Title	Supervised Learning of Semantic Class Disambiguation Classifiers for All Words Task
Author(s)	Ariyakornwijit, Patanan
Citation	
Issue Date	2012-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/10750
Rights	
Description	Supervisor: Shirai Kiyoaki, 情報科学研究科, 修士

Supervised Learning of Semantic Class Disambiguation Classifiers for All Words Task

By Patanan Ariyakornwijit

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Associate Professor Kiyoaki Shirai

September, 2012

Supervised Learning of Semantic Class Disambiguation Classifiers for All Words Task

By Patanan Ariyakornwijit (1010201)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Associate Professor Kiyoaki Shirai

and approved by
Professor Shimazu Akira
Professor Tojo Satoshi

August, 2012 (Submitted)

Acknowledgements

This thesis would not have been possible unless I did consult by my advisor, Assistant Professor Kiyoaki Shirai, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. It is a pleasure to thank those who made this thesis possible. Especially my two seniors: Kobkrit Viriyayudhakorn who is providing me a high performance server to compile the program, and Nuttapong Rattanaponglekha who advise me on difficult programming.

I would like to show my gratitude to JAIST, Asia Jinzai Program, and Japanese Government MEXT of providing me a chance to accomplish the Master Degree together with fully of experiences in Japan.

I would like to thank my lovely family and friends who always support and cheer me up when I feel down. And also Thais at JAIST who make my life in this country-side become enjoyable.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the thesis.

Patanan Ariyakornwijit

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Word Sense Disambiguation (WSD) is the task to finding the right meaning of a word in a given sentence. WSD is one of the important tasks in the Natural Language Processing, which is also significant to other issues such as Machine Translation, Language Understanding and Information Retrieval.

In order to solve WSD problem, many algorithms are proposed. The supervised learning shows better performance than others. But the supervised learning method still suffers from a serious problem, which is the difficulty of preparing training data, also known as: Knowledge Acquisition Bottleneck.

The goal of this research is to develop a method to train classifiers, which can be applicable to all words. The trained classifier can disambiguate the coarse grained sense of a given word in the context. Coarse grained senses are defined as an universal sense set for all words so that classifiers can disambiguate senses for all words, especially low frequent words. We believe that this method would alleviate the Knowledge Acquisition Bottleneck problem.

In the previous work, WSD classifiers are trained for individual target words. Therefore, it is necessary to train a bulk of classifiers in order to disambiguate senses of all words in a text. This is our main motivation of defining senses at the coarse level because these sense definitions, namely semantic class definitions, are common for all words.

Although semantic class disambiguation or the coarse grained WSD is not sufficient for some Natural Language Processing applications, but it is very useful in several applications such as Information Retrieval. For example, the word 'apple contains 3 senses, which are apple as a fruit, apple as a tree, and apple as a company. When we look up the coarse grained level of these 3 senses in WordNet, they would be noun.food, noun.plant and noun.group. As these coarse grained meaning, we can classify and retrieve the needed information.

In this research, we trained classifiers, which can be applicable for all words by using coarse-grained senses of WordNet, namely semantic class, as a common sense definition for disambiguating. We describe related work in Chapter 2. Chapter 3 explains the definition of the coarse grained sense, namely semantic class, the method of classifing the semantic class, and the used training data. We show several experiments and compare their results in Chapter 4. Then, we finally conclude the research and discuss the future work in the Chapter 5.

Chapter 2

Related Work

In Word Sense Disambiguation, using coarse grained senses as word senses has been carried out many times in various methods.

Levin proposed classification of English verbs[1]. She classified over 3,000 English verbs with the assumption that a verb's meaning influences its syntactic behavior. She first describes that verbs can express their arguments in alternate ways. Then, she presents the classes of verbs that share a kernel of meaning and discover in detail of the behavior for each class. Finally, she draws classes and their alternations, which become the verb inventory. At that time, the verb inventory of Levin has one drawback; her classification of verbs are based on syntactic properties unlike those in WordNet[2].

A method for mapping WordNet entries into Levin classes is proposed by Korhonen[3]. Words in WordNet are arranged in hierarchical, and each node contains a set of synonym called synset. 1,616 synsets were automatically mapped to one of 32 Levin classes, where the accuracy was 81%.

It is an open question how to define a set of common semantic classes for all words. It may depend on the applications requiring semantic class disambiguation. In this paper, WordNet is used for semantic class definition, however, any sets of semantic classes, including above verb classes, could be applicable for our method.

WSD with a coarse grained sense inventory has also been studied. Izquierdo et al. used Base Level Concepts (BLC) from WordNet in order to perform the class-based Word Sense Disambiguation[4]. He conducted the experiments under two different sets of BLC: all types of relations encoded in WordNet, and only the hyponymy relations. A naive most frequent classifier is able to perform a semantic tagging with accuracy figures over 75%.

Kohomban and Lee proposed a technique based on the similarity of word senses, which are coarser and more general concepts[5]. The general classes are mapped to fine grained senses with simple heuristics. Their proposed method trained a classifier for a word by using memory-based learner with 4 effective features: Local Context, Part-of-Speech, Collocation and Syntactic Relation. They reported that the accuracy was over 77%.

Semantic class disambiguation is not only well known in English but also another languages. Izquierdo et al. presented an approach of semantic disambiguation based on machine learning and semantic classes for Spanish[6]. They used semantic classes in order to collect a large number of examples for each class while the degree of polysemy is also reduced. Cast3LB, manually annotated corpus with Spanish WordNet senses, has been applied to Support Vector Machine with linear kernel in order to perform semantic disambiguation. The accuracy of disambiguation for nouns and verbs was 76.2%.

Resnik proposed an unsupervised WSD method based on selectional preferences [7]. Statistical model of selectinal restriction, which is an association score between a predicate and a conceptual class of a noun, is obtained from a corpus without sense tags and used for disambiguation of nouns. Although he evaluated his method for disambiguation of fine grained WordNet senses, his method could be used for coarse grained WSD using association scores for conceptual classes (i.e. coarse senses).

Past researches on coarse sense disambiguation tried to train classifiers for individual words. On the contrary, we aim to implement the universal model by training semantic class disambiguation classifiers that could be applicable to all words. We will further discuss the differences between previous work and our method in Subsection 3.2. Actually, our method is mainly based upon the Learning Semantic Class for Word Sense Disambiguation of Kohomban and Lee[5]. There are several points of differences, which could be categorized into 3 part of views: the learning algorithm, quantity of trained classifiers, and used corpus. We use Support Vector Machine while they used Memory Based Learner. We trained 32 classifiers (corresponding to number of noun and verb of WordNet unique beginners or semantic classes) while they trained one classifier for one word. Lastly, we use Senseval-3 and Yomiuri Shimbun newspaper articles 2003, while they use SemCor and Sensevel-2/3[13][14]. The reason of using the newspaper article corpus instead of SemCor is the larger-scaled data of Yomiuri Shimbun.

Chapter 3

Proposed Method

3.1 Semantic Class

WordNet, broadly cited as a sense repository, offers hierarchical structure of senses (meaning)[2]. WordNet develop synsets, which are organized into forty-five lexicographer files based on syntactic category and logical groupings. At this coarsest level of the senses in Wordnet, we defined as semantic classes, which are used in this research. This is a set of 45 semantic classes of:

- 26 semantic classes of noun,
- 15 semantic classes of verb,
- 3 semantic classes of adjective,
- and 1 semantic class of adverb.

Table 3.1 lists the mapping between IDs, name and contents of each semantic class. In this research, only nouns and verbs are disambiguated. Although the sense tagged corpus that we use for disambiguation contains adjectives, all of them are classified as one semantic class called 'adj.all'. For adverbs, it is only one class of adverb; therefore it is not necessary to be disambiguated.

3.2 Architecture

In this section, we describe about how do the classifiers determine semantic classes for given sentences. First, We show the abbreviation of important words, which will be frequently used in this paper.

Table 3.1: List of Semantic Classes in WordNet

ID	Name	Contents
03	noun.Tops	unique beginner for nouns
04	noun.act	nouns denoting acts or actions
05	noun.animal	nouns denoting animals
06	noun.artifact	nouns denoting man-made objects
07	noun.attribute	nouns denoting attributes of people and objects
08	noun.body	nouns denoting body parts
09	noun.cognition	nouns denoting cognitive processes and contents
10	noun.communication	nouns denoting communicative processes and contents
11	noun.event	nouns denoting natural events
12	noun.feeling	nouns denoting feelings and emotions
13	noun.food	nouns denoting foods and drinks
14	noun.group	nouns denoting groupings of people or objects
15	noun.location	nouns denoting spatial position
16	noun.motive	nouns denoting goals
17	noun.object	nouns denoting natural objects (not man-made)
18	noun.person	nouns denoting people
19	noun.phenomenon	nouns denoting natural phenomena
20	noun.plant	nouns denoting plants
21	noun.possession	nouns denoting possession and transfer of possession
22	noun.process	nouns denoting natural processes
23	noun.quantity	nouns denoting quantities and units of measure
24	noun.relation	nouns denoting relations between people or things or ideas
25	noun.shape	nouns denoting two and three dimensional shapes
26	noun.state	nouns denoting stable states of affairs
27	noun.substance	nouns denoting substances
28	noun.time	nouns denoting time and temporal relations
29	verb.body	verbs of grooming, dressing and bodily care
30	verb.change	verbs of size, temperature change, intensifying, etc.
31	verb.cognition	verbs of thinking, judging, analyzing, doubting
32	verb.communication	verbs of telling, asking, ordering, singing
33	verb.competition	verbs of fighting, athletic activities
34	verb.consumption	verbs of eating and drinking
35	verb.contact	verbs of touching, hitting, tying, digging
36	verb.creation	verbs of sewing, baking, painting, performing
37	verb.emotion	verbs of feeling
38	verb.motion	verbs of walking, flying, swimming
39	verb.perception	verbs of seeing, hearing, feeling
40	verb.possession	verbs of buying, selling, owning
41	verb.social	verbs of political and social activities and events
42	verb.stative	verbs of being, having, spatial relations
43	verb.weather	verbs of raining, snowing, thawing, thundering

1. Semantic Class(SC_i): a semantic class is a coarse grained meaning of a word. In general, one word contains one or more semantic classes.

2. Target word(Tw): A target word is an ambiguity word which potentially has 2 or more semantic classes.

3. Classifier (CL_i): A classifier is a system to judge whether a target word in a context has Semantic Class i or not.

As shown in Figure 3.1, in the most of previous work, WSD classifiers should be trained for individual target word w_i, since the sense inventories $\{\ldots, S_{ij}, \ldots\}$ are different. On the other hand, in our approach, we develop one system which can disambiguate all words in a text as shown in Figure 3.2. Note that our system choose semantic classes SC_i that are common for all words.

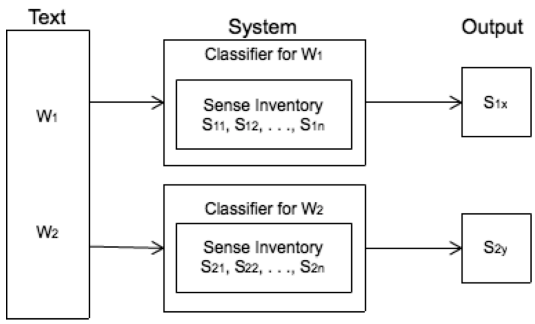

Figure 3.1: Previous Approach

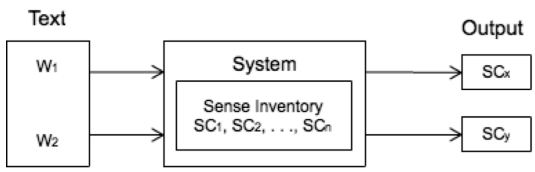

Figure 3.2: Our Approach

Our system choose semantic classes for a given target word as follows. Fig. 3.3 also illustrates our procedure.

i. Part-of-speech (POS) of the target word is identified by POS tagger. Only nouns and verbs could be disambiguated.

ii. By looking up WordNet, all possible candidates of semantic classes $\{\cdots, SC_k, \cdots\}$, which is a subset of all noun or verb semantic classes, for the target word are retrieved. For example, the target verb 'activate' has two semantic classes: verb.creation and verb.change.

iii. Each binary classifier CL_i judge if the target word has SC_i or not. The classifiers for individual semantic classes are trained in advance. For classification, features used for CL_i are extracted from a context of the target word. For example, 'activate' is applied to the classifiers namely verb.creation and verb.change which are correspond to semantic classes in ii.

iv. Finally the system outputs all SC_i where CL_i judges 'yes' as chosen semantic classes for the target word.

To illustrate the above procedure, let us consider how 'activate' in the context below is disambiguated:

Context: *Do you know what it is , and where I can get one ? We suspect you had seen the Terrex Autospade , which is made by Wolf Tools . It is quite a hefty spade , with bicycle - type handlebars and a sprung lever at the rear , which you step on to* ***activate*** *it . Used correctly , you should n't have to bend your back during general digging , although it wo n't lift out the soil and put in a barrow if you need to move it ! If gardening tends to give you backache , remember to take plenty of rest periods during the day , and never try to lift more than you can easily cope with .*

The system generates features as describe in Section 3.3.2 from the surrounding words. With the extracted features, CL_i of verb.creation and verb.change judge whether the word is SC_i or not. For Tw 'activate' in this context, the correct semantic class is verb.creation. So, CL_i of verb.creation is expected to judge as YES, while CL_i of verb.change is expected to judge as NO.

3.3 Classifier

In general, a classifier is a system that has ability to identify which category an instance belongs to. For this research, the classifier (CL_i) could judge whether the target word

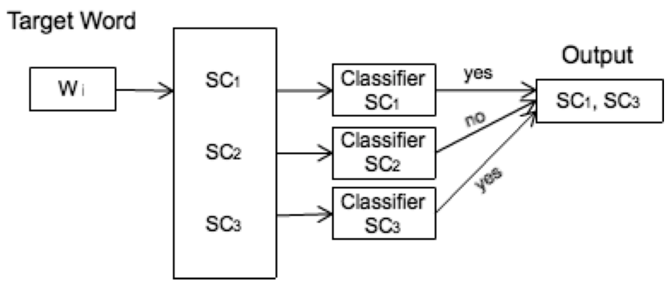

Figure 3.3: Architecture of Our System

has SC_i or not. In this section, we present the linear classifier and the features that we use to implement the classifier.

3.3.1 Learning Algorithm

In this research, Support Vector Machines (SVM)[15] is used as the classification algorithm. SVM is a kind of supervised learning, which can analyze data and recognize patterns. SVM is a binary classifier trained from a collection of positive and negative data. SVM is also applicable for multi-class classification by one-versus-rest method or pair-wise method. However, SVM is used as a binary classifier in this research. The binary SVM works as follows:

1. A set of training data by consisting of positive and negative samples is prepared.
2. The model is built by using SVM training algorithm. The model will separate the data to 2 side with the clearly gap.
3. The test data will be consulted with the model and placed positive or negative side.

For example, Figure 3.4 shows an example of training data and three separators. Black and white dots are positive and negative samples, respectively. The separator H_3(arrow line) does not separate black and white dots. It means that H_3 is a bad separator (bad model). The separator H_1 (bold line) can separate black and white dots with very small margin. So, H_1 could be used as a separator, but not so effective if compared to a H_2. The separator H_2 (dotted line) completely separates black and white dots with the maximum margin. Comparing these 3 separators, H_2 is the most suitable for this training data.

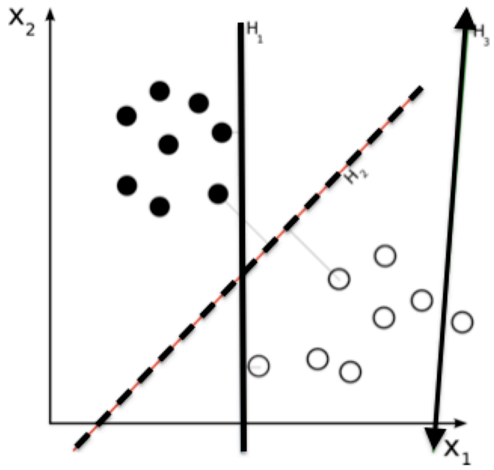

Figure 3.4: Outline of Support Vector Machine

In this paper, we use Liblinear as a supervised learning algorithm. We, first, tried to use Libsvm, but it is not a good option for evaluating such a large number of instances and features[8][9]. We change the learning algorithm from Libsvm to Liblinear. Without using kernels, liblinear can quickly train a much larger set via a linear classifier. Liblinear is an open source library for linear classification for data with millions of instances and features. It supports logistic regression and linear support vector machines. The main features of Libninear are same data formats as Libsvm, determination of weights for unbalance data and cross validation for model selection. We use L2-regularized L2-loss support vector classification with the default setting of Liblinear.

3.3.2 Features

The feature used in this research is fairly simple; we borrow machine learning features successfully used in WSD. Specifically, given an ambiguous target word, we use the following features from Kohomban and Lee[5] with some modifications.

