

# A Study for Evaluating the Importance of Various Parts of Speech (POS) for Information Retrieval (IR)

**Chirag Shah**

Dept. of CSE

IIT Bombay, Powai

Mumbai - 400 076,

Maharashtra, India.

Email: *chirag@cse.iitb.ac.in*

**Pushpak Bhattacharyya**

Dept. of CSE

IIT Bombay, Powai

Mumbai - 400 076,

Maharashtra, India.

Email: *pb@cse.iitb.ac.in*

## Abstract

Traditionally in the vector space model of document representation for various IR (Information Retrieval) tasks, all the content words are used without considering their individual significance in the language. Such methods treat a document as a bag-of-words and do not exploit any language related information. It is obvious that considering such information in representing the documents can help in improving the performance of various IR tasks, but how to obtain this information is considered to be difficult. One of the information that can be important is the knowledge about the role of various parts-of-speech (POS). Although importance of various POS is very subjective and depends on the application as well as the domain under consideration, it can be very useful to evaluate their importance even in a general setup. In this paper we present a study to understand this importance. We first generate the document vectors using particular POS. We then evaluate how *good* is this representation. This is done by measuring the information content provided by document vectors. This information is then used to reconstruct the document vectors. In order to show that these document vectors are *better* than those of generated by traditional methods, we consider text classification application. We show some improvement in classification accuracy, but more importantly, we demonstrate the consistency in the results and a step toward a new and promising direction for using semantics for IR tasks.

**KEYWORDS:** *Information Retrieval (IR), vector space model, text document representation, semantics, Parts of Speech (POS), entropy, information content, text classification*

## 1 Introduction

Vector space model [Salton *et al.*, 1975] is the most accepted and used method for document representation in Information Retrieval (IR) tasks. Typically after performing preprocessing on the documents, the document vectors are constructed according to some term-weighting scheme like Term Frequency (TF) [Salton, 1989], Inverse Document Frequency (IDF) [Tokunaga and Iwayama, 1994], TF with IDF (TFIDF) [Joachims, 1997], or Weighted IDF (WIDF) [Tokunaga and Iwayama, 1994]. Traditionally these methods do not use any language specific knowledge explicitly. They treat any document as a bag-of-words without being concerned about the role of a particular word that it plays in the language or the relations that may exist among the words. The focus of the work reported here is on the former case, *i.e.*, to understand the significance of various POS and use this knowledge to improve some IR tasks. Typically researchers use all the content words (nouns, verbs, adjectives, and verbs) after preprocessing to construct the document vectors, but it has been unclear if a particular POS is more important than the others. It is realized in some applications of IR that nouns are more important [Turney and Littman, 2002]. However, it is also found that sometimes even stop words can be useful. The importance of various POS is also very subjective. For instance, if the task is to differentiate between positive and negative opinions, then adjectives and adverbs become more important [Turney, 2002]. Biology is found to be adjective-rich (with words like *dorsal*, *cordate*, *ungulate*), while music uses significant adverbs (like *fortissimo*, *softly*). Some studies for determining the role of verbs in document analysis are conducted by Judith Klavans *et al.* [Klavans and Kan, 1998] for event profiles of news articles. However, to the best of our knowledge, there has not been a particular theory or measure for justifying the importance of various POS in a general setup irrespective of the domain.

In this paper we present a study to understand the importance of various parts of speech (POS) from IR point of view. The rest of the paper is organized in two parts. In the first part, we describe a method for evaluating the importance of POS. In this method we first represent the documents in terms of vectors with TFIDF method using a particular POS. The *goodness* of these document vectors is then found using the information content given by them. The importance of any POS is set equal to this *goodness*, which is used to weigh the document vectors according to the POS of their content words. In the second part of the paper, we show how this information about the importance of POS can be used. An application of text classification is considered to show the improvements obtained by considering POS information. Although the improvements are not very significant, they are consistent and follow the intuition.

