The Role of Parts-of-Speech in Feature Selection

Stephanie Chua

***Abstract*—This research explores the role of parts-of-speech (POS) in feature selection in text categorization. We compare the use of different POS, namely nouns, verbs, adjectives and adverbs with a feature set that contains all POS. The best results are obtained with the use of only nouns. Therefore, we make use of a WordNet-based POS feature selection approach using the nouns feature set to compare with popular feature selection methods, namely Chi Square (Chi2) and Information Gain (IG). We find that the WordNet-based POS approach using only nouns as features can outperform Chi2 and IG in categorization effectiveness. Here, a machine learning approach to text categorization is employed and the Reuters-21578 top ten categories are used as the dataset.**

***Index Terms*—feature selection, machine learning, text categorization, WordNet**

I. Introduction

In recent years, the number of digital documents has escalated tremendously. Therefore, there is a need for an automated system to categorize those digital documents. Automated text categorization is defined as assigning new documents to pre-defined categories based on the classification patterns suggested by a training set of categorized documents [1]. One of the methods for automated text categorization is the machine learning approach. The application of machine learning in the field of text categorization only emerged in the 1990s. Many machine-learning schemes have been applied to text categorization and among them are Naïve Bayes [2], support vector machines (SVM) [3], decision trees [4] and so on. The concept behind machine learning in the task of categorization is generally described as a learner that automatically builds a classifier by learning from a set of documents that has already been classified by experts [1].

In automatically categorizing documents, it is crucial that a set of words that accurately represent the contents be used in training the classifier. In text categorization, one of the major processes is feature selection. Feature selection is performed in text categorization to tackle the problem of the large dimensionality of the feature space. This process involves selecting a subset of features from the feature space to represent the category. A feature space can contain thousands of features; however, it is not computationally efficient to process a large feature space. Therefore, a good subset of features needs to be selected to represent each category.

Works by [5] show the use of POS in feature selection. Their cascaded feature selection extracts two sets of documents, one with all POS and another with only nouns from the Reuters-21578 dataset. These terms are then looked up in WordNet [6] and the synonyms associated with those terms are used as features for representing documents. This work however, was not benchmarked with any other feature selection methods.

In this research, we focus on the feature selection process and we aim to explore the effects of different POS on text categorization effectiveness. In feature selection, we explore the use of nouns, verbs, adjectives and adverbs and all four POS combined to see whether the different POS do actually make a difference in categorization effectiveness. We then choose the best POS feature set and employed a WordNet-based feature selection approach. We benchmark this approach with two popular feature selection approaches, Chi Square (Chi2) and Information Gain (IG) [7] to judge the performance of the WordNet-based POS approach.

The following sections are organised as follows. In Section II, we give an overview of POS. Section III will generally describe WordNet while Section IV will discuss the WordNet-based POS feature selection approach. In Section V, the experiments are described and the results and analysis are presented in Section VI. Finally, Section VII concludes the paper.

II. Parts-Of-Speech (POS)

This section gives a brief definition of the four POS that are used in this research. According to [8], the definitions for the following POS are as follows:

1. Nouns
   A word or group of words used for referring to a person, thing, place or quality. For example, *"postman", "rope"* and *"Queensland"*.
2. Verbs
   A type of word or phrase that shows an action or a state. For example, *"run", "stand"* and *"remain"*.
3. Adjectives
   A word used for describing a noun or pronoun. For example, *"pretty", "big"* and *"tall"*.
4. Adverbs
   A word used for describing a verb, an adjective, another adverb or a whole sentence. For example, *"slowly", "quickly"* and *"cheerfully"*.

Manuscript received December 24, 2007. I would like to acknowledge the postgraduate fellowship funding provided by the Malaysian Ministry of Science, Technology and Environment (now called the Ministry of Science, Technology and Innovation).

Stephanie Chua is with the Faculty of Computer Science and Information Technology (FCSIT), Universiti Malaysia Sarawak (UNIMAS) in Malaysia. She is a lecturer with the Department of Information Systems, FCSIT, UNIMAS. (phone: +6082-583728; fax: +6082-583764; e-mail: chlstephanie@fit.unimas.my).