Local Context

Local context is a feature represented as words in a context of a target word. It is also known as bag-of-words feature. We have tried several windows of n words to the left and n words to the write, where $n = \{3, 5, 10, 20\}$. The best value of n to our system is n=3. The punctuation marks and function words were removed. All words were converted into lower case. The window did not exceed the boundaries of a context; when there were not enough words to either side of the target word within the window, those remaining positions are ignored.

For instance, let us consider the noun 'immunisation' in the context below:

After ***immunisation*** *you must wait at least 1 month before becoming pregnant . Eat properly Eating well before and during pregnancy is very important . It keeps you fit and helps you to have a healthy baby . You do n't need a special diet and eating for two could mean you put on too much weight .*

The local context features are [after, you, must, wait] for window size $n = 3$.

Part-of-Speech

This feature consists of parts of speech (POS) of 2-gram, 3-gram and 4-gram including the target word itself. POSs of words in a sentence is determined by the public POS tagger developped by Tsujii laboratory in University of Tokyo[10]. POS feature can be represented as:

2-gram: $\{p_{-1}\ p_0\}$, $\{p_0\ p_1\}$

3-gram: $\{p_{-2}\ p_{-1}\ p_0\}$, $\{p_{-1}\ p_0\ p_1\}$, $\{p_0\ p_1\ p_2\}$

4-gram: $\{p_{-3}\ p_{-2}\ p_{-1}\ p_0\}$, $\{p_{-2}\ p_{-1}\ p_0\ p_1\}$, $\{p_{-1}\ p_0\ p_1\ p_2\}$, $\{p_0\ p_1\ p_2\ p_3\}$

Note: p_0 is POS of target word, p_1, p_2, p_3 is POS of 1,2,3 words after the target word. p_{-1}, p_{-2}, p_{-3} are POS of 1,2,3 words before the target word. When there was no word to either side of the target word, the value *"null"* was used to fill the vacancies.

For example, we used the POS tagger to extract POSs of the context that contains the noun "immunization":

After/IN ***immunisation/TN*** *you/PRP must/MD wait/VB at/IN least/JJS 1/CD month/TN before/IN becoming/VBG pregnant/JJ ./. Eat/NFP properly/RB Eating/VBG well/RB before/IN and/CC during/IN pregnancy/TN is/VBZ very/RB important/JJ ./. It/PRP keeps/VBZ you/PRP fit/TN and/CC helps/VBZ you/PRP to/TO have/VB a/DT healthy/JJ baby/TN ./. You/PRP do/VBP n't/RB need/VB a/DT special/JJ diet/TN and/CC eating/VBG for/IN two/CD could/MD mean/VB you/PRP put/VB on/RP too/RB much/JJ weight/TN ./.*

Symbols such as *IN, TN, PRP,* etc. after slashes are POS. The Part of Speech features are:

2-gram: {IN TN}, {TN PRP}

3-gram: {null IN TN}, {IN TN PRP}, {TN PRP MD}

4-gram: {null null IN TN}, {null IN TN PRP}, {IN TN PRP MD}, {TN PRP MD VB}

Collocation

A collocation feature is the connection between the words under consideration (target word) and its surrounding words, and it is used widely to solve the WSD task. No matter where the ambiguous word is appeared, a collocation has ability to determine the sense of the ambiguous word it contains. Aditi, showed that the ambiguous word "pound" represents examples of the benefitial of collocation: "pound of [something]" (unit of measure) shows that something always judges as noun.quantity[11].

In this paper, we consider 3 types of collocation, which are 2-gram, 3-gram and 4-gram including the target word itself. Therefore, the target word will replaced by '*'.In this paper, we consider 3 types of collocation, 2-gram, 3-gram and 4-gram including the target word itself. In our approach, the classifiers are applied to all words, i.e. they accept many kinds of words as the target word. Therefore, the target word is replaced by wild card symbol '*'.

2-gram: $\{w_{-1}$ $*\}$, $\{*$ $w_1\}$

3-gram: $\{w_{-2}$ w_{-1} $*\}$, $\{w_{-1}$ $*$ $w_1\}$, $\{*$ w_1 $w_2\}$

4-gram: $\{w_{-3}$ w_{-2} w_{-1} $*\}$, $\{w_{-2}$ w_{-1} $*$ $w_1\}$, $\{w_{-1}$ $*$ w_1 $w_2\}$, $\{*$ w_1 w_2 $w_3\}$

Note: '*' represents the target word, w_1, w_2, w_3 is 1,2,3 words after the target word. w_{-1}, w_{-2}, w_{-3} is 1,2,3 words before the target word. Similar to Part of Speech feature, if there is not enough word on either side of a context, we replace vacancies with "null".

For example, for the sentence we used in the Part-of-Speech example, the collocation features would be:

2-gram:	{after *}, {* you}
3-gram:	{null after *}, {after * you}, {* you must}
4-gram:	{null null after *}, {null after * you}, {after * you must}, {* you must wait}

Syntactic Relation Feature

Syntactic Relation feature represents direct grammatical relationships, such as subject-verb or noun-adjective, between the target word and its surrounding word. We use the typed dependency of the Stanford parser in order to extract the features[12]. We only analyze the sentence that contains the target word. There are 2 kinds of typed dependencies parsing: typed dependencies and collapsed typed dependencies.

Typed dependencies are representation in which each word in the sentence (except the head of the sentence) is the dependent of one other word.

Collapsed typed dependencies are obtained by collapsing a pair of typed dependencies into a single typed dependency, which is then labeled with a name based on the word between two dependencies.

Let us consider the case that the target word is 'immunization' in the same example sentence shown below:

> *"After **immunisation** you must wait at least 1 month before becoming pregnant ."*

Table 3.2 shows the result of Stanford Parser on both typed dependencies and collapsed typed dependencies. Accordeing to the table, *"prep(wait-5, After-1)"* and *"pobj(After-1, immunisation-2)"* in the typed dependencies are collapsed to *"rep_after(wait-5, immunisation-2)"* in collapsed typed dependencies

We will use the collapsed typed dependencies of Stanford parser as the syntactic relation features. The word indices in the output of Stanford Parser are removed and the target word is replaced as '*'. Therefore, the syntactic relation features of the sentence above are:

Table 3.2: Typed Dependencies and Collapsed Typed Dependencies that are extracted by Stanford Parser.

Stanford Parser	
Typed Dependencies	**Collapsed Typed Dependencies**
prep(wait-5, After-1)	prep_after(wait-5, immunisation-2)
pobj(After-1, immunisation-2)	nsubj(wait-5, you-3)
nsubj(wait-5, you-3)	aux(wait-5, must-4)
aux(wait-5, must-4)	root(ROOT-0, wait-5)
root(ROOT-0, wait-5)	quantmod(1-8, at-6)
quantmod(1-8, at-6)	mwe(at-6, least-7)
mwe(at-6, least-7)	pobj(at-6, least-7)
dobj(wait-5, 1-8)	dobj(wait-5, 1-8)
tmod(wait-5, month-9)	tmod(wait-5, month-9)
prep(wait-5, before-10)	prepc_before(wait-5, becoming-11)
pcomp(before-10, becoming-11)	acomp(becoming-11, pregnant-12)
acomp(becoming-11, pregnant-12)	

{*prep_after(wait, *), nsubj(wait, you), aux(wait, must), root(ROOT, wait), quantmod(1, at), mwe(at, least), dobj(wait, 1), tmod(wait, month), prepc_before(wait, becoming), acomp(becoming, pregnant)*}

3.4 Training Data

For our method, we use two kinds of training data a collection of monosemous words in a raw text and polysemous words in a sense tagged corpus.

3.4.1 Monosemous words

First, we use monosemous words, which have only one semantic class in WordNet, as training data. We extract all monosemous words from a raw text corpus.

In order to create training data, the specification of positive and negative samples are necessary. We used the target word that has a SC_i as positive samples. On the other hand, for the negative samples, we choose samples from other target words, which have semantic class other than SC_i. The number of positive training data and the number of negative training data will be changed for different semantic classes.

According to our All:All experiment in the Section 4.4, the performance of classifiers that were trained from monosemous words is not quite good. Outputs of the system tend to be judged as negative.

Random 1:1

We notice that the main cause of these bias judgments is the number of negative samples is far more than the positive samples. Therefore, we are more considering about the balance of number of positive and negative samples in training data.

We propose another method to construct the training data considering balance of the number of positive and negative data. In this method, all monosemous words that has a SC_i are used as positive samples. On the other hand, for the negative samples, monosemous words that has a semantic class other than SC_i are randomly chosen so that the ratio of the number of positive and negative samples would be 1:1.

At Most Method

Furthermore, we consider about the variety of target words in training data. We believe that the more target words are trained, the higher performance of system is. We propose a method called 'At Most Method', which has ability to gathering as much as difference target words for both positive and negative sample of training data. Since a target word contains one or more contexts, At Most Method will limit the maximum contexts per each target word for each classifier. Hereafter, T_m stand for the threshold of the number of context. It also means the maximum number of contexts of one target word in the training data. If the contexts of a target word is either equal to T_m or lower than T_m, all contexts are selected. On the other hand, if the number of contexts of a target word is larger than T_m, At Most Method will randomly select contexts up to the number of T_m. Table 3.3 shows how contexts are chosen in at most method ($T_m = 3$).

Table 3.3: Example of selecting context by using At Most Method

Target word	**Contexts**	**Selected Contexts**
T_1	C_1	$\{C_1\}$
T_2	C_2, C_3	$\{C_2, C_3\}$
T_3	C_4, C_5, C_6	$\{C_4, C_5, C_6\}$
T_4	C_7, C_8, C_9, C_{10}	$\{C_7, C_8, C_9\}$ or $\{C_7, C_9, C_{10}\}$ or $\{C_7, C_8, C_{10}\}$ or $\{C_8, C_9, C_{10}\}$

At Most Method could be applied to both positive and negative sample in the training data. Furthermore, it will control the ratio of positive and negative data as predefined ratio P:N.

For each semantic class, the most appropriate T_m of both positive and negative samples are determined by the following procedure.

i. f and $NTW(f)$, where f stands for frequency of target words, while $NTW(f)$ stands for number of types of target words whose frequency is equal to f. Table 3.4 shows an illustrative example of the first table.

Table 3.4: Example of Frequency Table

Positive Samples		Negative Samples	
Frequency(f)	Number of Words($NTW(f)$)	Frequency(f)	Number of Words($NTW(f)$)
1	100	1	150
2	10	2	50
4	5	3	10
10	1	20	1

ii. Second is table that is calculated from the infromation of the first table which contains T_m and $NC(T_m)$, where $NC(T_m)$ stands for the number of contexts chosen by At Most Method for a given threshold T_m. $NC(T_m)$ can be calculated from the statistics in the first table as Equation 3.1. It implies that all contexts are chosen for the target words such that $f <= T_m$, but only T_m contexts are chosen for target words whose frequency f is greater than T_m. An example of this table is shown in Table 3.5.

$$NC(T_m) = \sum_{f=1}^{T_m} f \times NTW(f) + \sum_{f=T_m+1}^{n} T_m \times NTW(f) \tag{3.1}$$

where n is the maximum frequency.

Table 3.5: Example of Calculation of T_m

Positive Samples		Negative Samples	
T_m	Number of Contexts($NC(T_m)$)	T_m	Number of Contexts($NC(T_m)$)
1	116	1	211
2	132	2	272
3	138	3	283
4	144	4	284
...	...	...	...
10	150	20	300

iii. These two tables are created for both positive and negative samples. Then, the we compare T_m of positive and negative side in table two, and select the best pair of T_m that closes to the predefined ratio P:N. Note that T_m of positive and negative could be different. For example, according to Table 3.4 and Table 3.5, given P:N = 1:2, T_m of positive samples is 10 and T_m of negative samples is 20.

There are some cases that even we pick all of the positive data (at most of positive data is maximum), the total number of positive data is less than the negative data chosen by At Most Method when $T_m = 1$. We solve this issue by randomly choosing the negative data until the number of positive data and negative data become P:N. For instance, let us suppose that P:N = 1:1, all positive data consists of 400 contexts while the negative data contains 6,400 contexts after at most one context is chosen for one target word. We will select only 400 contexts (for different target words) from 6,400 contexts for negative samples.

3.4.2 Polysemous words

The second data set that we use as a training data is polysemous words. The polysemous words, which are used as training data, are extracted from a sense tagged corpus. The correct semantic classes are used to distinguish the positive and negative samples. We use the technique called "cross-validation" in order to generate the training data and evaluate the performance of the system.

We separate the set of polysemous words into 5 parts with the equivalent number of words. Then, we used 5-fold cross validation to define the set of training data and test data. For these 5-fold cross validation, 4 parts are used as training data and 1 part as the test data. For the fairness of selecting which part is test data, we alter the set of test data for 5 times in order to make every part is the test data as shown in the Table 3.6.

Table 3.6: 5-fold cross validation of Polysemous words

Training	Training	Training	Training	Test
Training	Training	Training	Test	Training
Training	Training	Test	Training	Training
Training	Test	Training	Training	Training
Test	Training	Training	Training	Training

There are both advantages and disadvantages of using either monosemous words or polysemous words as training data. For monosemous words, we could create large number of training data because our system can extract all monosemous words from raw text corpus. However, words in training and test data are totally different, which is a possibility of causing negative impacts on semantic class disambiguation. On the other hand, the polysemous words are not able to build a large training data because sense or semantic class annotation is required. Nevertheless, a part of training data consists of sentences of the same target words in the test data, which lead to the higher performance of the classifiers.

Chapter 4

Evaluation

In this chapter, we report the experiments to evaluate the proposed method. In the following, we present how to prepare data, criteria of evaluation, definition the baseline of the system, various kinds of experiments and the results. Then, we will compare the results for various kinds of training data sets.

4.1 Preparing Data

We use a sense-tagged corpus called Senseval-3 English lexical sample task corpus as the tast data[13]. According the Senseval-3 English task corpus, only nouns, verbs, and adjectives are tagged with their senses. But, only nouns and verbs are ambiguous in the semantic classes level. So, we restricted the target words to nouns or verbs.

In this Senseval-3 task corpus, one word contains several instances (or example sentences) from different contexts. Each instance, there is a direct or indirect semantic class indicator which determines the semantic class of the target word. An instance contains one or more semantic classes due to its context. For nouns, there is a parameter namely "senseid" indecating the correct sense of the instance.

For example, Figure 4.1 shows an instance of the noun "argument":

From Figure 4.1 data, there are 2 sense IDs for this instance. One is "argument%1:09:00::" and the other is "argument%1:10:03::". The number after the first semi-colon of the senseid ('09' or '10') corresponds to an ID of semantic class in Table 3.1. As the consequence, the instance of noun 'argument' in the above context contains semantic classes of "noun.cognition" and "noun.communication".

For verbs, there is a parameter namely "senseid" but different format from the senseid of nouns. For instance, Figure 4.1 shows an instance of the verb call "appear":

Figure 4.1: Example of the Target Word Noun ‘argument’

<instance id=“argument.n.bnc.00069452” docsrc=“BNC”>
<answer instance=“argument.n.bnc.00069452” senseid=“argument%1:09:00::”/>
<answer instance=“argument.n.bnc.00069452” senseid=“argument%1:10:03::”/>
<context>
He believes that literary study has a cultural and humane rather than an intellectual value ; and the implication of his recent work is that there are other ways of attaining cultural value . This has some affinity with the Marxist position . In a published <head>argument</head>between Scholes and Hirsch , the former made the following statement , on the assumption that the conservative Hirsch would disagree with it : At the heart my belief is the conviction that no text is so trivial as to be outside the bounds of humanistic study . The meanest graffito , if fully understood , can be a treasure of human expressiveness
</context>
</instance>

Figure 4.2: Example of the Target Word Verb ‘appear’

<instance id=“appear.v.bnc.00067408” docsrc=“BNC”>
<answer instance=“appear.v.bnc.00067408” **senseid=“190903”**/>
<context>
A software product which runs exclusively on workstations is Signal Processing WorkSystem of SPW from Comdisco . This is a very comprehensive development tool allowing the engineer to make an entry at virtually any level in the design phase of a DSP based product . (A full review will be <head>appearing</head>in a later edition of EW + WW . References 1 . B.W . Kernighan and D M Ritchie .
</context>
</instance>

Figure 4.3: Correspondence between Two Formats of Sense IDs

<lexelt item="appear.v">
<sense id="190901" source="ws" **wn="appear%2:30:00::;appear%2:30:02::"** synset="appear arise emerge show" gloss="to come into view; become visible."/>
<sense id="190902" source="ws" **wn="appear%2:39:00::;appear%2:39:01::"** synset="appear look seem" gloss="to seem. "He appears smart, but I have doubts"."/>
<sense id="190903" source="ws" **wn="appear%2:30:01::;appear%2:36:00::"** synset="appear come_out" gloss="to come before the public, as a book or performer."/>
</lexelt>

The correct sense of this instance is "190903". Unfortunately, these senseid cannot directly map to the semantic class ID in Table 3.1. The mapping between "senseid in the format of verb" and "senseid in the format of noun" is provided in Senseval-3 English task corpus. In this case, the mapping of sense IDs of the verb 'appear' is written as in Figure 4.1:

According to the mapping dictionary above, the instance of verb 'appear' of the context from the former example with the senseid of "190903" can map to "appear%2:30:01::" and "appear%2:36:00::". From the sense IDs, we can obtain the semantic class ID in Table 3.1. Thus, the instance in this context contains 2 semantic classes: "verb.change" and "verb.creation".