## I Evaluating the Importance of POS using *Goodness* of Vectors Measure

In this part we describe our experiments for evaluating the *importance* of various POS. The *importance* of a particular POS is determined based on how *good* the document is represented using the words of that POS. This can be a subjective issue. However, we can follow an obvious intuition that *the more the information some representation provides, the better it is*. The idea of finding the information content of the document vectors is originally given by Rong Jin *et al.* [Jin *et al.*, 2001], which is derived from Latent Semantic Indexing (LSI) [Deerwester *et al.*, 1990] and information theory concepts [Shannon, 1948]. In the next couple of sections we elaborate this idea following the experiments.

## 2 Intuition Behind Using Mutual Information for Goodness Measurement

In this section we analyze how to measure the information content of document vectors for evaluating their *goodness*. In order to understand this, we need to enumerate few concepts from information theory. As shown by Shannon [Shannon, 1948] in his classical work on information theory, entropy of an event  $C$  can be defined using its probability distribution  $p_i$  as

$$H(C) = - \sum_{i=1}^l p_i \cdot \log(p_i) \quad (1)$$

where  $l$  is the total number of discrete states. The conditional entropy  $H(C|D)$  can be calculated as

$$H(C|D) = H(C, D) - H(D) \quad (2)$$

where  $H(C, D)$  is the joint entropy of events  $C$  and  $D$ . In particular, the conditional entropy can be found using

$$H(C|D) = - \sum_{i=1}^l \sum_{j=1}^l p_j \cdot p_{(i|j)} \cdot \log(p_{(i|j)}) \quad (3)$$

where  $p_j$  is the probability distribution of random variable  $D$  and  $p_{(i|j)}$  is the conditional probability distribution for random variable  $C$  given random variable  $D$ .

Now we describe how to measure information content using entropy. According to the definition [Press *et al.*, 1993], the mutual information  $I(C, D)$  can be evaluated as<sup>1</sup>

---

<sup>1</sup>More information about how the information content is related to the entropy can be found in [Shannon, 1948; Lathi, 1968].

$$I(C, D) = H(C) + H(D) - H(C, D) \quad (4)$$

$$= H(C) + H(D) - H(C|D) - H(D) \quad (5)$$

(from equation 2)

$$= H(C) - H(C|D) \quad (6)$$

*i.e.*, the difference between  $H(C)$ , the entropy of the random variable  $C$ , and  $H(C|D)$ , the average entropy of the random variable  $C$  given the value of the random variable  $D$ . The entropy of a random variable  $C$  represents the uncertainty in guessing the value of the random variable  $C$ . Therefore, the mutual information  $I(C,D)$  measures the *decrease of the uncertainty* in the value of the random variable  $C$  caused by knowing the value of the random variable  $D$ .

Linking this understanding to the document representation case, the two random variables  $C$  and  $D$  correspond to the *document content* (described in the next section), and the document vectors respectively. Therefore,  $H(C)$  represents the uncertainty in guessing the content of a document, given that we only know that the document is in the collection, while the conditional entropy  $H(C|D)$  measures the uncertainty about the document content given that we are provided the representation vectors for the documents. The difference between these two entropies, *i.e.*, the mutual information  $I(C,D)$ , indicates the decrease of the uncertainty about knowing the document content. In other words, it tells us how much more confidence we gain in guessing the document content after seeing the document vectors. Thus, the mutual information  $I(C,D)$  reflects the *informativeness* of the document vectors generated by the term weighing schemes giving the sense of *goodness* of these schemes.

### 3 Experimental Evidences by Jin, Falusos, and Hauptmann

The intuition that more informative vectors help in improving IR is verified by Rong Jin *et al.* [Jin *et al.*, 2001] through large scale experiments. The authors conducted their experiments on four different term weighing schemes over six different collections. These test collections were taken from TREC, the size of which varied from small collection with 20,000 documents to fairly large collection with 160,000 documents. The average size of the document also varied very much from collection to collection. In such diversity of the corpora also they found that the average precision measures were quite consistent with mutual information in every single case. These compelling experiments proves that *the more informative the document vectors are, the better will be the performance of IR using these vectors.*

### 4 Goodness Calculation

In this section we formalize the ideas expressed in the previous section. Here we are providing the necessary mathematical formulation only. The reader is referred to [Jin *et al.*, 2001] for more details.