III. Introduction To WordNet

WordNet is an online thesaurus and an online dictionary. It can be considered as a dictionary based on psycholinguistics principles. WordNet contains nouns, verbs, adjectives and adverbs as parts-of-speech (POS). Function words are omitted based on the notion that they are stored separately as part of the syntactic component of language [6]. The information in WordNet is organized into sets of words called synsets. Each synset in WordNet has a unique signature that differentiates it from other synsets. Each of the synset contains a list of synonymous words and semantics pointers that illustrate the relationships between it and other synsets.

In this research, WordNet is chosen over other alternatives, as it is able to provide semantics information, consistently structured and electronically available.

IV. WordNet-Based POS Feature Selection

In the WordNet-based POS feature selection, five sets of features are obtained. The nouns are first identified based on the nouns in the WordNet's dictionary. Synonyms that co-occur in a category are cross-referenced with the help of WordNet's dictionary. Cross-referencing is the process of comparing the synset sense signatures of two synsets. If the synset sense signatures of the two synsets are the same, this means that the two terms are synonymous and exist in the same synset. The terms obtained from cross-referencing will be the features that will be used to represent a category. The same approach is used to obtain sets of features that consist of only verbs, adjectives and adverbs in WordNet that appear in each category. The four sets of features contain nouns, verbs, adjectives and adverbs respectively. The fifth set of features consists of features that include all four POS in WordNet that appear in each category. The approach is shown in Fig. 1.

V. Experiments

The dataset used in this research is the top ten categories for the Reuters-21578 dataset. Two sets of experiments were carried out. The first set of experiment was carried out to compare the performance of the different POS in the WordNet-based POS approach. The performances of the WordNet-based POS approach using nouns, verbs, adjectives and adverbs respectively were compared with using all POS (nouns, verbs, adjectives and adverbs). The second set of experiment compared the best WordNet-based POS approach with statistical feature selection approaches, Chi2 and IG.

A. *Machine Learning for Automated Text Categorization*

The algorithm used for categorization in this research is the multinomial Naïve Bayes classifier from the Waikato Environment for Knowledge Analysis (WEKA) [9]. WEKA is chosen because it has a readily implemented multinomial Naïve Bayes algorithm and also automated computation of effectiveness measures, such as precision, recall and F_1 measure. We chose the multinomial Naïve Bayes classifier over other types of classifiers, as it has been widely used in text categorization tasks and it is simple and straightforward in its implementation. The core equation for the multinomial Naïve Bayes classifier is derived from the Bayes theorem. It is based on the naïve assumption that words in a document occur independently of each other given the class.

The multinomial Naïve Bayes learning scheme will learn from the training document representation and induce a classifier. This classifier will then be tested on the testing document representation to evaluate its effectiveness in classification. Fig. 2 shows the machine learning approach to automated text categorization.

B. *Performance Measures*

There are various methods to judge the effectiveness of the text categorization classifier built using the machine learning approach. In binary categorization, the contingency table is used to evaluate the classifier for each category. From a contingency table, which is also known as a confusion matrix, performance measures that can be computed are in the form of precision (P), recall (R), accuracy (Acc), error (Err) and f-measure (F_1). These measures have been used in most of the previous researches. Prior to carrying out the experiments, the appropriate metrics were determined.

Precision is defined as the number of documents retrieved that are relevant over the total number of documents retrieved. Recall is defined as the number of documents retrieved that is relevant from the total number of documents that are relevant [9]. Accuracy is defined as the number of documents correctly retrieved over the total number of documents while error is the number of documents incorrectly retrieved over the total number of documents.