For adjectives in Senseval-3 English task corpus, there are ambiguity in the fine grained meaning, while there is no ambiguity in the coarse grained (semantic classes) level. That is, all adjectives are mapped to the semantic class "adj.all".

Although there are 45 semantic classes in Wordnet, only 32 semantic classes appear in Senseval-3 English task corpus. These 32 semantic classes consist of 18 semantic classes of nouns and 14 semantic classes of verbs. The list of these semantic classes is shown in Figure 4.1.

Two experiments were conducted to evaluate our proposed method: one is 'monosemous words task' where the set of monosemous words are used as the training data, the other is 'polysemous words task' where the polysemous words are used.

We extract all monosemous words from Senserval-3 English task corpus [13] and Yomiuri Shimbun newspaper articles in 2003 [14]. We used the newspaper articles corpus in order to enlarge the number of training data for our system, because we believe that the larger training data is, the higher performance is. The polysemous words are gathered from the sense tagged corpus: Senseval-3 English task corpus.

Figure 4.4: Disambiguating Noun and Verb

Noun:
noun.act, noun.artifact, noun.attribute, noun.body, noun.cognition, noun.communication, noun.event, noun.group, noun.location, noun.object, noun.person, noun.possession, noun.process, noun.quantity, noun.relation, noun.shape, noun.state, and noun.substance

Verb:
verb.body, verb.change, verb.cognition, verb.communication, verb.competition, verb.consumption, verb.contact, verb.creation, verb.emotion, verb.motion, verb.perception, verb.possession, verb.social, and verb.stative

Table 4.1 shows number of target words (types), average number of instances (tokens) and average number of semantic classes per a target word in Senseval-3 data. It is used for the test data in monosemous words task, and both test and training data using 5-fold cross validation in polysemous words task.

Table 4.1: Statistics of Senseval-3 Corpus

	Words	Instances	Semantic Classes
Nouns	20	3593	3.90
Verbs	32	3953	4.18

Table 4.2 shows the average number of positive and negative samples in Senseval-3 and Yomiuri Shimbun corpus in monosemous words task. Note that the amount of the training data in monosemous words task is much greater than in polysemous words task.

Table 4.2: Training Data in Monosemous Words Task

	Senseval-3		Yomiuri	
	positive	negative	positive	negative
Nouns	163	4080	2370	59100
Verbs	67	4460	235	3290

4.2 Evaluation Criteria

Six kinds of evaluation criteria are used in the experiment. They are classified into two groups: Instance Based Evaluation and Judgment Base Evaluation. Instance Based Evaluation is capable of evaluating the outputs for instances or test sentences, while Judgment Based Evaluation is able to evaluate the judgment of clasifiers of semantic classes.

4.2.1 Instance Based Evaluation

Instance Based Evaluation is a measurement of the accuracy of instances. There are two sub-types of Instance Based Evaluation: Accuracy (Exact Match) and Accuracy (Partial Match).

Before describe about the formula, we would like to explain about three parameters that will be used for calculating the accuracies.

Exact Match (EM) = the judgment is EM when the set of semantic classes chosen by system is completely the same as correct semantic classes.

Partial Match (PM) = the judgment is PM when the semantic class(es) chosen by system contains at least one correct semantic class and it is not Exact Match.

Not Match (NM) = the judgment is NM when the semantic class chosen by the system DO NOT contains any correct semantic classes.

Table 4.3 shows the example of judgment of Exact Match, Partial Match and Not Match.

Table 4.3: Examples of Judgment on Target Instances

Target word	**Sentence**	**Correct Semantic Classes**	**Semantic Classes chosen by System**	**Judgment**
T_1	S_1	SC_1	SC_1	EM
T_1	S_2	SC_1, SC_2	SC_1, SC_2	EM
T_1	S_3	SC_1, SC_2	SC_1	PM
T_2	S_4	SC_1, SC_3	SC_1	PM
T_2	S_5	SC_3, SC_4	SC_3, SC_4, SC_5	PM
T_3	S_6	SC_5, SC_6	SC_3, SC_5	PM
T_4	S_7	SC_7	SC_8	NM
T_4	S_8	SC_7	-	NM

Then, two kinds of accuracies are defined as follows.

Accuracy (Exact Match)

The accuracy (exact match) is the ratio of Exact Match. The formula is:

$$Accuracy(ExactMatch) = \frac{N(EM)}{N(EM) + N(PM) + N(NM)} \tag{4.1}$$

OR

$$\frac{\text{Number of sentences of Exact Match}}{\text{Number of sentences}} \tag{4.2}$$

Where N() is the number of instances of EM, PM, or NM. For the test set of Table 4.3, accuracy (Exact Match) is $\frac{2}{8} = 25\%$.

Accuracy (Partial Match)

It is the ratio of Partial Match. The formula for calculating Accuracy (Partial Match) is:

$$Accuracy(PartialMatch) = \frac{N(EM) + N(PM)}{N(EM) + N(PM) + N(NM)} \tag{4.3}$$

OR

$$\frac{\text{Number of sentences of either Exact Match or Partial Match}}{\text{Number of sentences}} \tag{4.4}$$

For the test set of Table 4.3, accuracy (Partial Match) is $\frac{2+4}{8} = 75\%$.

In order to calculate the Instance Based Evaluation, there are two tables: the table of contexts and the table of judgment, which represent in Table 4.4 and Table 4.5, respectively. These two tables are linked by the index (primary key). From these tables, we can find a target word with its specific context for its correct semantic classes, the prediction semantic classes by system and the result of judgment written in EM/PM/NM.

Table 4.4: Tabel of Target Word and its Context

Index	Target Word	Sentence/Context
1	T1	S1
2	T1	S2
...	...	...

Table 4.5: Table of Judgment of Contexts

Index	Correct SC	Predicted SC	EM/PM/NM
1	SC1	SC1	EM
2	SC2, SC3	SC2	PM
...	...	...	...

4.2.2 Judgment Based Evaluation

Judgment Based Evaluation contains 4 types of measurements: Agreement Ratio, Precision, Recall and F-measure. To calculate these measures, a standard concept of true/false - positive/negative, as shown in Table 4.6, is used.

Table 4.6: Parameters of Judgment Based Evaluation

		Prediction (by System)	
		Positive	Negative
Reality (of the Instance)	Positive	TP	FN
	Negative	FP	TN

TP: true possitive (the system judges as positive, and it is correct)
TN: true negative (the system judges as negative, and it is incorrect)
FP: false possitive (the system judges as positive, and it is correct)
FN: false negative (the system judges as negative, and it is incorrect)

Agreement Ratio

This is a simple concept; how often did the Prediction and Reality in Table 4.6 are agreed, or what percent of the prediction are corrected. Agreement Ratio is calculated by the following formula:

$$AgreementRatio = \frac{TP + TN}{TP + TN + FN + FP} \tag{4.5}$$

OR

$$\frac{\text{Number of Correct Judgments}}{\text{Number of sentences}} \tag{4.6}$$

Precision

Precision shows percentage of correctness of positive prediction. In this case, Precision is calculated by the following formula:

$$Precision = \frac{TP}{TP + FP} \tag{4.7}$$

OR

$$\frac{\text{Number of instances that system judges as positive and it is correct}}{\text{Number of instances that system judges as positive}} \tag{4.8}$$

Recall

Recall is the percentage where the semantic classes in the test data are predicted by the system. In this paper, Recall is computed by the following formula:

$$Recall = \frac{TP}{TP + FN} \tag{4.9}$$

OR

$$\frac{\text{Number of instances that system judges as positive and it is correct}}{\text{Number of positive instances in test data}} \tag{4.10}$$

F-Measure

F-measure(F-1 or F-score) can be interpreted as a weighted average of the precision and recall, becoming the value from 0 to 1. It is defined as Equation (4.11).

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.11}$$

Example of calcualtion of Judgment Based Evaluation

For more clearly understanding of Agreement Ratio, Precision, Recall and F-measure, we show an example of computation of these measurements:

Table 4.7: Example of Judgment Based Evaluation

Tagged Test Data (Reality)	**System Judged (Prediction)**	Conclusion
1	1	TP
1	-1	FN
-1	1	FP
1	-1	FN
-1	-1	TN

From the Table 4.7, we can conclude that TP = 1, FN = 2, FP = 1 and TN = 1. Therefore, the Judgment Based Evaluation is:

Agreement Ratio	= (1+1)/(1+2+1+1)	= 40%
Precision	= 1/(1+1)	= 50%
Recall	= 1/(1+2)	= 33.3%
F-Measure	= (2*0.5*0.333)/(0.5+0.333)	= 40%

4.3 Baseline

We define the baseline for all experiments. The baseline of our system always selects "Most Frequent Semantic Class" which is calculated from number of semantic classes of training data. In the case of using monoseous words for the training data, we count the number of semantic classes from monosemous data of Senseval-3 English task corpus and Yomiuri Shimbun newspaper articles in 2003. For polysemous words task, we count the semantic classes from training data of each trial of cross validation. Table 4.8 shows the frequency of semantic classes of polysemous words from Senseval-3 and the monosemous words from these two corpora. The steps of selecting the "Most Frequent Semantic Class" is shown below:

1. Analyze Part-of-Speech(POS) of the target word by using POS-tagger[10].

2. Look up the target word with its POS in WordNet in order to get all candidates semantic classes of the target words (all possible semantic classes of the target word in any sentences).

3. The most frequent semantic class among the candidates in step 2 is chosen.

For instance, the word "argument" as a noun contains two candidates: "noun.cognition" and "noun.communication". Seeing Table 4.8, frequency of noun.cognition is 569, while frequency of noun.communication is 991 in monoseous words task. Therefore, the baseline of this word will always select "noun.communication".

4.4 Results

4.4.1 Monosemous Words Task

The monosemous words are the words that contain only one semantic class when look up that words with POS tagged in WordNet Dictionary.

We evaluate several training data in this task. On the other hand, the test data is the same for all experiments. We define the experiments corresponding to the ratio of P:N which is the ratio between number of positive training data and number of negative training data.

Before showing the results of experiments, we would like to present the meaning and method of naming each experiment. As we said in the previous paragraph, the names of experiments are corresponding to the ratio of P:N of training data. There are two kinds of P:N data. One is 'Random P:N' where negative samples are randomly chosen. The

Table 4.8: Frequency of Semantic Classes in Training Data

Semantic Class	Frequency					
	monosemous	5-fold crossvalidation of polysemous				
		1st	2nd	3rd	4th	5th
noun.act	953	137	137	134	148	149
noun.artifact	1910	382	382	384	390	400
noun.attribute	753	411	411	423	416	411
noun.body	231	167	167	163	161	162
noun.cognition	569	366	366	367	363	363
noun.communication	911	407	407	417	418	414
noun.event	158	62	62	61	67	67
noun.group	332	544	544	548	541	539
noun.location	242	43	43	41	44	40
noun.object	223	48	48	48	46	44
noun.person	1982	39	39	37	35	38
noun.possession	124	26	26	25	23	22
noun.process	57	33	33	33	27	29
noun.quantity	159	34	34	33	27	29
noun.relation	84	39	39	38	40	36
noun.shape	30	28	28	28	26	25
noun.state	431	204	204	205	202	199
noun.substance	538	55	55	53	53	51
verb.body	124	31	31	33	32	34
verb.change	224	539	539	526	544	521
verb.cognition	234	188	188	191	194	191
verb.communication	601	372	372	376	382	371
verb.competition	74	57	57	49	52	51
verb.consumption	43	258	258	252	255	255
verb.contact	238	59	59	62	63	61
verb.creation	104	324	324	326	327	323
verb.emotion	100	11	11	0	11	12
verb.motion	128	156	156	163	161	162
verb.perception	219	306	306	309	310	316
verb.possession	119	117	117	119	118	122
verb.social	197	161	161	169	156	166
verb.stative	81	210	210	203	201	208

other is 'At most P:N' where numbers of positive and negative samples are adjusted by At Most method in Subsection 3.4.1. 'All:All' means that all positive and negative samples are used.

All:All

For this experiment, we extract monosemous words from Senseval-3 English task corpus and Yomiuri Shimbun newspaper articles in 2003. The positive data is gathering from monosemous words that contain SC_i. Conversely, the negative data is collected from the monosemous words that contain other semantic classes rather than the SC_i. Table 4.9 and Table 4.10 show the results of the Instance Based Evaluation and Judgment Based Evaluation on this experiment.

Table 4.9: Results of Instance Based Evaluation of All : All

		Noun	Verb	All
System	Exact Match	3.1%	2.7%	2.9%
	Partial Match	3.9%	2.9%	3.4%
Baseline	Exact Match	24.2%	26.7%	25.4%
	Partial Match	30.0%	30.6%	30.3%

Table 4.10: Results of Judgment Based Evaluation of All : All

		Noun	Verb	All
System	Agreement Ratio	74.8%	74.1%	74.4%
	Precision	32.2%	24.2%	29.3%
	Recall	2.2%	2.3%	2.2%
	F-measure	3.6%	3.7%	3.6%
Baseline	Agreement Ratio	66.6%	65.8%	66.2%
	Precision	8.7%	14.1%	11.1%
	Recall	19.2%	21.4%	20.2%
	F-measure	9.5%	13.6%	11.3%

From Table 4.9 , both measurements of Instance Based Evaluation are about 10 times lower than the Baseline. For the results of Judgment Based Evaluation in Table 4.10 only Agreement Ratio and Precision are higher than the Baseline. On the contrary, other criteria of baseline are higher than our system.

According to our error analysis, almost all of the judgments by the system are negative. We analyze the number of positive and negative of the training data. The smallest ratio of number of positive to number of negative sample is 1:4. Furthermore, the largest ratio is 1:1564. These mean the number of negative samples is much more larger than the positive samples which lead to the bias to negative judgment and misclassification of the system.

Random 1:1

Due to the bad result from the All:All in the previous experiment, we assume that the cause of the low performance is the larger number of negative samples than positive samples. Hence, we randomly select the negative sample in order to make the number of both positive and negative sample of training data almost equal. Table 4.11 and Table 4.12 show the results of the Instance Based Evaluation and Judgment Based Evaluation.

Table 4.11: Results of Instance Based Evaluation of Random 1:1

		Noun	Verb	All
System	Exact Match	30.2%	25.3%	28.6%
	Partial Match	60.4%	42.5%	53.0%
Baseline	Exact Match	24.2%	26.7%	25.4%
	Partial Match	30.0%	30.6%	30.3%

Table 4.12: Results of Judgment Based Evaluation of Random 1:1

		Noun	Verb	All
System	Agreement Ratio	60.1%	55.2%	58.0%
	Precision	29.6%	25.7%	27.9%
	Recall	48.9%	41.0%	45.4%
	F-measure	34.4%	27.1%	31.2%
Baseline	Agreement Ratio	66.6%	65.8%	66.2%
	Precision	8.7%	14.1%	11.1%
	Recall	19.2%	21.4%	20.2%
	F-measure	9.5%	13.6%	11.3%

The closure quantity of negative samples and positive sample by using random selection method leads the scores higher than the Baseline in all evaluation criteria except for Agreement Ratio, which is a little bit lower. The system achieved better performance for nouns than verbs in terms of all criteria.

Comparing to All:All, the performance of Random 1:1 shows a great improvement of the measurements. For instance, Accuracy (Exact Match) in Random 1:1 is roughly 10 times better, and F-measure is about 9 times greater than All:All. On the other hand, Agreement Ratio is about 22% worse, and Precision is only 4.8% worse than All:All. In total, we can conclude that Random 1:1 is better than All:All.

These seem to be a good sign of improvement of the system. We can conclude that the unbalance of data could make the bias occurs in the judgment. All:All, which the negative data is overwhelmingly larger than the positive side, leads the judgment of the system tends to be negative.

At Most Method

Although the results of Random 1:1 is much greater than All:All, but its performance is still low, about 30% on both Exact Match and F-Measure. Naturally, a target word appears in one or more contexts. In the negative sample of training data, random selection of negative samples in Random 1:1 may not cover all target words or may contain contexts of the same target words. In other words, some words may appear in the training data many times, while some other words are not included by random sampling. We believe that such deviation may be a problem of low performance. Thus, we set up a new hypothesis that choosing various kinds of negative words in the training data can relieve the problem.