Let  $n$  be the number of documents in the collection. Let  $d_1, d_2, \dots, d_n$  be the document vectors in term space. Let  $M$  be the document-term matrix. Each number  $M_{ij}$  in the matrix  $M$  represents the weight

of the  $j^{th}$  word in the  $i^{th}$  document. Let  $D$  be the document-document matrix, which can be found as

$$D = MM^T \quad (7)$$

As defined earlier,  $C$  is the random variable for *document content*. We define *document content* as a set of weighted *concepts* and each *concept* corresponds to an eigenvector of the document-document matrix  $D$ . Thus, the random variable  $C$  is essentially related to and can be defined in the following way: the random variable  $C$  can only take one of the values from the set of eigenvectors  $v_1, v_2, \dots, v_n$  and the eigenvalue  $\lambda_i$  indicates the importance of the eigenvector  $v_i$ . Therefore, we can assume that the probability for the random variable  $C$  to be the eigenvector  $v_i$  is proportional to the eigenvalue  $\lambda_i$ , which enables us to define the probability distribution for random variable  $C$  as the following.

$$P(C = v_i) = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}, 1 \leq i \leq n \quad (8)$$

The random variable  $D$  corresponds to the document vectors. The possible values that it can take are the set of document vectors in the document collection, *i.e.*,  $d_1, d_2, \dots, d_n$ . Since every document in the collection is equiprobable, we can assume the uniform distribution for the random variable  $D$ , that is, the probability for the random variable  $D$  to be any document vector  $d_i$  is a constant, or

$$P(D = d_i) = \frac{1}{n}, 1 \leq i \leq n \quad (9)$$

Now, the document can be viewed as a set of *concepts* and the weight for each *concept* is given by the projection of the document vector on the corresponding axis. Therefore, we can assume that the probability for a document to contain some particular *concept* is proportional to the projection of the document vector on the corresponding *concept* axis. Thus, the conditional probability  $P(C = v_i | D = d_j)$  would be proportional to the projection of document vector  $d_j$  on the *concept* axis  $v_i$ , that is:

$$P(C = v_i | D = d_j) = \frac{|d_j^T v_i|}{\sum_{k=1}^n |d_j^T v_k|} \quad (10)$$

With all these probabilities defined, we can find their respective entropies and finally, the mutual information as defined in equation 6. The more this mutual information for a given method of vector generation, the better is that method.

## 5 Experiments and Results

In this section we provide the implementation details of the ideas proposed in the previous section. We used The British National Corpus (BNC)<sup>2</sup> for our experiments. It is a 100 million word collection of samples from a wide range of sources, designed to represent a wide cross-section of current British English. The text corpus includes extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memorandums, school and university essays, among many other kinds of text.

Each text is segmented into orthographic sentence units, within which each word is automatically assigned a POS (Part Of Speech) code. There are six and a quarter million sentence units in the whole corpus. Segmentation and word-classification was carried out automatically by the CLAWS [Garside and Smith, 1997] stochastic POS tagger developed at the University of Lancaster. The classification scheme used for the corpus distinguishes some 65 parts of speech. The information about the documents that we used for our experiments is given in the following table. For the experiments described in this section, we used training documents only. The results of the experiments are reported in Table 2.

