixed context (a sentence or a larger context). The underlying hypothesis of this model is that the distributional profile of words implicitly expresses their semantic. The similarity between two words *v* and *w* can then be measured geometrically, for

example, by the cosine between the corresponding vectorsA vector is computed for each word in a corpus. This kind of representation conflates senses: a vector includes all the senses of the word it represents.

Each instance constitutes a singleton cluster. Next, agglomerative clustering merges the most similar pair of clusters, and continues with successively less similar pairs until a stopping threshold is reached. The performance of the agglomerative clustering of context vectors was assessed in an unconstrained setting and in the biomedical domain (Savova et al. 2005). However problem in the construction of context vectors is that a large amount of (unlabeled) training data is required to determine a significant distribution of word cooccurrences.

A different view of word sense discrimination is provided by graph-based approaches, which have been recently explored with a certain success. These approaches are based on the notion of that is, a graph whose vertices correspond to words in a text and edges *E* connect pairs of words which cooccur in asyntactic relation, in the same paragraph, or in a larger context. The construction of a cooccurrence graph based on grammatical relations between words in context was described by Widdows and Dorow (2002)

Given a target ambiguous word a local graph is built around. By normalizing the adjacency matrix associated with.

As a result, all its neighbors are no longer eligible as hub candidates. The algorithm stops when the relative frequency of the word corresponding to the selected hub is below a fixed threshold. The entire set of hubs selected is said to represent the senses of the word of interest. Hubs are then linked to the target word with zero-weight edges and the minimum spanning tree (MST) of the entire graph is calculated.

**2.1.1.2 Semi-Supervised Word Sense Disambiguation**

Semi-supervised approach to word sense disambiguation uses both annotated and non-annotated data for finding the correct sense of the word.

**2.1.1.2.1 Bootstrapping**

The aim of *bootstrapping* is to build a sense classifier with little

training data, and thus overcome the main problems of supervision: the lack of

annotated data and the data sparsity problem. Bootstrapping usually starts from few annotated data *A*, a large corpus of unannotated data *U*, and a set of one or more basic classifiers. As a result of iterative applications of a bootstrapping algorithm, the annotated corpus *A* grows increasingly and the untagged data set *U* shrinks until some threshold is reached for the remaining examples in *U*. The small set of initial examples in *A* can be generated from hand-labeling or from the automatic selection with the aid of accurate heuristics (Yarowsky 1995).

There are two main approaches to bootstrapping in WSD: *cotraining* and *selftraining*. Both approaches create a subset unlabeled examples chosen at random. Each classifier is trained on the annotated data set *A* and applied to label the set of examples. As a result of labeling, the most reliable examples are selected according to some criterion, and added to *A*. The procedure is repeated several times (at each iteration includes a new subset of random examples from . Within this setting, the main difference between cot raining and self-training is that the former alternates two classifiers, whereas the latter uses only one classifier, and at each iteration retrains on its own output. An example of use of these methods in WSD was presented by Mihalcea (2004), where the two classifiers for cot raining use local and topical information, respectively, and a self-training single classifier combines the two kinds of information.

**Filipino Language**

Filipino language is the official name of tagalog it is best described as “Tagalog based”. Tagalog is used to differentiate it from the other language or dialects in the Philippines and Filipino is used to differentiate it from the language of other countries. Pedro de San Buenaventura wrote the first Tagalog Dictionary and successfully published on 1613. At the beginning of eighteenth century a latter book of the same name was written by [Czech](https://en.wikipedia.org/wiki/Czech_people) [Jesuit](https://en.wikipedia.org/wiki/Jesuit) missionary [Paul Klein](https://en.wikipedia.org/wiki/Pablo_Clain).

Tagalog Central Philippine Language within Austronesian family. Being Malayo-Polynesian it is related to other Austronesian languages, usch as Malagasy, Javanesse, (Malay and Indonesian), Tetum, and Yami. It is closely related to the languages spoken in Bicol Region and the Visayan Island, such as Bikol Group and Visayan Group including Hilagaynon and Cebuano.