We adjust ratio of number of positive and negative data in various cases. Thus the "at most" value for each classifier is automatically adjusted in order to make the number of positive to negative become close to the set ratio. "At most method" (Section 3.4.1) can be applied on both positive and negative sample of the training data.

We set up several experiment with different ratios that using "At most method" as ratio Positive:Negative consist of 1:1, 1:2, 1:3, 2:1 and All:1.

First we try to use the "At Most Method" with the smallest ratio which is 1:1. Then, we increase the ratio of negative part to 1:2 and 1:3 as the consequence. Table 4.13 and Table 4.14 show the performance of system by using at most method with the ratio of 1:1, 1:2, and 1:3.

Table 4.13: Results of Instance Based Evaluation (At Most Method 1:1, 1:2 and 1:3)

		Noun	Verb	All
At Most Method 1:1	Exact Match	24.3%	14.2%	19.3%
	Partial Match	36.8%	20.6%	28.8%
At Most Method 1:2	Exact Match	15.7%	8.8%	12.3%
	Partial Match	20.3%	10.3%	15.4%
At Most Method 1:3	Exact Match	11.3%	6.1%	8.8%
	Partial Match	13.9%	7.0%	10.5%
Baseline	Exact Match	24.2%	26.7%	25.4%
	Partial Match	30.0%	30.6%	30.3%

In accordance with the results of experiments At Most Method 1:1, 1:2 and 1:3, we can conclude some issues about increasing the number of negative sample of training data. The score of Agreement Ratio and Precision are slightly improved. On the contrary, the score of Recall is significantly reduced (around 40%-50%). These makes the score of F-measure lower. Instance Based Evaluation, both Exact Match and Partial Match, also

Table 4.14: Results of Judgment Based Evaluation (At Most Method 1:1, 1:2 and 1:3)

		Noun	Verb	All
At Most Method 1:1	Agreement Ratio	68.7%	67.5%	68.2%
	Precision	31.9%	27.0%	29.8%
	Recall	27.5%	16.7%	22.8%
	F-measure	26.9%	17.7%	22.9%
At Most Method 1:2	Agreement Ratio	73.1%	71.7%	72.5%
	Precision	34.1%	26.3%	30.7%
	Recall	13.6%	8.9%	11.6%
	F-measure	17.1%	11.7%	14.8%
At Most Method 1:3	Agreement Ratio	74.0%	73.1%	73.6%
	Precision	35.1%	30.5%	33.1%
	Recall	8.7%	5.7%	7.4%
	F-measure	12.0%	8.5%	10.5%
Baseline	Agreement Ratio	66.6%	65.8%	66.2%
	Precision	8.7%	14.1%	11.1%
	Recall	19.2%	21.4%	20.2%
	F-measure	9.5%	13.6%	11.3%

behave in same way as All:All: the more negative training data is, the more misjudgment occur.

As the result, the more negative data is, the lower performance is. We set another experiment, which arrange the ratio of positive and negative data to 2:1. We assume that the results would be improved if we set the number of negative lower than positive.

Table 4.15: Results of Instance Based Evaluation (At Most Method 2:1)

		Noun	Verb	All
System	Exact Match	26.6%	14.2%	20.6%
	Partial Match	58.5%	20.6%	39.9%
Baseline	Exact Match	24.2%	26.7%	25.4%
	Partial Match	30.0%	30.6%	30.3%

Table 4.15 and Table 4.16 shows the performance of the system which seems to be improved. At Most Method 2:1 is the best experiments among various setting we conducted. But this is likely to bias the training data. The quantity of the negative sample in training data is very low, which means the system may judge the test data as yes because it does not know what patterns of the negative data are.

Table 4.16: Results of Judgment Based Evaluation (At Most Method 2:1)

		Noun	Verb	All
System	Agreement Ratio	54.9%	67.5%	60.4%
	Precision	26.7%	27.0%	26.8%
	Recall	52.6%	16.7%	36.9%
	F-measure	33.0%	17.7%	26.3%
Baseline	Agreement Ratio	66.6%	65.8%	66.2%
	Precision	8.7%	14.1%	11.1%
	Recall	19.2%	21.4%	20.2%
	F-measure	9.5%	13.6%	11.3%

We tried one more experiment, namely All:1, which it is related to the At Most technique. We used all positive training data, which is monosemous words that contain SC_i, while the negative data is selected monosemous words that do not contain SC_i by At Most Method in order to make the ratio of positive to negative data become 1:1. The results of the experiment are shown in the Table 4.17 and Table 4.18.

Table 4.17: Results of Instance Based Evaluation (At Most Method All:1)

		Noun	Verb	All
System	Exact Match	24.7%	19.2%	22.0%
	Partial Match	39.7%	28.8%	34.4%
Baseline	Exact Match	24.2%	26.7%	25.4%
	Partial Match	30.0%	30.6%	30.3%

Table 4.18: Results of Judgment Based Evaluation (At Most Method All:1)

		Noun	Verb	All
System	Agreement Ratio	67.2%	63.8%	65.7%
	Precision	31.4%	27.7%	29.8%
	Recall	32.2%	25.5%	29.2%
	F-measure	28.5%	22.7%	26.0%
Baseline	Agreement Ratio	66.6%	65.8%	66.2%
	Precision	8.7%	14.1%	11.1%
	Recall	19.2%	21.4%	20.2%
	F-measure	9.5%	13.6%	11.3%

Accordance to Evaluations of At Most Method 1:1, 1:2, 1:3, 2:1, and All:1, At Most Method 2:1 and All:1 represent the best models for these method. In order to select the best model for At Most Method, we refer Accuracy (Exact Match) and F-Measure because of their strictly and most common use criteria. For Accuracy (Exact Match), At

Most Method All:1 shows 1.4 point of percentage higher than At Most Method 2:1. On the contrary, F-Measure of At Most Method 2:1 represents 0.3 point of percentage higher score than the All:1.

Moreover, At Most 2:1 is likely to be more bias than All:1 because of the number of positive sample is twice of negative sample in training data, and the small number of negative sample may not lead to the misjudgment; the system will judge as positive because of lack of information. In contrast, At Most Method All:1 has the equivalence number of both positive and negative sample of training data. The point of bias judgment can be ignored in All:1 experiment.

Therefore, due to close performance of At Most Method 2:1 and All:1, and in the more tendency of judging as positive of At Most Method 2:1, we conclude the At Most All:1 to be the best model among all of At Most methods. For all experiments of At Most methods, the performance for nouns was better than verbs.

Summary

In this part, we sum up the 3 methods that are used to train the model by using the monosemous words. These three methods are All:All, Random 1:1 and At Most Method All:1 (the best model of the experiments of At Most Method). The Figure 4.5. displays the comparison of both Instance Based Evaluation and Judgment Based Evaluation on these three methods.

According to the bar graph of Figure 4.5, the most strict criteria of Instance Based Evaluations is Accuracy (Exact Match) and the most common use and effective measurement of Judgment Based Evaluation is F-Measure. Random 1:1 represents the highest performance on both measurements. Moreover, Random 1:1 also shows the roughly 40-50% and 30-50% better than the At Most Method All:1 and the baseline in the criteria of Accuracy (Partial Match) and Recall.

Although All:All shows the best performance in the criteria of Agreement Ratio, but these occurred because of the bias judgment. All:All system mostly judges the test data as negative, while 70% of test data is tagged as negative. We are just about to say that the All:All model always judges the test data as negative. These lead to the high score of Agreement Ratio of All:All.

For Precision, the score of All:All, Random 1:1 and At Most Method are almost the same. All of these scores are roughly 60% higher than the Baseline.

Owing to the analysis of these evaluations, for this current work, we can conclude the Random 1:1 is the best model for disambiguating the semantic classes by using Liblinear together with the monosemous words as training data.

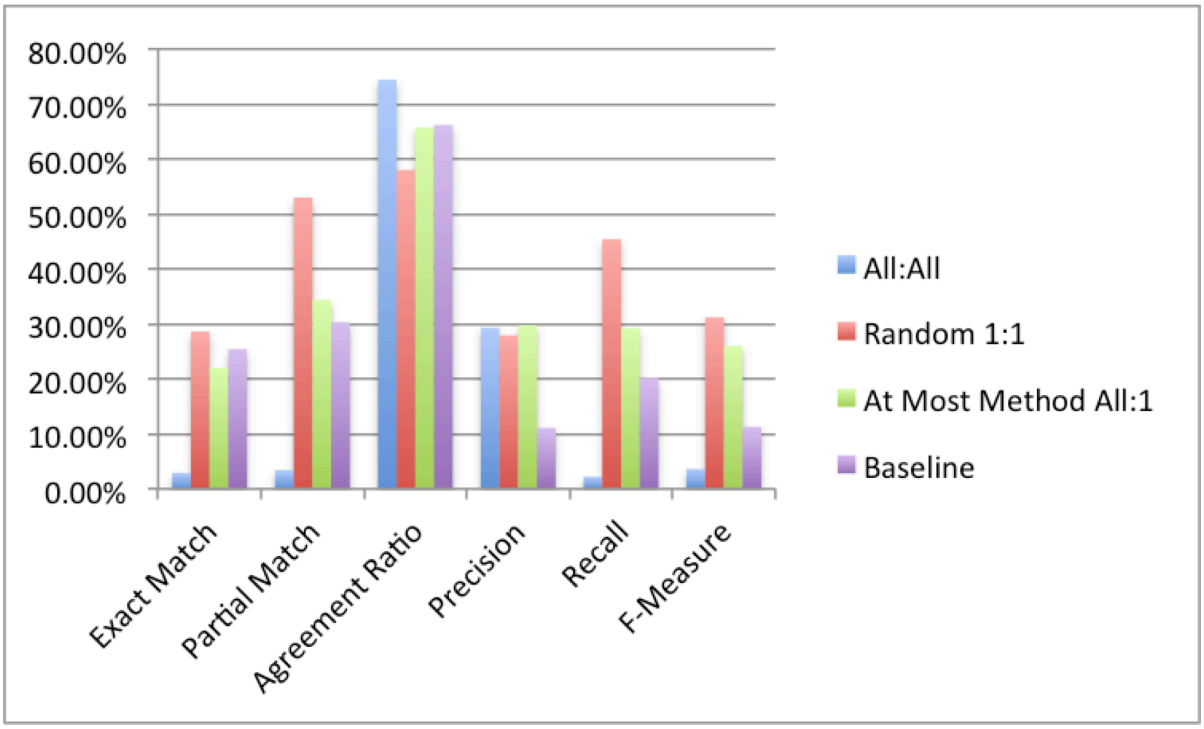

Figure 4.5: Comparision of Three Methods in Monosemous Words Task

Table 4.19 and Table 4.20 show examples of semantic classes predicted by the system of Random 1:1 monosemous words. For 'bank' in the context of index 3, the correct semantic class noun.act is predicted. For 'bank' in the contexts of index 8, no semantic class is chosen. It may be caused by data sparseness, that is, no effective feature in found from the contexts. For 'difficulty' in the context of index 10 and 11, many false positive errors are found. Finally, 'interest' in the context of index 26, the correct semantic class noun.act is not predicted. Further investigation is required to reveal the causes of errors and improve the performance of the proposed method.

4.4.2 Polysemous Words Task

In this section, we use the polysemous words from Senseval-3 English word corpus as the training data. We split the polysemous words from Senseval-3 English task corpus in to 5 parts and apply 5-fold cross validation technique in order to evaluate the performance of using polysemous words as training data for disambiguate semantic classes.

Table 4.19: Examples of Semantic Class Disambiguation (Contexts)

Index	Target Word	Context/Sentences
3	bank	Many types have a pot pitot in the nose and the ASI indicates a false reading during side − slip . On some , the rudder overbalances badly and there is a bad trim change . Most only need a small angle of <head>bank</head> , as the rudders on gliders are not very powerful . Normally you will want to use full side − slip for a few seconds rather than a small amount of slip for a much longer time . Provided that the need of slipping is spotted straight away , it should never be necessary to side − slip close to the ground .
8	bank	Then do the same on an easterly or westerly heading . Also try changing speeds . You will discover large errors flying near north or south with even small amounts of <head>bank</head> , and large errors on east and west if you vary the speed . Compass errors are an awful nuisance and it is well worth finding out a little about them . You also need to become accustomed to thinking and using degrees , and deciding whether you need to turn left or right to change the heading .
10	difficulty	Again , much interest attaches to interpretation , as an impassive hierarchical image of the Madonna is softened through the centuries into a more human and tender figure . The multitude of Madonnas for Italian worship in the Renaissance made this a fruitful theme for connoisseurship which has taken on the task of distinguishing authentic works from those by followers or copyists . An art critic may have <head>difficulty</head> in deciding how far the picture needs to be considered as devotional imagery , and how far discussion can be limited to artistic merits . An extremist view expressed by a twentieth − century artist is what Matisse had to say about his responses to murals by Giotto . When I see the Giotto frescoes at Padua I do not trouble to recognise which scene in the life of Christ I have before me , but I perceive instantly the sentiment which radiates from it and which is instinct in the composition in every line and color .
11	difficulty	In person , incidentally , Diderot was an encouraging critic . Sainte − Beuve tells us that David always spoke of Diderot with gratitude . It seems that David had at first great <head>difficulty</head> in making his way with the public , and was several times unsuccessful in his efforts after fame . It was at this time that Diderot , who often strolled into the artists ’ studios , paid a visit to David , and saw a picture which the artist was just finishing . He admired it and talked about the artist 's meaning , and the noble ideas he attributed to him.
26	interest	Our American friend 's cousin in London is an art student . Her college library has interesting books , as well as the latest art magazines . Her course teacher has given her a reading list , and the library staff are good at helping students with all sorts of <head>interests</head> . As part of the course , she has to choose a subject of her own about which to write a paper ; one of her difficulties is to know how to form her own views , not just copy already received opinions . She is looking for critical views against which to pitch her own ; it seems that she may have chosen the wrong sort of topic , since on a holiday in Italy she had been stunned by the newly renovated Michelangelo ceiling in the Sistine Chapel in Rome , and although there were plenty of books about it , many of them went into extravagant detail .

Table 4.20: Examples of Examples of Semantic Class Disambiguation (Results)

Index	**Correct SC**	**Predicted SC**	**EM/PM/NM**
3	noun.act	noun.act	EM
8	noun.act	null	NM
10	noun.act	noun.attribute, noun.cognition, noun.state	NM
11	noun.act	noun.act, noun.attribute, noun.cognition, noun.state	PM
26	noun.act, noun.cognition	noun.cognition	PM

In these 5 parts of data, four are used as the training data while the other one is used as the test data. We alter the part for using as test data 5 times, which means all of these 5 parts will be used as test data once. Absolutely, for each part of data, it contains more than one semantic class. We set the positive sample and negative sample for both training and test data as the former experiments. The target words that contain SC_i are placed to positive side, on the other hand, target words that do not contain SC_i are placed to the negative side. Anyway, for the polysemous words, one target word usually contains more than one semantic class. These lead to a target word could be positive sample for two or more semantic classes. Table 4.21 and Table 4.22 show the summarization of the results of Instance Based Evaluation and Judgment Based Evaluation of 5 trials of 5-fold cross validation.

Table 4.21: Results of Instance Based Evaluation (5 Trials of Cross-Validation)

		1st	2nd	3rd	4th	5th
System	Exact Match	42.1%	44.6%	44.2%	44.9%	40.3%
	Partial Match	48.8%	54.1%	50.8%	50.1%	49.2%
Baseline	Exact Match	37.6%	39.2%	40.2%	37.1%	38.0%
	Partial Match	42.3%	43.7%	41.1%	40.5%	42.5%

Table 4.22: Results of Judgment Based Evaluation (5 Trials of Cross-Validation)

		1st	2nd	3rd	4th	5th
System	Agreement Ratio	81.5%	83.7%	82.3%	83.0%	82.9%
	Precision	64.4%	56.4%	60.2%	66.1%	65.6%
	Recall	36.7%	36.4%	39.4%	37.4%	36.3%
	F-measure	42.5%	41.9%	44.3%	44.1%	43.8%
Baseline	Agreement Ratio	72.6%	74.5%	73.6%	73.0%	73.1%
	Precision	16.1%	15.6%	20.7%	15.6%	16.6%
	Recall	17.4%	17.4%	19.8%	16.9%	16.9%
	F-measure	14.1%	14.3%	17.1%	13.3%	13.9%

According to the five times of switching the partition of test data, the measurement of the Instance Based Evaluation: Accuracy (Exact Match) is around 40 $\sim$ 45% and the Accuracy (Partial Match) is about 50 $\sim$ 55%. These two evaluations are higher than the baseline roughly 1.8 times. In order to show the summary of results of 5-fold cross validation, we average these 5 results from the raw results. Table 4.23 and Table 4.24 show the evaluation on both Instance Based and Judgment Based Evaluation. The performance for nouns was better than verbs, however, differences were not so great as compared with monosemous words Random 1:1 and At Most Method.