Category name	Training docs	Testing docs set-1	Testing docs set-2
Arts	60	30	60
Imaginative	60	30	60
Natural science	60	30	60
Social science	60	30	60
World affairs	60	30	60
Total	300	150	300

Table 1: BNC Documents used for Experiments

POS	Vector size	$H(C)$	$H(C D)$	$I(C, D)$
Nouns	32,008	3.9467	4.2723e-6	3.9467
Verbs	9,837	3.8168	1.2129e-5	3.8168
Adjectives	14,608	3.7973	1.1996e-5	3.7973
Adverbs	2,982	3.4313	1.8965e-5	3.4312

Table 2: Mutual Information Given by TFIDF Vectors

Now, based on the *goodness*, *i.e.*,  $I(C, D)$  obtained for vectors of particular POS, we assigned weights to terms, which is nothing but the  $I(C, D)$  itself. The next part of the paper explains the application of finding these weights.

<sup>2</sup>Available at: <http://www.hcu.ox.ac.uk/BNC/>

## II Improving the Classification Accuracy using POS Information

In this part we describe how we used the weighted document vectors using POS information for improving classification accuracy. As shown in the previous section, we recreated document vectors using by weighing the terms according to their POS. Here we show the classification using individual POS, using all the content words, and using weighted document vectors.

### 6 Implementation Details

We implemented the following algorithm for creating document vectors, training category models, and finally, finding categories of new documents.

1. Perform preprocessing on documents. This includes removal of stop words [Yang and Wilbur, 1996] and stemming [Porter, 1997].
2. Prepare TFIDF vectors for training documents. IDF (Inverse Document Frequency) for a word  $w$  is defined as [Tokunaga and Iwayama, 1994]

$$IDF(w) = \log \left( \frac{N}{N_w} \right) \quad (11)$$

where  $N$  is the total number of documents and  $N_w$  is the number of documents in which word  $w$  is occurring.

3. Calculate the category vectors using the following formula [Joachims, 1997].

$$\vec{c}_i = \alpha \frac{1}{|C_i|} \sum_{\vec{d} \in C_i} \frac{\vec{d}}{\|\vec{d}\|} - \beta \frac{1}{|D - C_i|} \sum_{\vec{d} \in D - C_i} \frac{\vec{d}}{\|\vec{d}\|} \quad (12)$$

where

$\vec{c}_i$  = Category vector for  $i^{th}$  category

$C_i$  = Set of training documents assigned to  $i^{th}$  category

$\|\vec{d}\|$  = Euclidean length of a vector  $d$

$|D|$  = Total number of documents

Here  $\alpha$  and  $\beta$  are parameters that adjust the relative impact of positive and negative training examples. As recommended in [Buckley *et al.*, 1994],  $\alpha = 16$  and  $\beta = 4$  are taken in our experiments.

4. Prepare TFIDF vectors for testing documents. Since these vectors are created considering the words used in training, the size of the vectors for training and testing will be the same.
5. Find the cosine similarity between each testing document's vector and each category vector using the following formula.

$$Cos(d_i, d_j) = \frac{\sum_{k=1}^n d_{ik} \cdot d_{jk}}{\sqrt{\left(\sum_{k=1}^n d_{ik}^2\right) \left(\sum_{k=1}^n d_{jk}^2\right)}} \quad (13)$$

where  $d_i$  and  $d_j$  are any two given vectors. The size of these vectors is  $n$ . The category whose vector gives the maximum similarity value is the category for the given test document.

## 7 Experiments and Results

After implementing the algorithm given in the previous section, based on the results we found the classification accuracy for each vector set. The classification accuracy is defined as the following

$$\text{Classification accuracy} = \frac{\text{Number of documents correctly classified}}{\text{Total number of test documents}} \quad (14)$$

The results of classification accuracy are given in Table 3.

Vectors nature	Accuracy for set-1	Accuracy for set-2
Only Nouns	86.00%	83.33%
Only Verbs	81.33%	79.00%
Only Adjectives	80.67%	78.67%
Only Adverbs	65.33%	55.67%
All content words	86.67%	83.33%
Weighted All content words	87.33%	83.67%

Table 3: Classification Accuracy

## 8 Analysis of the Results

From the results given in the previous section, the following observations can be made.