According to the Philippine Statistics Authority, as of 2014 there were 100 million people living in the Philippines, where almost all of whom will have some basic level of understanding of the language. The tagalog homeland, Kataglugan covers roughly much of the central to southern parts of the island of Luzon. Particulary in Aurora, Bataan, Batangas, Bulacan, Cavite, Laguna, Metro Manila, Nueva Ecija, Rizal, and Large parts of Zambales. Tagalog is also spoken natively by inhabitants on the islands of Marindoque, Mindoro, and Medium areas of Palawan. It is spoken by approximately 64 million Filipinos, 96% of the household population; 22 million, or 28% of the total Philippine population, speak it as a native language.

Many Tagalog dialects, particularly those in the south, preserve the glottal stop found after consonants and before vowels. This has been lost in Standard Tagalog. For example, standard Tagalog *ngayon*. In many [southern dialects](https://en.wikipedia.org/wiki/Batangas_Tagalog), the progressive aspect infix of *-um-* verbs is *na.* For example standard Tagalog kumakain(eating) is *nakain* in Quezon and Batangas Tagalog ask *nakain ka ba ng Pating?* (“Do you eat Shark”). Would be understood as saying “Has a shark eaten you?” by the speaker of Manila Dialect. Some dialects have interjections which are considered a regional trademark. For example, the interjection *ala e!* usually identifies someone from [Batangas](https://en.wikipedia.org/wiki/Batangas" \o "Batangas) as. does *hane?!* in Rizal and Quezon provinces.

A homograph is one of a group of words that share the same spelling but have different meanings. Thus, a Filipino words with the same spelling but have different meanings or English translation are homographs. Take for instance the word *buhay.* If it is used as noun, it means *alive.* Thus, this words convey a different meanings depend on how you used it in the sentence.

## Related Studies

* + 1. **Application of Supervised Learning in Word Sense Disambiguation**
       1. **An Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation System.**

was used as a learning algorithm as a simple representation of statistical learning method. It used the semantically annotated DSO corpus containing sentences from two different corpora, namely Wall Street Journal(WSJ) and Brown Corpus. Both the corpora were both trained in their respected domain and tested cross sectional to test its portability. The trained WSJ corpus was tested in Base Corpus and vice versa. The corpora contained a group of twenty-one words which frequently appear in the WSD literature to perform the comparative experiment by treating its words as different classification problem. For the disambiguation they use two kinds of information: local text and topical task. In this manner, local context attributes have to be binarized in a process, while the topical context attributes remain as binary tests about the presence or absence of a concrete word in the sentence.

The result of the study showed that the performance of sense taggers was not guaranteed when moving from one domain to another since it used two types of corpus (WSJ and BC) so there is a need for an adaptation for cross-corpus application.

* + 1. **Application of Unsupervised Learning in Word Sense Disambiguation**
       1. **A Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge.**

This study by (Chen, Bowes, Ding, & Brown) used a knowledge based approach but used an unsupervised learning for extracting information from the text corpus using Dependency Knowledge Method. It is based on the insight that.

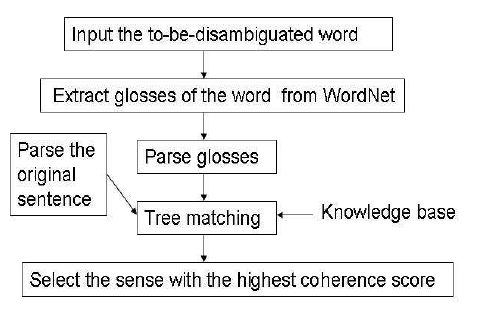
*“If a word is semantically coherent with its context then at least one sense of this word is semantical coherent with its context.*

The corpus building was done through Web Searches to collect as many as possible valid sample sentences for each instance. Each word that need to be disambiguated are compiled and saved in a text file. Merging for dependency relation was straightforward using node dependency. One head or node and one dependent word/node or the child node. WSD was done using tree matching function that matches the dependent words of to-be-disambiguated word.