Table 4.23: Results of Instance Based Evaluation (average)

		Noun	Verb	All
System	Exact Match	42.3%	45.3%	43.8%
	Partial Match	50.7%	49.5%	50.1%
Baseline	Exact Match	40.2%	36.6%	38.4%
	Partial Match	45.7%	39.4%	42.6%

Table 4.24: Results of Judgment Based Evaluation (average)

		Noun	Verb	All
System	Agreement Ratio	83.2%	82.3%	82.8%
	Precision	63.1%	61.6%	62.4%
	Recall	37.1%	36.4%	36.8%
	F-measure	43.1%	42.8%	43.0%
Baseline	Agreement Ratio	74.1%	72.7%	73.5%
	Precision	10.4%	24.1%	16.4%
	Recall	15.9%	19.2%	17.3%
	F-measure	12.1%	16.7%	14.1%

Using polysemous words as a training data shows a better performance than using the monosemous one. Although the number of positive training data that uses polysemous words is about a hundred or more, while the number of positive training that use monosemous words is roughly a thousand or more. But the performance of the polysemous words is better. If we have more polysemous words to train the system, the performance is expected to be improved. In another view, there must be some gaps between the characteristic of monosemous words and polysemous words, which lead to the difficulty of disambiguating the semantic classes of the target words.

Chapter 5

Conclusion and Future Plan

5.1 Conclusion

We have proposed the universal model for classifying semantic classes, which could be applicable to all words. We compare two kinds of experiments, which are differentiated by the source of training data.

The first model is training by using monosemous words from two corpora: Senseval-3 English Task Corpus and Yomiuri Shimbun newspaper articles in 2003. The monosemous word contains only one semantic class for each word. There are several experiments that we conducted under these training data. The best performance is obtained when the training data is prepared with all monosemous words that contain specific semantic classes and randomly selecting the monosemous words that contain another semantic classes, where the ratio of these positive and negative samples are set to one-to-one. F-Measure, which is the most well-known criteria, of this system shows roughly 2 times better than the baseline.

The other model is using polysemous words as the training data. Polysemous words are extracted from Senseval-3 English Task corpus. We conduct the famous method namely 5-fold cross validation on these polysemous words in order to separate the test and training data. After dividing the polysemous words into 5 parts, we choose 4 parts to use as the training data, and the remaining as the test data. The test data is altered 5 times in order to make all parts are used as the test data. The average of these 5 times of experiment are the result of 5-fold cross validation. The performance seems to be better than using monosemous words as training data. Agreement Ratio is over 80%, and Precision is about 68.6%. Moreover, F-Measure is roughly 3 times better than the beseline, which means about 1.4 times greater than the monosemous method.

We also show the results of Instance Based Evaluation and Judgment Based Evaluation for nouns and verbs. In our system, most of nouns show better performance than verbs. The main cause of these is the more samples of nouns than verbs in training data. However, nouns in baseline are slighly lower performance than verbs.

From these results, we can conclude as the follows:
1. The ratio of positive and negative sample in training data has an effect to the judgment of the system.

2. Even the number of training data that uses the monosemous words is much more greater than the polysemous words, the polysemous system shows better performance.

5.2 Future Plans

According to all experiments and observations of the characteristics of the data, we are planning to add another corpus to the system that uses monosemous words as training data in order to enlarge the number of positive samples. We plan to use the Yomiuri Shimbun newspaper articles in 2002. The reason that we cannot apply this corpus to the system immediately because of the time consuming of extracting monosemous words, converting the format and also generating the features.

We are planning to use another learning algorithm such as K-nearest Neighbors algorithm to classify the semantic classes. Then, we could compare the performance of Support Vector Machine to K-nearest Neighbors.

Bibliography

[1] Beth Levin: English Verb Classes and Alternations, University of Chicago Press, Chicago, IL, 1993.

[2] Christiane Fellbaum, editor: WordNet An Electronic Lexical Database, *The MIT Press*, Cambridge, MA ; London, May 1998.

[3] Anna Korhonen: Assigning Verbs to Semantic Classes via WordNet, *In Proceedings of the COLING Workshop on Building and Using Semantic Networks*, 2002.

[4] Ruben Izquierdo, Armando Suarez and German Rigau: A Proposal of Automatic Selection of Coarse-Grained Semantic Classes for WSD, *In Procesamiento del lenguaje natural. N. 39 (sept. 2007)*, pages 189–196, 2007.

[5] Upali S. Kohomban and Wee Sun Lee: Learning Semantic Classes for Word Sense Disambiguation, *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 34-41, 2005.

[6] Rubén Izquierdo-Beviá, Lorenza Moreno-Monteagudo, Borja Navarro and Armando Suárez: Spanish all-words semantic class disambiguation using Cast3LB corpus, *In Proceedings of the 5th Mexican international conference on Artificial Intelligence*, MICAI'06, pages 879–888, 2006.

[7] Philip Resnik: Selectional Preference and Sense Disambiguation. *In Proceeding of ACL Siglex Workshop on Tagging Text with Lexical Semantics*, 1997.

[8] Rong-En Fan and Kai-Wei Chang and Cho-Jui Hsieh and Xiang-Rui Wang and Chih-Jen Lin: LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research*, pages 1871–1874, 2008

[9] Chih-Chung Chang and Chih-Jen Lin: LIBSVM: A Library for Support Vector Machines, *ACM Transactions on Intelligent Systems and Technology*, pages 27:1–27:27, 2011.

[10] Yoshimasa Tsuruoka: An English Pos–Tagger, `http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/postagger/` [2008/07/28].

[11] Aditi Shrikumar: Two Approaches to All-Words Sense Disambiguation, *CS288 Final Project*, Berkley University, 2009.

[12] Marie catherine De Marneffe, Bill Maccartney, and Christopher D. Manning: Generating Typed Dependency Parses from Phrase Structure Parses, *In LREC 2006.*

[13] R. Mihalcea, T. Chklovski, and A. Kilgarriff: The Senseval-3 English lexical sample task, *In Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25-28, 2004.

[14] 'The Daily Yomiuri' newspaper articles in English, 2003.

[15] Cortes, Corinna; and Vapnik, Vladimir N.; Support-Vector Networks, Machine Learning, 20, 1995. http://www.springerlink.com/content/k238jx04hm87j80g/

Appendix A

Detail Results of Judgment Based Evaluation

This appendix shows the results of Judgment Based Evaluation (Agreement Ratio, Precision, Recall and F-Measure) of test data in Senseval-3 English Task corpus. The test data covers 32 semantic classes, which consist of 18 semantic classes of nouns and 14 semantic classes of verbs as explained in Section 4.1. We show results of each semantic class in several experiment setting.

A.1 Monosemous Words Task

A.1.1 All:All

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.79156627	1	0.02808989	0.05464481	0.51566265	0.30141844	0.95505618	0.45822102	178	652	830	5	5	564	170	7562	58180
noun.artifact	0.67898036	0.52083333	0.05165289	0.09398496	0.5206909	0.38469087	0.80991736	0.52162342	484	1015	1499	48	25	1019	392	8130	57612
noun.attribut	0.62257235	0.77777778	0.01340996	0.02636535	0.6151467	0.47787611	0.10344828	0.17007874	522	846	1368	9	7	113	54	2149	63593
noun.body	0.22077614	0	0	0	0.22074769	0	0	0	204	57	261	0	0	0	0	409	65333
noun.cognitio	0.70484309	0.68539326	0.13406593	0.22426471	0.68175	0	0	0	455	974	1429	89	61	0	0	3307	62435
noun.commu	0.67098958	0.68421053	0.02509653	0.04841713	0.46916941	0.33333333	0.59073359	0.42618384	518	1033	1551	19	13	918	306	6690	59052
noun.event	0.82593553	0	0	0	0.82548704	0	0	0	79	371	450	0	0	0	0	1590	64152
noun.group	0.58114121	0.71428571	0.02205882	0.04279601	0.57551593	0	0	0	680	920	1600	21	15	0	0	5020	60722
noun.locatio	0.76657936	0	0	0	0.76655331	0	0	0	51	165	216	0	0	0	0	1803	63939
noun.object	0.87740332	0.5	0.01694915	0.03278689	0.87740327	0	0	0	59	416	475	2	1	0	0	691	65051
noun.person	0.87824539	0.1	0.09090909	0.0952381	0.07273485	0.0713128	1	0.13313162	44	573	617	40	4	617	44	10752	54990
noun.posses	0.94651763	0	0	0	0.95152466	0	0	0	31	607	638	4	0	0	0	4910	60832
noun.process	0.91156687	0	0	0	0.91157882	0	0	0	38	381	419	0	0	0	0	135	65607
noun.quantit	0.94150089	0	0	0	0.94150091	0	0	0	38	596	634	0	0	0	0	1266	64476
noun.relatio	0.79177655	0	0	0	0.79177655	0	0	0	48	178	226	0	0	0	0	1862	63880
noun.shape	0.83650648	0	0	0	0.83650648	0	0	0	33	164	197	0	0	0	0	42	65700
noun.state	0.7002789	0	0	0	0.70263739	0	0	0	253	595	848	2	0	0	0	2568	63174
noun.substa	0.70905104	1	0.01515152	0.02985075	0.70453682	0	0	0	66	155	221	1	1	0	0	1934	63808
verb.body	0.93810914	0	0	0	0.93810221	0	0	0	41	610	651	0	0	0	0	126	4405
verb.change	0.6702879	0.66666667	0.00601504	0.0119225	0.47357475	0.30866602	0.47669173	0.37470449	665	1343	2008	6	4	1027	317	559	3972
verb.cognitio	0.77518116	0.703125	0.18907563	0.29801325	0.74412082	0.47540984	0.12184874	0.19397993	238	702	940	64	45	61	29	426	4105
verb.commu	0.67197935	0.3	0.08297872	0.13	0.29625181	0.29578351	1	0.4565323	470	1119	1589	130	39	1589	470	631	3900
verb.compet	0.84424208	0	0	0	0.84575364	0	0	0	64	349	413	1	0	0	0	229	4302
verb.consum	0.47757675	0	0	0	0.47757923	0	0	0	319	290	609	0	0	0	0	45	4486
verb.contact	0.85015392	0	0	0	0.50888409	0.2345679	1	0.38	76	428	504	0	0	324	76	256	4275
verb.creatio	0.53794102	0	0	0	0.53755277	0	0	0	407	472	879	0	0	0	0	236	4295
verb.emotio	0.94979851	0	0	0	0.953044	0	0	0	15	293	308	1	0	0	0	185	4346
verb.motion	0.58653403	0	0	0	0.56775166	0.46153846	0.3030303	0.36585366	198	281	479	1	0	130	60	311	4220
verb.percept	0.5753101	0.25	0.00514139	0.01007557	0.42377462	0.11290323	0.05398458	0.07304348	389	535	924	8	2	186	21	202	4329
verb.possess	0.8245571	1	0.00666667	0.01324503	0.82319595	0	0	0	150	696	846	1	1	0	0	174	4357
verb.social	0.82671591	0.13888889	0.02475248	0.04201681	0.78247383	0.08823529	0.04455446	0.05921053	202	1109	1311	36	5	102	9	943	3588
verb.stative	0.83894368	0.33333333	0.00387597	0.00766284	0.83953991	0	0	0	258	1345	1603	3	1	0	0	206	4325
Average:	0.7444863	0.29295358	0.02237155	0.03629015	0.66226633	0.11080424	0.20185204	0.11289259									

Using all positive data and negative data (All : All)

A.1.2 Random 1:1

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	# Positive Output	# Positive and Correct	# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-Measure							
noun.act	0.60361446	0.26332288	0.47191011	0.33802817	0.51566265	0.30141844	0.95505618	0.45822102	178	652	830	319	84	7562	5842
noun.artifact	0.503002	0.35085714	0.63429752	0.4518028	0.5206909	0.38469087	0.80991736	0.52162342	484	1015	1499	875	307	8130	6159
noun.attribut	0.5380117	0.43341404	0.68582375	0.53115727	0.6151467	0.47787611	0.10344828	0.17007874	522	846	1368	826	358	2149	1846
noun.body	0.7164751	0.86931818	0.75	0.80526316	0.22074769	0	0	0	204	57	261	176	153	409	389
noun.cognitic	0.46745976	0.32494279	0.62417582	0.42738901	0.68175	0	0	0	455	974	1429	874	284	3307	2765
noun.commu	0.53255964	0.37751479	0.61583012	0.46808511	0.46916941	0.33333333	0.59073359	0.42618384	518	1033	1551	845	319	6690	5206
noun.event	0.72	0.22352941	0.24050633	0.23170732	0.82548704	0	0	0	79	371	450	85	19	1590	1385
noun.group	0.59875	0.5264624	0.55588235	0.54077253	0.57551593	0	0	0	680	920	1600	718	378	5020	4027
noun.locatior	0.63425926	0.26666667	0.31372549	0.28828829	0.76655331	0	0	0	51	165	216	60	16	1803	1597
noun.object	0.54947368	0.13953488	0.50847458	0.2189781	0.87740327	0	0	0	59	416	475	215	30	691	634
noun.person	0.5834684	0.06882591	0.38636364	0.11683849	0.07273485	0.0713128	1	0.13313162	44	573	617	247	17	10752	7836
noun.posses	0.71316614	0.08695652	0.51612903	0.14883721	0.95152466	0	0	0	31	607	638	184	16	4910	3973
noun.proces	0.38663484	0.08045977	0.55263158	0.14046823	0.91157882	0	0	0	38	381	419	261	21	135	133
noun.quantit	0.77602524	0.04385965	0.13157895	0.06578947	0.94150091	0	0	0	38	596	634	114	5	1266	1117
noun.relation	0.63716814	0.05263158	0.04166667	0.04651163	0.79177655	0	0	0	48	178	226	38	2	1862	1623
noun.shape	0.45685279	0.1754386	0.60606061	0.27210884	0.83650648	0	0	0	33	164	197	114	20	42	42
noun.state	0.57665094	0.34036145	0.44664032	0.38632479	0.70263739	0	0	0	253	595	848	332	113	2568	2217
noun.substar	0.82352941	0.70149254	0.71212121	0.70676692	0.70453682	0	0	0	66	155	221	67	47	1934	1677
verb.body	0.3655914	0.06542056	0.68292683	0.11940299	0.93810221	0	0	0	41	610	651	428	28	126	111
verb.change	0.58017928	0.33143939	0.26315789	0.29337804	0.47357475	0.30866602	0.47669173	0.37470449	665	1343	2008	528	175	559	465
verb.cognitio	0.38829787	0.25544267	0.7394958	0.37971953	0.74412082	0.47540984	0.12184874	0.19397993	238	702	940	689	176	426	371
verb.commu	0.3971051	0.30967239	0.84468085	0.45319635	0.29625181	0.29578351	1	0.4565323	470	1119	1589	1282	397	631	522
verb.compet	0.80387409	0.05263158	0.015625	0.02409639	0.84575364	0	0	0	64	349	413	19	1	229	204
verb.consum	0.51067323	0.53710247	0.47648903	0.50498339	0.47757923	0	0	0	319	290	609	283	152	45	45
verb.contact	0.49007937	0.15057915	0.51315789	0.23283582	0.50888409	0.2345679	1	0.38	76	428	504	259	39	256	221
verb.creatior	0.56200228	0.57746479	0.2014742	0.29872495	0.53755277	0	0	0	407	472	879	142	82	236	208
verb.emotior	0.73051948	0.06410256	0.33333333	0.10752688	0.953044	0	0	0	15	293	308	78	5	185	167
verb.motion	0.5302714	0.392	0.24747475	0.30340557	0.56775166	0.46153846	0.3030303	0.36585366	198	281	479	125	49	311	272
verb.percept	0.40367965	0.37801205	0.64524422	0.47673314	0.42377462	0.11290323	0.05398458	0.07304348	389	535	924	664	251	202	176
verb.possess	0.67021277	0.18536585	0.25333333	0.21408451	0.82319595	0	0	0	150	696	846	205	38	174	158
verb.social	0.66056445	0.14985591	0.25742574	0.18943534	0.78247383	0.08823529	0.04455446	0.05921053	202	1109	1311	347	52	943	741
verb.stative	0.63568309	0.1487069	0.26744186	0.19113573	0.83953991	0	0	0	258	1345	1603	464	69	206	188
Average:	0.57955734	0.2788558	0.45422121	0.3116805	0.66226633	0.11080424	0.20185204	0.11289259							

Randomly select data around 1:1 (Random 1:1)