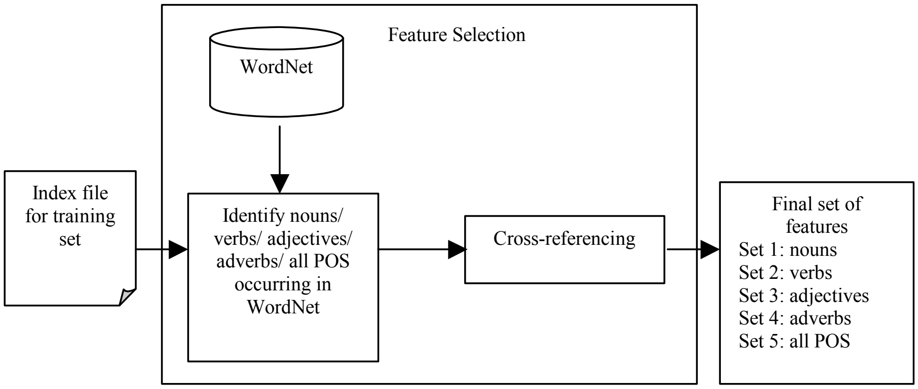

Fig 1. The WordNet-based POS feature selection approach

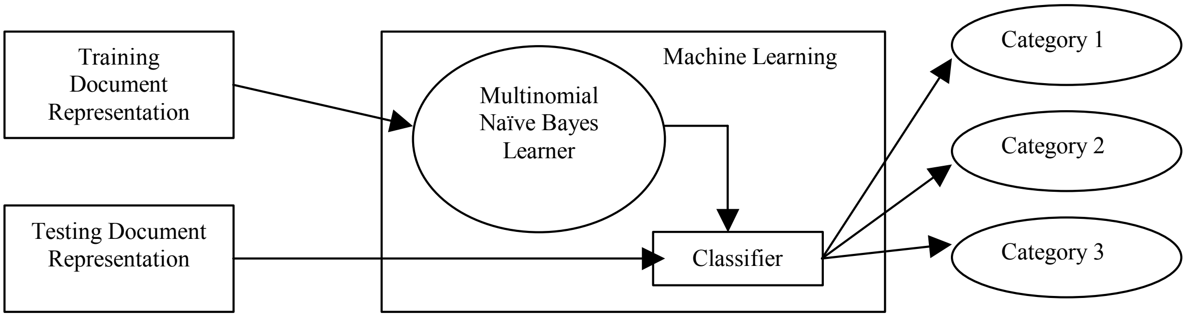
Training Document Representation
Testing Document Representation
Machine Learning
Multinomial Naïve Bayes Learner
Classifier
Category 1
Category 2
Category 3

Fig 2. The machine learning approach to automated text categorization

Both accuracy and error are not widely used as measures in text categorization [10]. The number of instances being tested for each category is always small as compared to the total number of instances in the test set. With a large denominator in accuracy and error, any variations in the correct instances will not have a significant impact on the value of accuracy and error. Therefore, it is not a good evaluator for a classifier.

TABLE I
CONTINGENCY TABLE FOR BINARY CATEGORIZATION (EXTRACTED FROM [9])

Category Set		Classifier Judgments	
		Yes	No
Expert Judgments	Yes	TP	FN
	No	FP	TN

From Table I, TP, FP, FN and TN represent the number of true positives, false positives, false negatives and true negatives respectively. True positives and true negatives are correctly classified instances. False positives are instances that are wrongly classified as positive when it is actually negative while false negatives are instances that are wrongly classified as negative when it is actually positive. These values can then be used to calculate the performance metrics based on the formula shown in Table II.

TABLE II
COMPUTATION FOR THE PERFORMANCE METRICS (EXTRACTED FROM [1])

Performance Metrics	Computation Formula
Precision	$P = TP / (TP + FP)$
Recall	$R = TP / (TP + FN)$
Accuracy	$Acc = (TP + TN) / (TP + FP + FN + TN)$
Error	$Err = (FP + FN) / (TP + FP + FN + TN)$
F-measure	$F_\beta = (\beta^2 + 1) \cdot P \cdot R / \beta^2 \cdot P + R$

In binary categorization, precision and recall are the common measures used in most researches [1]. However, neither precision nor recall makes an effective measure by itself. To have a high value of precision would mean that there is a trade-off of low recall and vice-versa. Often, the combination of both precision and recall is used to evaluate the effectiveness of a classifier.

In this research, the performance of the classifier was measured using the F_1 measure. The F_1 measure is the combination of both precision and recall to obtain a single value to measure the performance of a classifier. The formula for F_1 measure is shown in (1).

$$F_\beta = (\beta^2 + 1) \cdot P \cdot R / \beta^2 \cdot P + R \qquad (1)$$

Typically, the value 1 is used for β to give equal weight to both precision and recall, thus resulting in the F_1 measure formula shown in (2).

$$F_1 = 2 \cdot P \cdot R / P + R \qquad (2)$$

VI. RESULTS AND ANALYSIS

From the first set of experiments, we found that the use of nouns gives the best performance as compared to all other POS and is also slightly better than the use of all four POS. The results are shown graphically in Fig. 3. We report the micro-averaged F_1 measure for all the experiments.

Let us analyse at term level by looking at the top 20 terms in each of five feature sets for category "acquisition". This is shown in Table III.