“*Computer Programmer write software”*

*“Many companies hire computer programmer”*

The resulting outputs of the experiment showed that accuracy is directly proportional to the number of word instances. After merging dependency relations, the weight was obtained for each directed graph with word as node and dependency relation as an edge. All the weights are assigned to the nodes/ words in the parsing trees. The original parsing sentence of the weight of the node is reciprocal to the distance between its node and to-be disambiguated node/word. In the parsing tree of a gloss that was retrieved using WordNet is also reciprocal to the level of its node in the parsing tree. The indicates the strength of semantic relevancy of head word and dependent word. The gloss most semantically coherent with the original sentence will be chosen as the correct sense. The combination of more matching strategies results in disambiguation quality.



**Figure 1.10 System flow for dependency Knowledge based WSD**

The method is purely based on dependency of the word in the sentences need to-be disambiguated and does not consider the surrounding words that contribute to the true meaning of the target word. The performance of WSD could be improved by further collection and incorporation of more context knowledge.

* + - 1. **Automatic Approach for Word Sense Disambiguation Using Genetic Algorithm.**

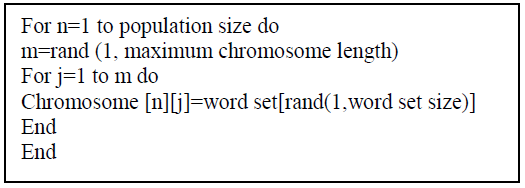
Utilized Unsupervised Learning in Word Sense Disambiguation using Genetic Algorithm. This study also web as source of data. Words that requires disambiguation used as a query to retrieve great number of Web Pages. Web Pages are cleaned and stored into a text file. After pre-processing Genetic algorithm was implemented.

1. *Population Representation*

GA runs on a number of potential solutions, termed as population, containing many encoded parameters. The for each individual in the population.

1. *Pick a random number representing the length of the chromosome in the population.*
2. *Chose words randomly from the set of word resulted from the preprocessing step in number equal to the length of the individual.*
3. *Repeat steps a and b until the population size reached.*

The fitness function of each individual measures how many documents covered by a particular word represented in that individual in another word. The actual meaning of the word was represented by the number of senses generated by the algorithm.



**Figure 1.11Pseudo Code for population selection for each senses of word**

The termination of algorithm can be specified by the number of generations and then all the individuals in the population that have fitness function more than specified threshold represent the discovered sense of the word.

* + 1. **Application of Semi-Supervised Learning Word Sense Disambiguation**
       1. **Word Sense Disambiguation of Opinionated Word Using Extended Gloss Overlap**

The study focused on integrating word sense disambiguation into the opinion classification task, and seeing how this will affect the overall results. The algorithm was integrated using VoxPop, as system that detects opinions in commentaries and classifies these by topic and polarity. Using machine learning features for WSD. Terms are labeled as subjective or objective depending on how it is used in the sentence, and whether it has an impact on the overall. Two sets of data were used which consist of commentaries taken from Inbox World section of the Philippine Stars. These commentaries were manually classified as being positive or negative. Then the linguist evaluation was compared with the results yielded by the system. There is a lot of factors that contributed to the low improvement of the system.

1. Certain in commentary may have polarities that are too high, rendering the polarities of other words.

“*Maybe this could be attributed t Chief Justice Puno’s*

*relentless effort in cleansing the ranks of the Judiciary*”.

The linguist classified this commentary as positive , while the system classified them as negative.

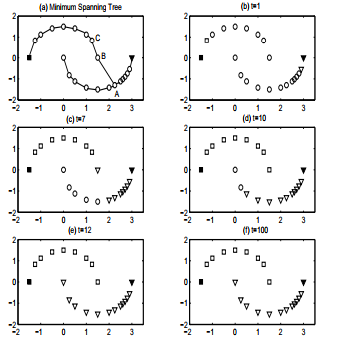
Words such as *not, too, should* and *although* change the polarity of the words next to them.