A.1.3 At Most Method 1:1

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.71927711	0.32484076	0.28651685	0.30447761	0.51566265	0.30141844	0.95505618	0.45822102	178	652	830	157	51	564	170	5768	5789
noun.artifact	0.58820499	0.37683824	0.42355372	0.39883268	0.5206909	0.38469087	0.80991736	0.52162342	484	1015	1499	544	205	1019	392	7958	6461
noun.attribut	0.5844943	0.44139651	0.33908046	0.38353196	0.6151467	0.47787611	0.10344828	0.17007874	522	846	1368	401	177	113	54	2149	2651
noun.body	0.66890611	0.87741935	0.66666667	0.75766017	0.22074769	0	0	0	204	57	261	155	136	0	0	409	553
noun.cognitic	0.58829175	0.37818182	0.45714286	0.41393035	0.68175	0	0	0	455	974	1429	550	208	0	0	3307	3687
noun.commu	0.63545344	0.44	0.33976834	0.38344227	0.46916941	0.33333333	0.59073359	0.42618384	518	1033	1551	400	176	918	306	5867	5869
noun.event	0.78585656	0.2	0.07594937	0.11009174	0.82548704	0	0	0	79	371	450	30	6	0	0	1590	1696
noun.group	0.60299116	0.56077348	0.29852941	0.38963532	0.57551593	0	0	0	680	920	1600	362	203	0	0	5018	6196
noun.locatior	0.748162	0.4375	0.2745098	0.3373494	0.76655331	0	0	0	51	165	216	32	14	0	0	1803	1962
noun.object	0.75946981	0.1744186	0.25423729	0.20689655	0.87740327	0	0	0	59	416	475	86	15	0	0	691	835
noun.person	0.73704938	0.07801418	0.25	0.11891892	0.07273485	0.0713128	1	0.13313162	44	573	617	141	11	617	44	4987	5180
noun.posses	0.75350635	0.06802721	0.32258065	0.11235955	0.95152466	0	0	0	31	607	638	147	10	0	0	4910	4994
noun.process	0.69630908	0.08333333	0.23684211	0.12328767	0.91157882	0	0	0	38	381	419	108	9	0	0	135	173
noun.quantit	0.81971027	0.03614458	0.07894737	0.04958678	0.94150091	0	0	0	38	596	634	83	3	0	0	1266	1369
noun.relatior	0.64521996	0.07692308	0.0625	0.06896552	0.79177655	0	0	0	48	178	226	39	3	0	0	1862	1920
noun.shape	0.61748843	0.13559322	0.24242424	0.17391304	0.83650648	0	0	0	33	164	197	59	8	0	0	42	61
noun.state	0.67171874	0.35869565	0.13043478	0.19130435	0.70263739	0	0	0	253	595	848	92	33	0	0	2568	2854
noun.substar	0.74059601	0.7	0.21212121	0.3255814	0.70453682	0	0	0	66	155	221	20	14	0	0	1934	2284
verb.body	0.88132196	0.1372549	0.17073171	0.15217391	0.93810221	0	0	0	41	610	651	51	7	0	0	126	200
verb.change	0.63888512	0.37751004	0.14135338	0.20568928	0.47357475	0.30866602	0.47669173	0.37470449	665	1343	2008	249	94	1027	317	559	679
verb.cognitio	0.62089243	0.35308642	0.60084034	0.44479005	0.74412082	0.47540984	0.12184874	0.19397993	238	702	940	405	143	61	29	426	486
verb.commur	0.53343039	0.27850163	0.36382979	0.31549815	0.29625181	0.29578351	1	0.4565323	470	1119	1589	614	171	1589	470	631	827
verb.compet	0.82695746	0.16666667	0.03125	0.05263158	0.84575364	0	0	0	64	349	413	12	2	0	0	229	277
verb.consum	0.47754837	0.5	0.10031348	0.16710183	0.47757923	0	0	0	319	290	609	64	32	0	0	45	73
verb.contact	0.77872529	0.15384615	0.10526316	0.125	0.50888409	0.2345679	1	0.38	76	428	504	52	8	324	76	256	390
verb.creatior	0.54696101	0.58	0.07125307	0.12691466	0.53755277	0	0	0	407	472	879	50	29	0	0	236	298
verb.emotior	0.8589187	0	0	0	0.953044	0	0	0	15	293	308	29	0	0	0	185	235
verb.motion	0.56546747	0.40350877	0.11616162	0.18039216	0.56775166	0.46153846	0.3030303	0.36585366	198	281	479	57	23	130	60	311	419
verb.percept	0.50385873	0.38255034	0.29305913	0.33187773	0.42377462	0.11290323	0.05398458	0.07304348	389	535	924	298	114	186	21	202	273
verb.posses	0.76064286	0.23232323	0.15333333	0.18473896	0.82319595	0	0	0	150	696	846	99	23	0	0	174	245
verb.social	0.69775793	0.08510638	0.0990099	0.09153318	0.78247383	0.08823529	0.04455446	0.05921053	202	1109	1311	235	20	102	9	887	889
verb.stative	0.76150827	0.13017751	0.08527132	0.1030445	0.83953991	0	0	0	258	1345	1603	169	22	0	0	206	250
Average:	0.68173692	0.29776975	0.2276086	0.22909848	0.66226633	0.11080424	0.20185204	0.11289259									

At most method with the ratio of 1:1 (At most 1:1)

A.1.4 At Most Method 1:2

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.77590361	0.38235294	0.07303371	0.12264151	0.51566265	0.30141844	0.95505618	0.45822102	178	652	830	34	13	564	170	6242	12485
noun.artifact	0.64694857	0.385	0.15909091	0.2251462	0.5206909	0.38469087	0.80991736	0.52162342	484	1015	1499	200	77	1019	392	7646	15297
noun.attribut	0.61012204	0.47058824	0.18390805	0.26446281	0.6151467	0.47787611	0.10344828	0.17007874	522	846	1368	204	96	113	54	2149	4298
noun.body	0.51574759	0.98734177	0.38235294	0.55123675	0.22074769	0	0	0	204	57	261	79	78	0	0	409	818
noun.cognitic	0.65536441	0.43708609	0.29010989	0.34874505	0.68175	0	0	0	455	974	1429	302	132	0	0	3038	6076
noun.commu	0.67031294	0.51595745	0.18725869	0.27478754	0.46916941	0.33333333	0.59073359	0.42618384	518	1033	1551	188	97	918	306	6377	12752
noun.event	0.8126007	0	0	0	0.82548704	0	0	0	79	371	450	6	0	0	0	1590	3180
noun.group	0.58488288	0.52525253	0.22941176	0.31934493	0.57551593	0	0	0	680	920	1600	297	156	0	0	4687	9373
noun.locatior	0.77122631	0.55555556	0.09803922	0.16666667	0.76655331	0	0	0	51	165	216	9	5	0	0	1803	3606
noun.object	0.86688679	0.33333333	0.08474576	0.13513514	0.87740327	0	0	0	59	416	475	15	5	0	0	691	1382
noun.person	0.84743418	0.0952381	0.13636364	0.11214953	0.07273485	0.0713128	1	0.13313162	44	573	617	63	6	617	44	10683	21363
noun.posses	0.82891447	0.1010101	0.32258065	0.15384615	0.95152466	0	0	0	31	607	638	99	10	0	0	4758	9517
noun.proces	0.84923369	0.09375	0.07894737	0.08571429	0.91157882	0	0	0	38	381	419	32	3	0	0	135	270
noun.quantit	0.88146567	0.04761905	0.05263158	0.05	0.94150091	0	0	0	38	596	634	42	2	0	0	1266	2532
noun.relatior	0.72513923	0.05882353	0.02083333	0.03076923	0.79177655	0	0	0	48	178	226	17	1	0	0	1862	3724
noun.shape	0.71941695	0.07407407	0.06060606	0.06666667	0.83650648	0	0	0	33	164	197	27	2	0	0	42	84
noun.state	0.68716912	0.27586207	0.03162055	0.05673759	0.70263739	0	0	0	253	595	848	29	8	0	0	2568	5136
noun.substar	0.71804149	0.8	0.06060606	0.11267606	0.70453682	0	0	0	66	155	221	5	4	0	0	1934	3868
verb.body	0.91661758	0.0625	0.02439024	0.03508772	0.93810221	0	0	0	41	610	651	16	1	0	0	126	252
verb.change	0.65932102	0.35294118	0.03609023	0.06548431	0.47357475	0.30866602	0.47669173	0.37470449	665	1343	2008	68	24	1027	317	515	1029
verb.cognitio	0.71985034	0.43809524	0.38655462	0.41071429	0.74412082	0.47540984	0.12184874	0.19397993	238	702	940	210	92	61	29	426	916
verb.commur	0.62663301	0.30745342	0.2106383	0.25	0.29625181	0.29578351	1	0.4565323	470	1119	1589	322	99	1589	470	558	1116
verb.compet	0.84413228	0	0	0	0.84575364	0	0	0	64	349	413	1	0	0	0	229	458
verb.consum	0.48250268	0.55555556	0.04702194	0.0867052	0.47757923	0	0	0	319	290	609	27	15	0	0	45	90
verb.contact	0.81444941	0.23529412	0.10526316	0.14545455	0.50888409	0.2345679	1	0.38	76	428	504	34	8	324	76	256	512
verb.creatior	0.53334977	0.3	0.00737101	0.01438849	0.53755277	0	0	0	407	472	879	10	3	0	0	236	472
verb.emotior	0.94004334	0.16666667	0.06666667	0.0952381	0.953044	0	0	0	15	293	308	6	1	0	0	185	370
verb.motion	0.57189988	0.33333333	0.04040404	0.07207207	0.56775166	0.46153846	0.3030303	0.36585366	198	281	479	24	8	130	60	311	622
verb.percept	0.52551071	0.34939759	0.14910026	0.20900901	0.42377462	0.11290323	0.05398458	0.07304348	389	535	924	166	58	186	21	202	404
verb.posses	0.80676774	0.30555556	0.07333333	0.11827957	0.82319595	0	0	0	150	696	846	36	11	0	0	174	348
verb.social	0.79695406	0.12643678	0.05445545	0.07612457	0.78247383	0.08823529	0.04455446	0.05921053	202	1109	1311	87	11	102	9	690	1380
verb.stative	0.80523827	0.14285714	0.04263566	0.06567164	0.83953991	0	0	0	258	1345	1603	77	11	0	0	206	412
Average:	0.72531502	0.30671661	0.11550203	0.14752986	0.66226633	0.11080424	0.20185204	0.11289259									

At most method with the ratio of 1:2 (At most 1:2)

A.1.5 At Most Method 1:3

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.77349398	0.30769231	0.04494382	0.07843137	0.51566265	0.30141844	0.95505618	0.45822102	178	652	830	26	8	564	170	7537	22611
noun.artifact	0.6549523	0.38028169	0.11157025	0.17252396	0.5206909	0.38469087	0.80991736	0.52162342	484	1015	1499	142	54	1019	392	8113	24339
noun.attribut	0.6210928	0.5112782	0.1302682	0.20763359	0.6151467	0.47787611	0.10344828	0.17007874	522	846	1368	133	68	113	54	2001	6001
noun.body	0.38935285	1	0.21568627	0.35483871	0.22074769	0	0	0	204	57	261	44	44	0	0	409	1227
noun.cognitio	0.66297365	0.43891403	0.21318681	0.28698225	0.68175	0	0	0	455	974	1429	221	97	0	0	3067	9201
noun.commu	0.67869953	0.58878505	0.12162162	0.2016	0.46916941	0.33333333	0.59073359	0.42618384	518	1033	1551	107	63	918	306	6218	18654
noun.event	0.82595267	0	0	0	0.82548704	0	0	0	79	371	450	0	0	0	0	1590	4770
noun.group	0.58926622	0.54910714	0.18088235	0.27212389	0.57551593	0	0	0	680	920	1600	224	123	0	0	4512	13536
noun.locatior	0.76661697	0.5	0.03921569	0.07272727	0.76655331	0	0	0	51	165	216	4	2	0	0	1803	5409
noun.object	0.86477183	0.2	0.03389831	0.05797101	0.87740327	0	0	0	59	416	475	10	2	0	0	691	2073
noun.person	0.86525895	0.1	0.11363636	0.10638298	0.07273485	0.0713128	1	0.13313162	44	573	617	50	5	617	44	9403	28209
noun.posses	0.85402078	0.09090909	0.22580645	0.12962963	0.95152466	0	0	0	31	607	638	77	7	0	0	4586	13757
noun.process	0.90179957	0.25	0.05263158	0.08695652	0.91157882	0	0	0	38	381	419	8	2	0	0	135	405
noun.quantit	0.91151703	0.04761905	0.02631579	0.03389831	0.94150091	0	0	0	38	596	634	21	1	0	0	1266	3798
noun.relation	0.76067043	0.11111111	0.02083333	0.03508772	0.79177655	0	0	0	48	178	226	9	1	0	0	1862	5586
noun.shape	0.79573944	0	0	0	0.83650648	0	0	0	33	164	197	8	0	0	0	42	126
noun.state	0.69197611	0.23529412	0.01581028	0.02962963	0.70263739	0	0	0	253	595	848	17	4	0	0	2040	6120
noun.substar	0.70901347	1	0.01515152	0.02985075	0.70453682	0	0	0	66	155	221	1	1	0	0	1934	6077
verb.body	0.92889249	0.125	0.02439024	0.04081633	0.93810221	0	0	0	41	610	651	8	1	0	0	126	378
verb.change	0.66231519	0.42391304	0.05864662	0.10303831	0.47357475	0.30866602	0.47669173	0.37470449	665	1343	2008	92	39	1027	317	287	861
verb.cognitio	0.75602374	0.52739726	0.32352941	0.40104167	0.74412082	0.47540984	0.12184874	0.19397993	238	702	940	146	77	61	29	305	916
verb.commur	0.67196729	0.27586207	0.06808511	0.10921502	0.29625181	0.29578351	1	0.4565323	470	1119	1589	116	32	1589	470	372	1116
verb.compet	0.84182074	0	0	0	0.84575364	0	0	0	64	349	413	2	0	0	0	229	687
verb.consum	0.48414092	0.75	0.01880878	0.03669725	0.47757923	0	0	0	319	290	609	8	6	0	0	45	135
verb.contact	0.8323098	0.23529412	0.05263158	0.08602151	0.50888409	0.2345679	1	0.38	76	428	504	17	4	324	76	256	768
verb.creatior	0.53678306	0.4	0.004914	0.00970874	0.53755277	0	0	0	407	472	879	5	2	0	0	236	708
verb.emotior	0.94005449	0	0	0	0.953044	0	0	0	15	293	308	4	0	0	0	185	555
verb.motion	0.584426	0.41666667	0.02525253	0.04761905	0.56775166	0.46153846	0.3030303	0.36585366	198	281	479	12	5	130	60	301	902
verb.percept	0.53310003	0.33333333	0.11053985	0.16602317	0.42377462	0.11290323	0.05398458	0.07304348	389	535	924	129	43	186	21	202	606
verb.possess	0.81623298	0.35	0.04666667	0.08235294	0.82319595	0	0	0	150	696	846	20	7	0	0	174	522
verb.social	0.84120231	0.37931034	0.05445545	0.0952381	0.78247383	0.08823529	0.04455446	0.05921053	202	1109	1311	29	11	102	9	720	2160
verb.stative	0.8077612	0.05263158	0.01162791	0.01904762	0.83953991	0	0	0	258	1345	1603	57	3	0	0	206	618
Average:	0.73606871	0.33063751	0.07378146	0.10478398	0.66226633	0.11080424	0.20185204	0.11289259									

At most method with the ratio of 1:3 (At most 1:3)