From Table III, we can see that, when comparing between the nouns feature set and the all POS feature set, there is one term that differ. All the other 19 terms are the same. The nouns feature set has the term "investment" but the all POS feature set has the term "bank". Both terms are relevant to the category "acquisition". The performances of both these feature sets in the text categorization task are almost the same. The only advantage of using the nouns feature set is that it has a much smaller feature space as compared to using all POS. This is our aim in dimensionality reduction. When looking at the feature set for verbs, we find that a number of terms like "acquire", "buy", "sell", "bid" and "purchase" are relevant to the category "acquisition". However, the performance of this verbs feature set in the text categorization task is a little lower, as compared to the nouns and the all POS feature sets. Although most of the verbs mentioned are relevant to the category, it can also be used in many other categories. This is due to the nature of verbs, where it is used to describe an action. It decreases the ability of those terms to discriminate one category from another. The same case is observed in the use of the adjectives and adverbs feature sets in the text categorization tasks.

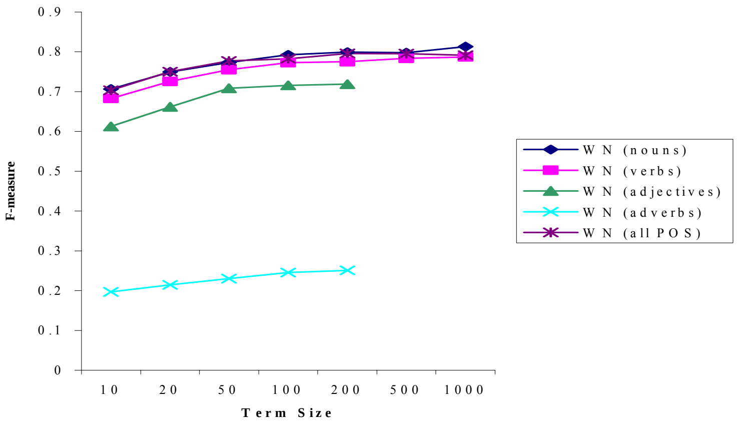

Fig. 3. Comparison of the performance of the WordNet-based POS feature selection approach using nouns, verbs, adjectives, adverbs and all POS

TABLE III
THE TOP 20 TERMS IN EACH FEATURE SET FOR THE CATEGORY "ACQUISITION"

Nouns	Verbs	Adjectives	Adverbs	All POS
title	said	stock	firm	title
company	title	common	earlier	company
shares	company	tender	currently	shares
pct	shares	outstanding	close	pct
corp	offer	international	wholly	corp
offer	share	subsidiary	late	offer
share	stock	financial	newly	share
stock	stake	firm	immediately	stock
stake	buy	total	soon	stake
acquisition	bank	based	fair	acquisition
merger	board	proposed	substantially	merger
common	sell	completed	approximately	common
unit	acquire	subject	overseas	unit
buy	bid	expected	short	buy
agreement	agreed	federal	originally	bank
board	tender	national	highly	agreement
shareholders	price	holding	shortly	board
sell	exchange	earlier	direct	shareholders
american	purchase	approved	near	sell
investment	terms	disclosed	course	american

It is even more obvious in the case of these two feature sets. Adjectives are commonly used to describe a noun or pronoun, while adverbs are used to describe a verb, an adjective, another adverb or a whole sentence. Again, these can be found throughout the dataset, across all categories. This explains the poor categorization results obtained when the adjectives and adverbs feature sets are used in the text categorization task.

The results of the second set of experiments show that the WordNet-based POS approach using the nouns feature set is able to perform better than both Chi2 and IG with exception of term size 10 and 20. This is shown in Fig. 4.

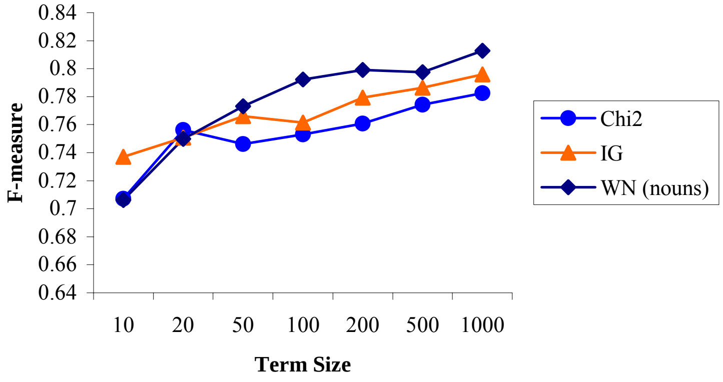

Fig. 4. Comparison of the performance of the WordNet-based POS feature selection approach using nouns, Chi2 and IG