1. The resources used also have their own deficiencies such as words tagged with incorrect part-of-speech.
2. The word sense disambiguation algorithm assigned the incorrect sense to a word.

One way to improve the system is the creation of resources to help process commentaries that are written in both English and Tagalog. The wordlist should include both English and Filipino terms used in commentaries and other forms of media. There are many aspects that are not covered by VoxPop formula since it does not take into account how combination of words can have a different meaning as opposed to when they are taken as separate lexical item.

* + - 1. **Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning.**

(Dong Hong-Ji & Chew-Lim, 2005)uses LP algorithm wherein label information of any vertex in a graph is propagated to nearby vertices through weighted edges until a global stable stage is achieved. Larger edge weights allow labels to travel through easier. Thus, the closer the examples, more likely they have similar labels.

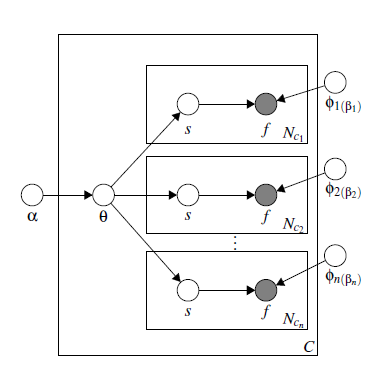


**Figure 1.12Classification result of LP on two-moon pattern dataset. (a) Minimum spanning tree of this dataset. The convergence process of LP algorithm with t varying from 1 to 100 is shown from (b) to (f).**

During the propagation process he soft label of each initial labeled example is clamped in each iteration to replenish label sources from these labeled data. Thus, the labeled data act like sources to push out la bels through unlabeled data With this push from labeled examples, the class boundaries will be pushed through edges with large weights and settle in gaps along edges with small weights. In learning process, the labels of unlabeled examples are determined not only by nearby labeled examples, but also by nearby unlabeled examples.

* + - 1. **Bayesian Word Sense Induction**

Sense induction is typically treated as an unsupervised clustering problem. The input to the clustering algorithm are instances of the ambiguous word with their accompanying contexts (represented by co-occurrence vectors) and the output is a grouping of these instances into classes corresponding to the induced senses (Brody & Lapata, 2009).

For each ambiguous word the distribution was drawn over senses, and then generate context words according to this distribution. Assuming that different senses will correspond to distinct lexical distributions. In this framework, sense distinctions arise naturally through the generative process. The words in the document are generated by repeatedly sampling a topic according to the topic distribution, and selecting a word given the chosen topics.

**Figure 1.13 Extended sense induction model; inner rectangles represent different sources (layers) of information.**

The solution is to treat each information source individually and then combine all of them together in a unified model. The underlying assumption is that the context window around the target word can have multiple representations all of which share the same sense distribution. independence between multiple layers was naively assume even though this is clearly not the case in our task Sense induction was placed in a probabilism probabilistic setting by modeling the context words around the ambiguous target as samples from a multinomial sense distribution. Each element in each layer is a variable, and is assigned a sense label where distinct layers correspond to different representations of the context around the target word. The sense distribution of the instance must be determined as a whole.

The procedure begins by randomly initializing all unobserved random variables. At each iteration, each random variable (senses of the word) is sampled from the conditional distribution P(si|s−i) where s−i refers to all variables. The distribution over samples drawn from this process will converge to the unconditional joint distribution P(s) of the unobserved variables. Each element in each layer is a variable, and is assigned a sense label.

* + - 1. **AI-KU: Using Substitute Vector s and Co-Occurrence Modeling for Word Sense Induction and Disambiguation.**

(Baskaya, Sert, & Volkan, AI-KU Using Substitute Vector and Co-Occurrence Modelling for Word Sense Induction and Disambiguation, 2013) Uses statistical language model to represent the context of the word senses. They developed a system that represents the context of each target words by using high probability substitutes according to statistical language model. These substitute words and their probabilities are used to create word pairs to feed our co-occurrence model. The output of the co-occurrence model is clustered by k-means algorithm.