A.1.6 At Most Method 2:1

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.57108434	0.23511905	0.44382022	0.307393	0.51566265	0.30141844	0.95505618	0.45822102	178	652	830	336	79	564	170	7562	3781
noun.artifact	0.55675189	0.3605547	0.48347107	0.41306267	0.5206909	0.38469087	0.80991736	0.52162342	484	1015	1499	649	234	1019	392	8130	5253
noun.attribut	0.52233681	0.42067308	0.67049808	0.51698671	0.6151467	0.47787611	0.10344828	0.17007874	522	846	1368	832	350	113	54	2149	1075
noun.body	0.79127332	0.83111111	0.91666667	0.87179487	0.22074769	0	0	0	204	57	261	225	187	0	0	409	205
noun.cognitio	0.47921013	0.34065934	0.68131868	0.45421245	0.68175	0	0	0	455	974	1429	910	310	0	0	3307	1654
noun.commu	0.60121161	0.42383107	0.54247104	0.47586791	0.46916941	0.33333333	0.59073359	0.42618384	518	1033	1551	663	281	918	306	6690	3345
noun.event	0.65689158	0.17241379	0.25316456	0.20512821	0.82548704	0	0	0	79	371	450	116	20	0	0	1590	795
noun.group	0.53291056	0.45211268	0.47205882	0.4618705	0.57551593	0	0	0	680	920	1600	710	321	0	0	5020	2510
noun.locatior	0.65061533	0.33333333	0.49019608	0.3968254	0.76655331	0	0	0	51	165	216	75	25	0	0	1803	902
noun.object	0.47294866	0.12204724	0.52542373	0.19808307	0.87740327	0	0	0	59	416	475	254	31	0	0	691	346
noun.person	0.66040997	0.07216495	0.31818182	0.11764706	0.07273485	0.0713128	1	0.13313162	44	573	617	194	14	617	44	10360	5180
noun.posses	0.65620754	0.07623318	0.5483871	0.13385827	0.95152466	0	0	0	31	607	638	223	17	0	0	4910	2455
noun.process	0.30944202	0.08	0.63157895	0.14201183	0.91157882	0	0	0	38	381	419	300	24	0	0	135	68
noun.quantit	0.71026726	0.03205128	0.13157895	0.05154639	0.94150091	0	0	0	38	596	634	156	5	0	0	1266	633
noun.relatior	0.56951446	0.06896552	0.08333333	0.0754717	0.79177655	0	0	0	48	178	226	58	4	0	0	1862	931
noun.shape	0.31253561	0.16774194	0.78787879	0.27659574	0.83650648	0	0	0	33	164	197	155	26	0	0	42	21
noun.state	0.53220818	0.31538462	0.48616601	0.38258165	0.70263739	0	0	0	253	595	848	390	123	0	0	2568	1284
noun.substar	0.30105072	0.29864253	1	0.45993031	0.70453682	0	0	0	66	155	221	221	66	0	0	1934	967
verb.body	0.88046315	0.1372549	0.17073171	0.15217391	0.93800698	0	0	0	41	609	650	51	7	0	0	126	63
verb.change	0.63888469	0.37751004	0.14135338	0.20568928	0.4735747	0.30866602	0.47669173	0.37470449	665	1343	2008	249	94	1027	317	559	280
verb.cognitio	0.62089243	0.35308642	0.60084034	0.44479005	0.74412082	0.47540984	0.12184874	0.19397993	238	702	940	405	143	61	29	426	213
verb.commur	0.53343039	0.27850163	0.36382979	0.31549815	0.29625181	0.29578351	1	0.4565323	470	1119	1589	614	171	1589	470	631	316
verb.compet	0.82695746	0.16666667	0.03125	0.05263158	0.84575364	0	0	0	64	349	413	12	2	0	0	229	115
verb.consum	0.47754837	0.5	0.10031348	0.16710183	0.47757923	0	0	0	319	290	609	64	32	0	0	45	23
verb.contact	0.77872529	0.15384615	0.10526316	0.125	0.50888409	0.2345679	1	0.38	76	428	504	52	8	324	76	256	128
verb.creatior	0.54696101	0.58	0.07125307	0.12691466	0.53755277	0	0	0	407	472	879	50	29	0	0	236	118
verb.emotior	0.8589187	0	0	0	0.953044	0	0	0	15	293	308	29	0	0	0	185	93
verb.motion	0.56546747	0.40350877	0.11616162	0.18039216	0.56775166	0.46153846	0.3030303	0.36585366	198	281	479	57	23	130	60	311	156
verb.percept	0.50385873	0.38255034	0.29305913	0.33187773	0.42377462	0.11290323	0.05398458	0.07304348	389	535	924	298	114	186	21	202	101
verb.posses	0.76064286	0.23232323	0.15333333	0.18473896	0.82319595	0	0	0	150	696	846	99	23	0	0	174	87
verb.social	0.69775793	0.08510638	0.0990099	0.09153318	0.78247383	0.08823529	0.04455446	0.05921053	202	1109	1311	235	20	102	9	943	472
verb.stative	0.76150827	0.13017751	0.08527132	0.1030445	0.83953991	0	0	0	258	1345	1603	169	22	0	0	206	103
Average:	0.60434021	0.26823661	0.36868325	0.26319543	0.66226336	0.11080424	0.20185204	0.11289259									

At most method with the ratio of 2:1 (At most 2:1)

A.1.7 At Most Method All:1

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.70963855	0.29677419	0.25842697	0.27627628	0.51566265	0.30141844	0.95505618	0.45822102	178	652	830	155	46	564	170	7562	7562
noun.artifact	0.59353545	0.37695313	0.39876033	0.3875502	0.5206909	0.38469087	0.80991736	0.52162342	484	1015	1499	512	193	1019	392	8130	8130
noun.attribut	0.57572627	0.44444444	0.45210728	0.44824311	0.6151467	0.47787611	0.10344828	0.17007874	522	846	1368	531	236	113	54	2149	2149
noun.body	0.73400661	0.88068182	0.75980392	0.81578947	0.22074769	0	0	0	204	57	261	176	155	0	0	409	409
noun.cognitio	0.57154234	0.36472603	0.46813187	0.41000962	0.68175	0	0	0	455	974	1429	584	213	0	0	3307	3307
noun.commu	0.62899519	0.42602041	0.32239382	0.36703297	0.46916941	0.33333333	0.59073359	0.42618384	518	1033	1551	392	167	918	306	6690	6690
noun.event	0.77917554	0.18181818	0.07594937	0.10714286	0.82548704	0	0	0	79	371	450	33	6	0	0	1590	1590
noun.group	0.59298698	0.53301887	0.33235294	0.40942029	0.57551593	0	0	0	680	920	1600	424	226	0	0	5020	5020
noun.location	0.73885642	0.41666667	0.29411765	0.34482759	0.76655331	0	0	0	51	165	216	36	15	0	0	1803	1803
noun.object	0.67734496	0.16783217	0.40677966	0.23762376	0.87740327	0	0	0	59	416	475	143	24	0	0	691	691
noun.person	0.72881255	0.06944444	0.22727273	0.10638298	0.07273485	0.0713128	1	0.13313162	44	573	617	144	10	617	44	10752	10752
noun.posses	0.76133043	0.07042254	0.32258065	0.11560694	0.95152466	0	0	0	31	607	638	142	10	0	0	4910	4910
noun.process	0.6509817	0.09022556	0.31578947	0.14035088	0.91157882	0	0	0	38	381	419	133	12	0	0	135	135
noun.quantity	0.81017505	0.03370787	0.07894737	0.04724409	0.94150091	0	0	0	38	596	634	89	3	0	0	1266	1266
noun.relation	0.64517777	0.07692308	0.0625	0.06896552	0.79177655	0	0	0	48	178	226	39	3	0	0	1862	1862
noun.shape	0.50073694	0.19444444	0.63636364	0.29787234	0.83650648	0	0	0	33	164	197	108	21	0	0	42	42
noun.state	0.6621471	0.3559322	0.16600791	0.22641509	0.70263739	0	0	0	253	595	848	118	42	0	0	2568	2568
noun.substar	0.73602781	0.66666667	0.21212121	0.32183908	0.70453682	0	0	0	66	155	221	21	14	0	0	1934	1934
verb.body	0.65704459	0.06220096	0.31707317	0.104	0.93810221	0	0	0	41	610	651	209	13	0	0	126	126
verb.change	0.62831526	0.36601307	0.16842105	0.23069001	0.47357475	0.30866602	0.47669173	0.37470449	665	1343	2008	306	112	1027	317	559	559
verb.cognitio	0.59747693	0.34573304	0.66386555	0.45467626	0.74412082	0.47540984	0.12184874	0.19397993	238	702	940	457	158	61	29	426	426
verb.commur	0.46796569	0.26732673	0.45957447	0.33802817	0.29625181	0.29578351	1	0.4565323	470	1119	1589	808	216	1589	470	631	631
verb.compet	0.82437764	0.2	0.046875	0.07594937	0.84575364	0	0	0	64	349	413	15	3	0	0	229	229
verb.consum	0.50217468	0.54901961	0.26332288	0.3559322	0.47757923	0	0	0	319	290	609	153	84	0	0	45	45
verb.contact	0.69940908	0.1779661	0.27631579	0.21649485	0.50888409	0.2345679	1	0.38	76	428	504	118	21	324	76	256	256
verb.creation	0.55597202	0.59302326	0.12530713	0.20689655	0.53755277	0	0	0	407	472	879	86	51	0	0	236	236
verb.emotion	0.82648043	0.04651163	0.13333333	0.06896552	0.953044	0	0	0	15	293	308	43	2	0	0	185	185
verb.motion	0.56331207	0.42105263	0.16161616	0.23357664	0.56775166	0.46153846	0.3030303	0.36585366	198	281	479	76	32	130	60	311	311
verb.percept	0.46705986	0.38392857	0.44215938	0.41099164	0.42377462	0.11290323	0.05398458	0.07304348	389	535	924	448	172	186	21	202	202
verb.possess	0.71331804	0.21818182	0.24	0.22857143	0.82319595	0	0	0	150	696	846	165	36	0	0	174	174
verb.social	0.70611237	0.08558559	0.09405941	0.08962264	0.78247383	0.08823529	0.04455446	0.05921053	202	1109	1311	222	19	102	9	943	943
verb.stative	0.72533132	0.16483516	0.1744186	0.16949153	0.83953991	0	0	0	258	1345	1603	273	45	0	0	206	206
Average:	0.65723586	0.29775253	0.2923984	0.259765	0.66226633	0.11080424	0.20185204	0.11289259									

At most method with the ratio of all:1 (At most all:1)

A.2 Polysemous Words Task

A.2.1 1^{st} Cross Validation

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.81437126	0.6	0.26470588	0.36734694	0.79640719	0	0	0	34	133	167	15	9	0	0	144	519
noun.artifact	0.68585601	0.55714286	0.375	0.44827586	0.76170431	0.63716814	0.69230769	0.66359447	104	199	303	70	39	113	72	380	816
noun.attribut	0.74067584	0.62637363	0.6	0.61290323	0.51186074	0.39247312	0.76842105	0.51957295	95	180	275	91	57	186	73	427	666
noun.body	0.8441637	0.83333333	0.97560976	0.8988764	0.23607284	0	0	0	41	12	53	48	40	0	0	163	45
noun.cognitic	0.69390335	0.54385965	0.32978723	0.41059603	0.67443081	0	0	0	94	194	288	57	31	0	0	361	780
noun.commu	0.68692326	0.51724138	0.44117647	0.47619048	0.69004596	0.52040816	0.5	0.51	102	212	314	87	45	98	51	416	821
noun.event	0.81874359	0.71428571	0.25	0.37037037	0.78544496	0	0	0	20	70	90	7	5	0	0	59	301
noun.group	0.72699301	0.672	0.63636364	0.6536965	0.41110045	0.40866873	1	0.58021978	132	191	323	125	84	323	132	548	729
noun.locatior	0.69834075	1	0.06666667	0.125	0.6684341	0	0	0	15	29	44	1	1	0	0	36	136
noun.object	0.92394105	1	0.11111111	0.2	0.91321286	0	0	0	9	87	96	1	1	0	0	50	329
noun.person	0.87139153	0	0	0	0.8713057	0	0	0	17	108	125	0	0	0	0	27	465
noun.posses	0.98337025	0	0	0	0.98336958	0	0	0	3	125	128	0	0	0	0	28	482
noun.process	0.94098083	1	0.25	0.4	0.91745141	0	0	0	8	77	85	2	2	0	0	30	304
noun.quantit	0.94441717	1	0.11111111	0.2	0.93635789	0	0	0	9	118	127	1	1	0	0	29	478
noun.relatior	0.95432038	0.8	0.88888889	0.84210526	0.82080795	0	0	0	9	36	45	10	8	0	0	39	142
noun.shape	0.79885801	0.33333333	0.125	0.18181818	0.8205202	0	0	0	8	32	40	3	1	0	0	25	132
noun.state	0.71058152	0.51282051	0.39215686	0.44444444	0.70482659	0	0	0	51	119	170	39	20	0	0	202	476
noun.substar	0.83801292	0.71428571	0.71428571	0.71428571	0.7045517	0	0	0	14	31	45	14	10	0	0	52	124
verb.body	0.96059552	1	0.14285714	0.25	0.95194314	0	0	0	7	124	131	1	1	0	0	34	486
verb.change	0.69372987	0.5625	0.33333333	0.41860465	0.35296776	0.33919598	1	0.5065666	135	270	405	80	45	398	135	530	1073
verb.cognitio	0.86007303	0.87096774	0.54	0.66666667	0.73592004	0	0	0	50	138	188	31	27	0	0	188	564
verb.commu	0.85581273	0.86666667	0.57142857	0.68874172	0.60854975	0.38853503	0.67032967	0.49193548	91	229	320	60	52	157	61	379	890
verb.compet	0.79092455	0	0	0	0.80010427	0	0	0	17	65	82	1	0	0	0	47	284
verb.consum	0.7032043	0.6969697	0.73015873	0.71317829	0.49836151	0.6	0.04761905	0.08823529	63	59	122	66	46	5	3	256	231
verb.contact	0.82703204	0	0	0	0.83498362	0	0	0	17	83	100	1	0	0	0	59	345
verb.creatior	0.75901161	0.73684211	0.70886076	0.72258065	0.73619991	0.80769231	0.53164557	0.64122137	79	96	175	76	56	52	42	328	376
verb.emotior	0.96385503	1	0.25	0.4	0.94735806	0	0	0	4	58	62	1	1	0	0	11	235
verb.motion	0.60787479	0.69565217	0.33333333	0.45070423	0.51492122	0	0	0	48	49	97	23	16	0	0	150	232
verb.percept	0.77381352	0.73529412	0.67567568	0.70422535	0.60390667	0.54545455	0.08108108	0.14117647	74	109	183	68	50	11	6	315	426
verb.possess	0.87659194	0.625	0.38461538	0.47619048	0.85234829	0	0	0	26	146	172	16	10	0	0	124	550
verb.social	0.87357478	0.83333333	0.32608696	0.46875	0.76661202	0.13636364	0.06521739	0.08823529	46	216	262	18	15	22	3	156	893
verb.stative	0.85941924	0.55555556	0.20833333	0.3030303	0.8341639	0.38461538	0.20833333	0.27027027	48	273	321	18	10	26	10	210	1072
Average:	0.81504242	0.64385805	0.36676708	0.42526818	0.72644517	0.16126797	0.17390484	0.14065712									

1st cross-validation

A.2.2 2^{nd} Cross Validation

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.75449102	0.5	0.07317073	0.12765957	0.75449102	0	0	0	41	126	167	6	3	0	0	137	526
noun.artifact	0.70486452	0.60655738	0.3627451	0.45398773	0.7449983	0.60683761	0.69607843	0.64840183	102	197	299	61	37	117	71	382	818
noun.attribut	0.72619951	0.70588235	0.54054054	0.6122449	0.54816363	0.46276596	0.78378378	0.5819398	111	164	275	85	60	188	87	411	682
noun.body	0.93823018	0.90243902	1	0.94871795	0.3122295	0	0	0	37	16	53	41	37	0	0	167	41
noun.cognitic	0.75066046	0.62318841	0.48314607	0.5443038	0.68881484	0	0	0	89	196	285	69	43	0	0	366	778
noun.commu	0.7088731	0.62195122	0.45945946	0.52849741	0.7086736	0.60638298	0.51351351	0.55609756	111	199	310	82	51	94	57	407	834
noun.event	0.81898748	0.5	0.23529412	0.32	0.81898526	0	0	0	17	73	90	8	4	0	0	62	298
noun.group	0.74477946	0.6962963	0.69117647	0.69372694	0.42490368	0.42236025	1	0.59388646	136	186	322	135	94	322	136	544	734
noun.locatior	0.82725665	0.5	0.25	0.33333333	0.81964056	0	0	0	8	34	42	4	2	0	0	43	131
noun.object	0.91397112	1	0.18181818	0.30769231	0.89283832	0	0	0	11	84	95	2	2	0	0	48	332
noun.person	0.96650796	0	0	0	0.96633474	0	0	0	5	117	122	0	0	0	0	39	456
noun.posses	0.96897314	0	0	0	0.96897181	0	0	0	5	125	130	0	0	0	0	26	482
noun.proces	0.97523138	1	0.4	0.57142857	0.95084112	0	0	0	5	77	82	2	2	0	0	33	304
noun.quantit	0.976369	0	0	0	0.97617845	0	0	0	4	124	128	0	0	0	0	34	472
noun.relation	0.88836376	0.8	0.44444444	0.57142857	0.82169285	0	0	0	9	36	45	5	4	0	0	39	142
noun.shape	0.84329138	0.25	0.2	0.22222222	0.89286392	0	0	0	5	34	39	4	1	0	0	28	130
noun.state	0.73164704	0.51923077	0.55102041	0.53465347	0.72030735	0	0	0	49	123	172	52	27	0	0	204	472
noun.substar	0.92571925	0.81818182	0.81818182	0.81818182	0.76637062	0	0	0	11	33	44	11	9	0	0	55	122
verb.body	0.93019784	0.5	0.1	0.16666667	0.92897208	0	0	0	10	120	130	2	1	0	0	31	490
verb.change	0.71942977	0.57692308	0.35714286	0.44117647	0.33232996	0.31818182	1	0.48275862	126	277	403	78	45	396	126	539	1066
verb.cognitio	0.8541146	0.89285714	0.5	0.64102564	0.73439749	0	0	0	50	137	187	28	25	0	0	188	565
verb.commur	0.83966807	0.81944444	0.60204082	0.69411765	0.60418306	0.40993789	0.67346939	0.50965251	98	221	319	72	59	161	66	372	898
verb.competi	0.924874	0.5	0.14285714	0.22222222	0.92200223	0	0	0	7	75	82	2	1	0	0	57	274
verb.consum	0.78800711	0.74285714	0.85245902	0.79389313	0.49530083	0.33333333	0.03278689	0.05970149	61	62	123	70	52	6	2	258	228
verb.contact	0.83125497	0	0	0	0.83818922	0	0	0	17	85	102	1	0	0	0	59	343
verb.creatior	0.81154122	0.80246914	0.78313253	0.79268293	0.71498971	0.80769231	0.5060241	0.62222222	83	93	176	81	65	52	42	324	379
verb.emotior	0.94685902	0	0	0	0.94524983	0	0	0	4	56	60	0	0	0	0	11	237
verb.motion	0.65569645	0.64285714	0.42857143	0.51428571	0.57234635	0	0	0	42	54	96	28	18	0	0	156	227
verb.percept	0.78732809	0.83076923	0.65060241	0.72972973	0.54903971	0.46153846	0.07228916	0.125	83	102	185	65	54	13	6	306	433
verb.possess	0.83106184	0.75	0.18181818	0.29268293	0.80568287	0	0	0	33	134	167	8	6	0	0	117	562
verb.social	0.86026715	0.58823529	0.24390244	0.34482759	0.78122437	0.09090909	0.04878049	0.06349206	41	225	266	17	10	22	2	161	884
verb.stative	0.83957005	0.35294118	0.125	0.18461538	0.84855761	0.46153846	0.25	0.32432432	48	277	325	17	6	26	12	210	1068
Average:	0.83700896	0.56384628	0.36432888	0.41893764	0.74530515	0.15567119	0.17427268	0.14273365									