Generally, the WordNet-based POS approach using the nouns feature set performs better from 50 terms onwards because it is capable of choosing constitutive terms that are more reflective of a category. For a smaller number of features, the WordNet-based POS approach is not able to capture adequate representative categorical features. This approach stabilizes as the number of features exceeds 50 features, as a larger number of features are needed to overcome the effects of overfitting that arises due to selecting statistically unique terms with a lesser semantic significance. With fewer terms at 10 and 20 terms, both Chi2 and IG have a lot of terms that have strong statistical values. Although those terms are not in general, representative of the category concerned, it is able to perform well in machine learning because of its statistical patterns. We can see in Table IV, both Chi2 and IG have the terms "lt" and "cts". These two terms are not relevant to the category "acquisition" but are still ranked in the top 20 terms because of its statistical significance. On the other hand, we can see that the WordNet-based POS approach using the nouns feature set has terms that are closely related to the category "acquisition". Here, we can conclude that all the 20 terms listed are in one way or another, reflective of the category "acquisition".

TABLE IV
THE TOP 20 TERMS IN THE WORDNET-BASED POS APPROACH USING THE NOUNS FEATURE SET, CHI2 AND IG FOR THE CATEGORY "ACQUISITION"

WordNet (Nouns)	Chi2	IG
title	shares	cts
company	offer	shares
shares	lt	net
pct	stake	loss
corp	merger	lt
offer	cts	shr
share	acquisition	offer
stock	company	stake
stake	inc	company
acquisition	acquire	merger
merger	net	tonnes
common	loss	wheat
unit	corp	acquisition
buy	usair	trade
agreement	common	inc
board	mln	mln
shareholders	unit	profit
sell	shr	qtr
american	stock	corp
investment	sell	acquire

Therefore, the WordNet-based POS approach using the nouns feature set exhibits superiority in two factors in comparison to both Chi2 and IG. These factors are effective categorization results and a set of features that reflect a category's contents. The combination of these factors substantiates the effectiveness of the WordNet-based POS approach using the nouns feature set.

However, one drawback seen in this WordNet-based POS approach is that the features chosen are constricted to the size of the WordNet's dictionary. In order to overcome this drawback, future works can combine multiple word databases or dictionaries in order to expand the vocabularies to cover a wider domain.

VII. CONCLUSION

We have explored the roles of the different POS in feature selection and we found that nouns best describe a category's contents. It is also able to perform slightly better than all four POS combined, with a much smaller feature set, which is our aim in dimensionality reduction.

The WordNet-based POS approach that uses only nouns is able to choose a set of terms that are more reflective of a category's content. This is in line with inducing a classifier that will act more like a human expert rather than having a classifier rely only on statistical findings, as is the case if Chi2 and IG are used to select features. This research also highlights the potential of the WordNet-based POS approach to be used for other areas of research, such as spam filtering and web document categorization.

REFERENCES

[1] Sebastiani, F. (1999). A tutorial on automated text categorization. In *Proceedings of ASAI-99, 1st Argentinean Symposium on Artificial Intelligence* (Analia Amandi and Ricardo Zunino, eds), pp 7 – 35, Buenos Aires, AR.

[2] McCallum A. and Nigam, K. (1998). A comparison of event model for naive Bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*.

[3] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *In Proceedings of the Tenth European Conference on Machine Learning (ECML)*, pp 137 – 142.

[4] Holmes G. and Trigg L. (1999). A diagnostic tool for tree based supervised classification learning algorithms. In *Proceedings of the Sixth International Conference on Neural Information Processing (ICONIP)*, Perth, Western Australia, Volume II, pp 514 – 519.

[5] Masuyama, T. and Nakagawa, H., "Applying Cascaded Feature Selection to SVM Text Categorization", in the DEXA Workshops, 2002, pp. 241-245.

[6] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3, 4: 235 – 244.

[7] Debole, F. and Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, Melbourne, AS, pp 784 – 788.

[8] Macmillan English Dictionary for Advanced Learners, International Student Edition, Macmillan Education, 2006.

[9] Witten, I. H. and Frank, E. *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.

[10] Scott, S. and Matwin. S. (1998). Text classification using WordNet hypernyms. *Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems*, pp 45 – 51.