The description of data enrichment procedure will answer each instances substitute vector constructed. The pairs of pairs of word and its meaning were used in the co-occurrence modelling by using substitute word sampling. Then co-occurrence model were clustered using the k-means clustering algorithm. The use of data enrichment aims to increase the number of instances of target words.

* 1. **Synthesis of the Studies**

Supervised methods undoubtedly perform better than other approaches. However, relying

on the availability of large training corpora for different domains, languages, and tasks is not a realistic assumption (Navigli, 2009). The use of unlabeled data is the main goal of Semi-Supervised approach to WSD. Semi-supervised uses to mapping function the first function that uses inference rule to find the hidden sense among a group of text or corpus while the latter uses a mapping function to find the hidden sense among group of words.

Word Sense Induction as stated in the study of (Baskaya, Sert, & Volkan, AI-KU Using Substitute Vector and Co-Occurrence Modelling for Word Sense Induction and Disambiguation, 2013) aims to discover different senses of word from a corpus by using unsupervised approach The main compensation of Semi-supervised approach is overcoming the bottleneck of labeled data scarcity for training. With this information the researcher came up with the study of using the Word Sense Induction as a constituent to Word Sense Disambiguation by the implementation of label propagation (Klapaftis & Manandar, 2010).

# CHAPTER 3 Research Methodology



## RESEARCH DESIGN

The researchers used the True Experimental Research Design. It is a blueprint of the procedure that enables the researcher to test his hypothesis by reaching valid conclusions about relationships between independent and dependent variables [True Experimental Design 2016]. It can also be used to describe all studies with at least one independent variable that is experimentally manipulated and with at least one dependent or outcome variable [Dawes 2010]. The researchers used this research design to manipulate the independent variable to observe its effect on some of the behavior or cognitive process (dependent variable) while using random assignment/s of participants to groups to control external factors from influencing the results. In doing this, the researchers attempt to determine or predict what may occur [Key 1997].

## SOURCES OF DATA

The population of the study consist of the data that will be coming from a sentence with ambiguous words. The Tagalog-English Parallel Corpus will be used as a Training Data. The corpora that will be used are also called bilingual corpora, one serving as a primary language and the other as secondary language a bilingual corpus was used since different senses of some words often translate differently in another language. In our experiment we used Filipino-English parallel corpus as a training. There is no Filipino – English Corpus that contain ambiguous words manually. So we are going to create Filipino-English corpus with ambiguous words.

## INSTRUMENTATION

## SOFTWARE/HARDWARE TOOLS

The researchers will be using an experiment paper to determine the accuracy of the system by finding its Precision, Recall and F1-Score. This will be used for each test case to determine the percentages for each sentiment. Once the paper is completed, we count the total number for each state and then we can compute the final Precision, Recall and F1-Score.

### SYSTEM ARCHITECTURE

Pre-processing

Base Corpus

Text Normalization

Tokenization

Sentence Splitter

Word Sense Induction

(AI-KU)

WSD

Induced Sense

Calculate Probability

Training Data

Classifier

Sense Definition

Sense Mapping Function

Reference Inventory

SEALang Dictionary

Label Propagation

Correct Sense Meaning

**Figure 3.1System Architecture for Word Sense Disambiguation**

Figure 3.1 shows the block diagram of the designed system. There are two major blocks used in the system. First is Pre-processing Block that deals with tokenization Sentence Splitting and Text Normalization. The second block is the Disambiguation Block which deals inducing the correct meaning of the words.

* 1. **Pre-processing**

**3.1.1. Sentence Splitter**

For complex sentences sentence splitter is needed to remove conjunction such as *and, but, although,* and *while* using the NLTK Python Library.

**3.1.2. Text Normalization**

Each stop word in the sentences will be removed for easy processing of text.

Stop words such as period ‘.’ , semicolon ‘;’ question mark‘?’ and exclamation point ‘!’ will be removed.

**3.1.3. Tokenization**

For efficiency each word will be tokenized. A mnemonic value will be assigned for each word.