1. Among the content words, following relation can be observed regarding the mutual information given by the document vectors prepared using them (see section 5).

$$I(C, D)_{Nouns} > I(C, D)_{Verbs} > I(C, D)_{Adjectives} > I(C, D)_{Adverbs}$$

2. Among the content words, following relation can be observed regarding the classification accuracy given by the vectors generated by them (see section 7).

$$Accuracy_{Nouns} > Accuracy_{Verbs} > Accuracy_{Adjectives} > Accuracy_{Adverbs}$$

This consistency confirms our hypothesis that *the more the information the document vectors give, the better they are.*

3. Combining all the content words gives better classification accuracy, which also follows the intuition.
4. Considering POS information further improves the classification accuracy, which was our claim and we proved it using theoretical tools as well as by experiments.

## 9 Conclusion

In this paper we presented a study to understand the importance of various POS in IR tasks. Importance of POS is considered to be a subjective issue. However, we tried to evaluate it in a general setup considering a particular corpus. This *importance* was derived using the *ability* of various POS in representing the documents. In order to show that such information can be useful in IR, we considered the text classification application. We showed the consistency in the importance that we measured for every POS with the classification accuracy given by the vectors prepared using them. We further demonstrated that using all the content words improves the results, and when we use the POS information, the results are still improved.

Although we have used tagged corpus, one can use some POS tagger for tagging an untagged corpus. Considering the average size of a web-page to be 4KB, Brill Tagger [Brill, 1992] can tag it in about a couple of seconds on a standard home PC. The results that we obtained may not be so significant, but are consistent and promising. Many more explorations can be done in this direction.

## References

- [Brill, 1992] Eric Brill. A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT, 1992.
- [Buckley *et al.*, 1994] C. Buckley, G. Salton, and J. Allan. The Effect of Adding Relevance Information in a Relevance Feedback Environment. In *International ACM SIGIR Conference*, pages 292–300, 1994.
- [Deerwester *et al.*, 1990] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 1990.
- [Garside and Smith, 1997] R. Garside and N. Smith. A hybrid grammatical tagger: Claws4. In R. Garside, G. Leech, and A. McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 102–121. Longman, London, 1997.
- [Jin *et al.*, 2001] Rong Jin, Christos Faloutsos, and Alex G. Hauptmann. Meta-scoring: automatically evaluating term weighting schemes in IR without precision-recall. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–89. ACM Press, 2001.
- [Joachims, 1997] Thorsten Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 143–151, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

- [Klavans and Kan, 1998] Judith Klavans and Min-Yen Kan. Role of verbs in document analysis. In *COLING-ACL*, pages 680–686, 1998.
- [Lathi, 1968] B. P. Lathi. *An Introduction to Random Signals and Communication Theory*. International Textbook Company, Scanton, Pennsylvania, 1968.
- [Porter, 1997] Martin Porter. An algorithm for suffix stripping. In Karen Sparck Jones and Peter Willett, editors, *Readings in Information Retrieval*, chapter 6, pages 313–316. San Francisco: Morgan Kaufmann, 1997.
- [Press *et al.*, 1993] W.H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1993.
- [Salton *et al.*, 1975] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.
- [Salton, 1989] Gerald Salton, editor. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, 1989.
- [Shannon, 1948] Claude Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, pages 379–423, 623–656, July, October, 1948.
- [Tokunaga and Iwayama, 1994] T. Tokunaga and M. Iwayama. Text categorization based on weighted inverse document frequency. Technical Report 94 TR0001, Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, 1994.
- [Turney and Littman, 2002] P. D. Turney and M. L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report ERB-1094 (NRC #44929, National Research Council, Institute for Information Technology, 2002.
- [Turney, 2002] P. D. Turney. Mining the Web for Lexical Knowledge to Improve Keyphrase Extraction: Learning from Labeled and Unlabeled Data. Technical Report ERB-1096, National Research Council Canada, 2002.
- [Yang and Wilbur, 1996] Yiming Yang and John Wilbur. Using corpus statistics to remove redundant words in text categorization. *Journal of the American Society of Information Science*, 47(5), 1996.