2nd cross-validation

A.2.3 3^{rd} Cross Validation

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.76687117	1	0.13636364	0.24	0.73006135	0	0	0	44	119	163	6	6	0	0	134	533
noun.artifact	0.71107414	0.64583333	0.31	0.41891892	0.75840699	0.62280702	0.71	0.6635514	100	195	295	48	31	114	71	384	820
noun.attribut	0.70633731	0.59223301	0.61616162	0.6039604	0.4842904	0.38918919	0.72727273	0.50704225	99	171	270	103	61	185	72	423	675
noun.body	0.89819879	0.87234043	1	0.93181818	0.22085174	0	0	0	41	11	52	47	41	0	0	163	46
noun.cognitic	0.70677928	0.53731343	0.40909091	0.46451613	0.68650304	0	0	0	88	192	280	67	36	0	0	367	782
noun.commu	0.74740778	0.64197531	0.51485149	0.57142857	0.68525001	0.5212766	0.48514851	0.5025641	101	205	306	81	52	94	49	417	828
noun.event	0.80615065	0.5	0.05555556	0.1	0.80545225	0	0	0	18	71	89	2	1	0	0	61	300
noun.group	0.73335305	0.71818182	0.59848485	0.65289256	0.4256585	0.42307692	1	0.59459459	132	180	312	110	79	312	132	548	740
noun.locatior	0.81212166	0.66666667	0.2	0.30769231	0.78240133	0	0	0	10	34	44	3	2	0	0	41	131
noun.object	0.93417151	1	0.36363636	0.53333333	0.89130214	0	0	0	11	83	94	4	4	0	0	48	333
noun.person	0.94945143	0	0	0	0.94909418	0	0	0	7	113	120	0	0	0	0	37	460
noun.posses	0.95959561	0	0	0	0.95959275	0	0	0	6	119	125	0	0	0	0	25	488
noun.process	0.94070112	0.4	0.4	0.4	0.9524658	0	0	0	5	80	85	5	2	0	0	33	301
noun.quantity	0.96803702	0	0	0	0.96812965	0	0	0	5	122	127	0	0	0	0	33	474
noun.relatior	0.97756602	0.9	0.9	0.9	0.80365499	0	0	0	10	36	46	10	9	0	0	38	142
noun.shape	0.81993759	0.2	0.2	0.2	0.89240141	0	0	0	5	34	39	5	1	0	0	28	130
noun.state	0.69535677	0.45238095	0.39583333	0.42222222	0.71959763	0	0	0	48	120	168	42	19	0	0	205	475
noun.substar	0.83398538	0.77777778	0.53846154	0.63636364	0.72089995	0	0	0	13	31	44	9	7	0	0	53	124
verb.body	0.9371845	0	0	0	0.94400692	0	0	0	8	122	130	1	0	0	0	33	488
verb.change	0.69043858	0.56842105	0.38848921	0.46153846	0.36372279	0.35012594	1	0.51865672	139	265	404	95	54	397	139	526	1078
verb.cognitio	0.88133212	0.85294118	0.61702128	0.71604938	0.7519347	0	0	0	47	141	188	34	29	0	0	191	561
verb.commur	0.82462916	0.7761194	0.55319149	0.64596273	0.6460998	0.43670886	0.73404255	0.54761905	94	226	320	67	52	158	69	376	893
verb.compet	0.84315035	0.66666667	0.13333333	0.22222222	0.82912024	0	0	0	15	69	84	3	2	0	0	49	280
verb.consum	0.69791179	0.73770492	0.67164179	0.703125	0.48641561	0.8	0.05970149	0.11111111	67	56	123	61	45	5	4	252	234
verb.contact	0.87939129	1	0.07142857	0.13333333	0.86751388	0	0	0	14	88	102	1	1	0	0	62	340
verb.creatior	0.74651343	0.68817204	0.79012346	0.73563218	0.74082873	0.81481481	0.54320988	0.65185185	81	97	178	93	64	54	44	326	375
verb.emotior	NO TEST DATA IN THIS CROSSVALIDATION																
verb.motion	0.74735951	0.70833333	0.48571429	0.57627119	0.64313363	0	0	0	35	61	96	24	17	0	0	163	220
verb.percept	0.79009283	0.83333333	0.625	0.71428571	0.59167451	0.61538462	0.1	0.17204301	80	107	187	60	50	13	8	309	428
verb.possess	0.86935349	0.83333333	0.32258065	0.46511628	0.8211275	0	0	0	31	139	170	12	10	0	0	119	557
verb.social	0.88586876	0.55	0.33333333	0.41509434	0.82508003	0.18181818	0.12121212	0.14545455	33	231	264	20	11	22	4	169	878
verb.stative	0.82627204	0.45454545	0.18181818	0.25974026	0.85704359	0.65384615	0.30909091	0.41975309	55	268	323	22	10	26	17	203	1077
Average:	0.82290961	0.60195142	0.39381995	0.44272342	0.73576702	0.20699572	0.19790275	0.17144042									

3rd cross-validation

A.2.4 4^{th} Cross Validation

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.81547619	0.46666667	0.23333333	0.31111111	0.82142857	0	0	0	30	138	168	15	7	0	0	148	514
noun.artifact	0.73271825	0.61818182	0.36170213	0.45637584	0.76940476	0.60169492	0.75531915	0.66981132	94	206	300	55	34	118	71	390	809
noun.attribut	0.72595099	0.66315789	0.59433962	0.62686567	0.50468415	0.30136986	0.20754717	0.24581006	106	165	271	95	63	73	22	416	681
noun.body	0.91780675	0.91304348	0.97674419	0.94382022	0.18278239	0	0	0	43	9	52	46	42	0	0	161	48
noun.cognitic	0.75495737	0.67213115	0.44565217	0.53594771	0.67896078	0	0	0	92	194	286	61	41	0	0	363	780
noun.commu	0.70147235	0.54666667	0.41	0.46857143	0.49087042	0.3592233	0.74	0.48366013	100	209	309	75	41	206	74	418	824
noun.event	0.84112747	0.2	0.08333333	0.11764706	0.87212078	0	0	0	12	78	90	5	1	0	0	67	293
noun.group	0.7205051	0.72072072	0.57553957	0.64	0.4384706	0.43573668	1	0.6069869	139	180	319	111	80	319	139	541	740
noun.locatior	0.92001232	0.71428571	0.71428571	0.71428571	0.8399627	0	0	0	7	34	41	7	5	0	0	44	131
noun.object	0.91311841	1	0.30769231	0.47058824	0.86924691	0	0	0	13	80	93	4	4	0	0	46	336
noun.person	0.94238308	1	0.11111111	0.2	0.93389632	0	0	0	9	114	123	1	1	0	0	35	459
noun.posses	0.94442821	0.5	0.125	0.2	0.94436139	0	0	0	8	119	127	2	1	0	0	23	488
noun.proces	0.85648129	0	0	0	0.88029002	0	0	0	11	73	84	2	0	0	0	27	308
noun.quantit	0.94239098	1	0.27272727	0.42857143	0.91838944	0	0	0	11	113	124	3	3	0	0	27	483
noun.relation	0.88760869	0.75	0.375	0.5	0.84263088	0	0	0	8	37	45	4	3	0	0	40	141
noun.shape	0.82219022	0.33333333	0.14285714	0.2	0.84606577	0	0	0	7	33	40	3	1	0	0	26	131
noun.state	0.72947419	0.57142857	0.39215686	0.46511628	0.69967704	0	0	0	51	116	167	35	20	0	0	202	479
noun.substar	0.80766219	0.83333333	0.38461538	0.52631579	0.71394598	0	0	0	13	30	43	6	5	0	0	53	125
verb.body	0.93698202	0	0	0	0.93626112	0	0	0	9	121	130	0	0	0	0	32	489
verb.change	0.71303803	0.54117647	0.38016529	0.44660194	0.32724939	0.3126615	1	0.47637795	121	273	394	85	46	387	121	544	1070
verb.cognitio	0.88207957	0.86206897	0.56818182	0.68493151	0.76892725	0	0	0	44	145	189	29	25	0	0	194	557
verb.commur	0.84898102	0.8030303	0.60227273	0.68831169	0.64028502	0.41830065	0.72727273	0.53112033	88	224	312	66	53	153	64	382	895
verb.competi	0.87915454	1	0.08333333	0.15384615	0.8647653	0	0	0	12	72	84	1	1	0	0	52	277
verb.consum	0.74899295	0.75384615	0.765625	0.75968992	0.49887304	0.8	0.0625	0.11594203	64	56	120	65	49	5	4	255	234
verb.contact	0.88860389	0.66666667	0.15384615	0.25	0.87622647	0	0	0	13	88	101	3	2	0	0	63	340
verb.creatior	0.77650631	0.73255814	0.7875	0.75903614	0.67357844	0.68965517	0.5	0.57971014	80	95	175	86	63	58	40	327	377
verb.emotior	0.96354928	1	0.25	0.4	0.9454685	0	0	0	4	57	61	1	1	0	0	11	236
verb.motion	0.72878697	0.61904762	0.7027027	0.65822785	0.62443196	0	0	0	37	59	96	42	26	0	0	161	222
verb.percept	0.71591732	0.6969697	0.58227848	0.63448276	0.5523067	0.33333333	0.05063291	0.08791209	79	105	184	66	46	12	4	310	430
verb.posses	0.82080424	0.53846154	0.21875	0.31111111	0.81391897	0	0	0	32	137	169	13	7	0	0	118	559
verb.social	0.85589459	0.78571429	0.23913043	0.36666667	0.76672062	0.17391304	0.08695652	0.11594203	46	212	258	14	11	23	4	156	897
verb.stative	0.83182245	0.63636364	0.12280702	0.20588235	0.83470481	0.57692308	0.26315789	0.36144578	57	259	316	11	7	26	15	201	1086
Average:	0.83021491	0.66058915	0.37383385	0.44137514	0.73034083	0.15633786	0.16854332	0.13358496									

4th cross-validation

A.2.5 5^{th} Cross Validation

	System				Baseline				# Positive Test Data	# Negative Test Data	# Test Data	System		Baseline		# Positive Train Data	# Negative Train Data
	Agreement Ratio	Precision	Recall	F-measure	Agreement Ratio	Precision	Recall	F-measure				# Positive	# Positive and Correct	# Positive	# Positive and Correct		
noun.act	0.8	0.25	0.06896552	0.10810811	0.82424242	0	0	0	29	136	165	8	2	0	0	149	516
noun.artifact	0.73112583	0.51785714	0.3452381	0.41428571	0.74445113	0.52542373	0.73809524	0.61386139	84	218	302	56	29	118	62	400	797
noun.attribut	0.71383078	0.66666667	0.55855856	0.60784314	0.52615325	0.37037037	0.27027027	0.3125	111	166	277	93	62	81	30	411	680
noun.body	0.95517315	0.93333333	1	0.96551724	0.18678732	0	0	0	42	9	51	45	42	0	0	162	48
noun.cognitic	0.74812129	0.63636364	0.45652174	0.53164557	0.68340271	0	0	0	92	198	290	66	42	0	0	363	776
noun.commu	0.71393629	0.6	0.40384615	0.48275862	0.47334424	0.35121951	0.69230769	0.46601942	104	208	312	70	42	205	72	414	825
noun.event	0.84301029	0.28571429	0.16666667	0.21052632	0.87333345	0	0	0	12	79	91	7	2	0	0	67	292
noun.group	0.73716978	0.73109244	0.61702128	0.66923077	0.43788066	0.43518519	1	0.60645161	141	183	324	119	87	324	141	539	737
noun.locatior	0.83860377	1	0.27272727	0.42857143	0.76528624	0	0	0	11	34	45	3	3	0	0	40	131
noun.object	0.90555262	1	0.33333333	0.5	0.85325037	0	0	0	15	82	97	5	5	0	0	44	334
noun.person	0.96776026	1	0.16666667	0.28571429	0.95947441	0	0	0	6	121	127	1	1	0	0	38	452
noun.posses	0.94506063	1	0.11111111	0.2	0.93718339	0	0	0	9	119	128	1	1	0	0	22	488
noun.process	0.93909712	1	0.33333333	0.5	0.90285763	0	0	0	9	74	83	3	3	0	0	29	307
noun.quantit	0.9604617	1	0.33333333	0.5	0.93674108	0	0	0	9	119	128	3	3	0	0	29	477
noun.relatior	0.88801026	0.875	0.58333333	0.7	0.7541498	0	0	0	12	33	45	8	7	0	0	36	145
noun.shape	0.86892334	1	0.25	0.4	0.81420897	0	0	0	8	31	39	2	2	0	0	25	133
noun.state	0.6951399	0.50909091	0.51851852	0.51376147	0.68897198	0	0	0	54	117	171	55	28	0	0	199	478
noun.substar	0.83766978	0.88888889	0.53333333	0.66666667	0.68197716	0	0	0	15	30	45	9	8	0	0	51	125
verb.body	0.95259746	0	0	0	0.95139982	0	0	0	7	123	130	0	0	0	0	34	487
verb.change	0.69142437	0.62025316	0.34027778	0.43946188	0.37798856	0.36455696	1	0.53432282	144	258	402	79	49	395	144	521	1085
verb.cognitio	0.86006077	0.83333333	0.53191489	0.64935065	0.75201058	0	0	0	47	141	188	30	25	0	0	191	561
verb.commur	0.82660396	0.7654321	0.62626263	0.68888889	0.65330821	0.46052632	0.70707071	0.55776892	99	219	318	81	62	152	70	371	900
verb.compet	0.83736548	0	0	0	0.84757171	0	0	0	13	68	81	1	0	0	0	51	281
verb.consum	0.68460633	0.71186441	0.65625	0.68292683	0.50287249	0.8	0.0625	0.11594203	64	57	121	59	42	5	4	255	233
verb.contact	0.85540006	0.5	0.06666667	0.11764706	0.85356437	0	0	0	15	84	99	2	1	0	0	61	344
verb.creatior	0.74774514	0.71910112	0.76190476	0.73988439	0.71916322	0.81481481	0.52380952	0.63768116	84	91	175	89	64	54	44	323	381
verb.emotior	0.96424992	0	0	0	0.96379624	0	0	0	3	60	63	0	0	0	0	12	233
verb.motion	0.72302394	0.69565217	0.44444444	0.54237288	0.62727443	0	0	0	36	58	94	23	16	0	0	162	223
verb.percept	0.74985418	0.73214286	0.56164384	0.63565891	0.61960689	0.58333333	0.09589041	0.16470588	73	112	185	56	41	12	7	316	423
verb.posses	0.84374913	0.55555556	0.17857143	0.27027027	0.83702147	0	0	0	28	140	168	9	5	0	0	122	556
verb.social	0.86530172	0.5	0.22222222	0.30769231	0.78864759	0.04545455	0.02777778	0.03448276	36	225	261	16	8	22	1	166	884
verb.stative	0.84234372	0.47368421	0.18	0.26086957	0.85782594	0.57692308	0.3	0.39473684	50	268	318	19	9	26	15	208	1077
Average:	0.8291554	0.65628207	0.36320834	0.43811416	0.73111712	0.166494	0.1693038	0.13870228									

5th cross-validation