**3.2. Word Sense Induction**

For the word Sense Induction we will use the AI-KU tools by (Baskaya, Sert, & Volkan, AI-KU Using Substitute Vector and Co-Occurrence Modelling for Word Sense Induction and Disambiguation, 2013).

**3.3. Label Propagation**

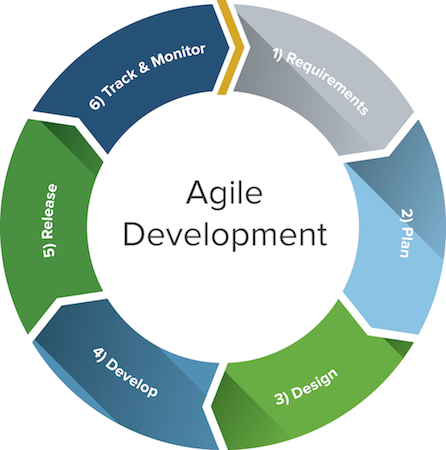
Using the induced sense generated by AI-KU tool together with the SEALang dictionary which provide as a substitute for WordNet. The training data which consist of unlabeled Tagalog text will be trained by implementing the SVM algorithm. This tools will be used in building a classifier a statistical that will be used WSD engine.

**3.4. Word Sense Disambiguation**

The WSD engine will calculate the probability of the given ambiguous word. Then the correct sense meaning will be retrieved Using the SEALang Dictionary.

### DEVELOPMENT DETAILS

Agile Method will be used for the Development of the System for a quick development, resourceful and adaptive to change in the design of the system. The software development will be divided into two phase the development of each module and the integration of the module. Agile development model is suitable for software project that is prone to change in plan in different stages of development.



**Figure 3.2Agile Development Model**

### RESEARCH INSTRUMENT

The proponents will provide a set of training data and sample set of Filipino articles in text files that will be used in the training and testing phase of the study. The proponents will use Python Programming Language and NLTK Library in creating the tool. A researcher-made test plan will serve as the tool for gathering data in the testing and implementation phase of the study. There will be an experiment paper to be used for Word Sense Disambiguation. There will be two sets of data: the Training Data and the Test Data. The procedures are there will be a set manually annotated polysemous word with its corresponding pseudosense . Then, the proponents will count the number of retrieved, retrieved, number of correct sense, and incorrect. Afterwards, it will compute Recall, Precision, and F-Measure. And lastly, the proponents will average the F-measure the system.

1. **Experiment Paper**

Each Sentence with polysemous word in the test data serve as the input of the system.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sentence with Polysemous word | NE | RNE | UNE | CT | WT | Recall | Precision | F-Measure |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

.

## DATA GENERATION/GATHERING PROCEDURE

The training data for word sense induction will be generated by manual annotating more than 140 Tagalog polysemous word with pseudo sense. The data generated by the WSI will serve as the co-training data for label propagation together with large training data that will be get from SEALang Corpus an online Tagalog dictionary and Corpus. The data generated by the label propagation will be used for building a classier that will b used for word sense disambiguation.

The test data consist of 200 sentences with polysemous. This sentence will be loaded in the system. The researcher recorded the result of the experimentation. The system’s classification will be evaluated using statistical tools.

## STATISTICAL TREATMENT OF DATA

In a lecture by Christopher Manning of Stanford University (Manning, 2012) about “Evaluation of text Classification”. He stated that the concepts of getting the Precision Recall and F1-Score are way of evaluation in different task of Natural Language Processing. According also to an article by Jacob Perkins (Perkins, 2010) for StreamHacker entitled “Text Classification for Sentiment Analysis – Precision and Recall”. Accuracy is not the only metric for evaluating the effectiveness of a classifier. Two other useful metrics are precision and recall. These two metrics can provide much greater insight into the performance characteristics of a binary classifier.

The performance of the system would be also measured using the F measure or F1 score. F1 score refers to the harmonic mean of precision and recall (F1 Score, 2016). The formula for the F1-score is as follows

Precision is the fraction of retrieved instances that are relevant. It also called as the positive predictive value (Precision and Recall, 2016)). The formula for the Recall is as follows:

In a lecture by Dan Jurafsky of Stanford University (Jurafsky, 2012) about “Precision recall and the F measure”. He presented a contingency table where he defined how to classify the results of a test into True Positives, True Negatives, False Positives and False Negatives

|  |  |  |  |
| --- | --- | --- | --- |
|  | | System | |
| **Correct** | **Not Correct** |
| Human | **Correct** | True Positive (TP) | False Negative (FN) |
| **Not Correct** | False Positive (FP) | True Negative (TN) |

**Table 3.1: 2-by-2 Contingency Table by** (Jurafsky, 2012)

As seen at table 3.1 True Positive (TP): Number of sentiments classified positively correct by both system and human while True Negative (TN): Number of sentiments that classified positively incorrect by both system and human.

False Positive (FP): Number of sentiments classified correctly by system but Incorrect by human. While False Negative (FN): Number of sentiments classified correctly by human but incorrect or neutral by system. Table 3.2 show the truth table in determining the right statistical variable for each result.

**Table 3.1 Truth Table**

|  |  |  |  |
| --- | --- | --- | --- |
| Human Answer | System Answer | Positive | Negative |
| Correct | Correct | TP | TN |
| Not Correct | FN | FP |
| Not Correct | Correct | FP | FN |
| Not Correct | TN | TP |

In 2013, (Eboña, et al., 2013) generated a rating system for the parameters: precision, recall and F1 score. The researchers will be using this interpreting the values from the parameters measured. The researchers will be using the rating system for precision, recall and F1 score. The rating system is shown on Table 3.3.

In evaluating the sentiment analyzer and sarcasm detection module of the system, the researchers would be using an experiment paper to be checked by the administrators that also tag the word sense seed.

# REFERENCES

Alok, R. P., Anirban, K., Singh, A., Raj, S., & Sinha, K. (2013). A Hybrid Approach to Word Sense Disambiguation Combining Supervised and Usuopervised Learning. *International Journal of Artificial Intelligence & Applications*, 1-13.

Alsaidi, B. K. (2016). Automatic Approach for Word Sense Disambiguation Using Genetic Algorithms. *International Journal of Advance Computer Science and Applications. Vol. 7, No.1 2016*, 41-44.

Baskaya, O., & Jurgens, D. (2016). Semi-supervised Learning with Induced Word Senses for State of the Art Word Sense Disambiguation. *Journal of Artificial Intelligence Research*, 1-34.

Baskaya, O., Sert, E., & Volkan, C. (2013). AI-KU Using Substitute Vector and Co-Occurrence Modelling for Word Sense Induction and Disambiguation. *In Proceeding of the Seventh International Workshop on Semantic Evaluation*, 300-306.

Brigole, J., & Roxas, R. (2009). A FIlipino-English Dictionary Designed for Word Sense Disambiguaton. 1-5.

Brody, S., & Lapata, M. (2009). Bayesian Word Sense Induction. *In Proceedings of the 12th Conference of the European Chapter of the Association for the Computational Linguistic(EACL)*, 103-111.

Chen, P., Bowes, C., Ding, W., & Brown, D. (n.d.). Fully Unsupervised Word Sense Disambiguation Using Dependency Method. 28-36.

Corazon, B. R., & Cheng, C. (2011). Word Sense Disambiguation of Opinionated Word Words Using Extended Gloss Lap. 1-4.

D, Y. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association fo Computational Linguistic(Cambridge, MA)*, 189-196.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, Eric Altendorf. (2016). Semi-Supervised Word Sense Disambiguation with Neural Model. 2-8.

Dong Hong-Ji, Z.-Y. N., & Chew-Lim, T. (2005). Word Sense Disambiguation Using Label Propagatin Based Semi-Supervised Learning. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics(ACL)*, 395-402.

Eboña, K. M., O, S., Perez, G. P., Roldan, J. M., Sagum, R. A., & Domingo, I. V. (2013). *Named Entity Recognizer (NER) for Filipino Novel Excerpts using Maximum Entropy Approach.* Undergraduate Thesis , Polytechnic University of the Philippines, Department of Computer Science, Manila City.

Edmonds, P., & Cotton, C. (2001). SenseVal-2: Overview. *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguaton Systems*, 1-6.

Escudro, G., Marquez, L., & Rigau, G. (2005). Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation.

F1 Score. (2016). *F1 Score*. Retrieved September 12, 2016, from Wikipedia: https://en.wikipedia.org/wiki/F1\_score

Guzman-Cabrera, R., Rosso, P., Montes-y-Gomez, M., Villasenior-Pineda, L., & Pinto-Avendano, D. (2009). Semi-supervised Word Sense Disambiguation Using Web as Web Corpus. 1-10.

Jurafsky, D. (2012). *Precision and Recall and F measure.* Standford University Coursera.

Kalyani, M., & Handanwar, L. (2015). Marathi Word Sense Disambiguation Using Genetic Agorithm- A Review. *International Journal of Advanced Computational Engineering and Networking, ISSN: 2320-2106*, 1-3.

Klapaftis, I., & Manandar, S. (2010). Word Sense Induction and Disambiguation Using Hierarchichal Random Graphs. 745-756.

Lapata, M., & F, K. (2007). An Information Retrieval Approach to Sense Ranking . *Proceedings of the Human Language Technology Conference of the Association for Computational Linguistics*, 348-355.

Llanes, M., & Ramos, A. (2016). FiCobu: A Semi-Supervised Constrcution of Filipino WordNet Through Web Crawling Using Decsion Tree Learning and Language Modeling for Synonyms and Part of Speech. 50-52.

Llanes, M., & Ramos, A. (2016). FiCoBu: A Semi-Supervised Construction of Filipino WordNet Through Web Crawling Using Decision Tree Learning and Language Modelling for Synonyms and Part of Speech. 50-52.

Manning, C. (2012). *Evaluation of Text Classification.* Standford University Coursera.

Mihalcea, R., & D, M. (2004). SenseLearner: Minimally Supervised word sense disambiguation for all words in an open text . *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 155-158.

Perkins, J. (2010). *TEXT CLASSIFICATION FOR SENTIMENT ANALYSIS – PRECISION AND RECALL*. Retrieved September 12, 2016, from StreamHacker: http://streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision-recall/

Precision and Recall. (2016). *Precision and Recall*. Retrieved September 12, 2016, from Wikipedia: https://en.wikipedia.org/wiki/Precision\_and\_recall

Resnik, P., & Lin, J. (1990). *Evaluation of NLP Systems.* San Francisco.

Rigau, G., Rodriguez, H., & Agirre, E. (2007). Building Accurate Semantics Taxonmonies from Monolingual MRDS. *In Proceedings of the 17th International Conference on Computational Linguistics*, 1103-1109.

Sagum, R. A., De Vera, J. G., Lansang, P. J., Narciso, D. S., & Respeto, J. K. (2015, March ). Application of Language Modelling in Sentiment Analysis for Faculty Comment Evaluation. *Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2015, 1*.

Savova, G., Petersen, T., Purandare, A., & A., K. (2005). Resolving ambiguities in Biomedical Text with UNsupervised Clustering Approaches.

Vickrey, D., Biewald, L., & Koller, D. (2005). WORD Sense Disambiguation for Machine Translation. 1-7.

Widdows, D., & Dorrow, B. (2005). A graph Model of Unsuopervised Lexical Aquisition. *Proccedings of the 19th International Conference on Computatonal Linguistics*, 1-7.

Yuret, D., & Mehmet, A. Y. (2010). The Noisy Channel Model for Unsupervised Word Sense Disamabiguation. 1-18.

# APPENDIX

## APPENDIX A: SAMPLE RESEARCH INSTRUMENT

The tool/s to be used for collecting information. (Experiment paper or Survey Questionnaire)

## APPENDIX B: COMMUNICATIONS

Letters

## APPENDIX C: SCREENSHOTS

Screen capture of the proposed technology.