NLP Found Helpful (at least for one Text Categorization Task)

Carl Sable and **Kathleen McKeown**
Department of Computer Science
Columbia University
New York, NY 10027
[sable,kathy]@cs.columbia.edu

Kenneth W. Church
AT&T Shannon Laboratory
180 Park Avenue
Florham Park, NJ 07932
kwc@research.att.com

Abstract

Attempts to use natural language processing (NLP) for text categorization and information retrieval (IR) have had mixed results. Nevertheless, there is a strong intuition that NLP is important, at least for some tasks. In this paper, we discuss a task involving captioned images for which the subject and the predicate are critical. The usefulness of NLP for this task is established in two ways. In addition to the standard method of introducing a new system and comparing its performance with others in the literature, we also present evidence from experiments with human subjects showing that NLP generally improves speed and accuracy.

1 Introduction

Most information retrieval (IR) and text categorization research reported in literature relies on "bag of words" approaches; i.e. each text document is represented as a vector of weighted words. Systems generally do not rely on syntax or semantics when computing statistics and making decisions. The use of Natural Language Processing (NLP) to aid text categorization and other IR applications has received a lot of attention, and many believe that there is tremendous potential in this area, but results have been mixed at best (Strzalkowski et al., 1998; Strzalkowski, 1999; Voorhees, 1993; Smeaton, 1999; Elworthy, 2000). Many of these attempts have been applied to specific tasks involving lengthy textual documents for which standard methods have been performing adequately.

Our research has focused on the categorization of multimedia documents based on associated text. The categories applied to multimedia documents can be quite different than categories applied to full-length text documents such as articles, e-mails, or web pages. The experiments discussed in this paper concern the categorization of captioned images that were embedded in news documents concerning disasters, and the possible categories for the images were *Workers Responding*, *Affected People*, *Wreckage*, and *Other*, defined to be mutually exclusive.

Figure 1: Philippine rescuers carry a fire victim March 19 who perished in a blaze at a Manila disco.

Figure 1 shows a sample image from our corpus along with the first sentence of its caption. This caption contains words that a standard bag of words approach would associate with at

least two categories (e.g. "rescuers" → *Workers Responding* and "victim" → *Affected People*). However, the predicate structure of the sentence emphasizes the rescuers, and this particular image was labeled as a member of the *Workers Responding* category, although you can also see wreckage and a victim within the image.

On the other hand, consider an image with a different caption, reading "*A fire victim who perished in a blaze at a Manila disco is carried by Philippine rescuers.*" This caption suggests a focus on the victim as opposed to the rescuers, which implies that the image would be more appropriate for the *Affected People* category. However, the words in the caption are nearly identical. A typical bag of words approach does not have the capacity to distinguish between this hypothetical image and the example shown; each word is either present a certain number of times or it is not, and there is no way to capture predicate structure. For certain tasks involving categories such as the ones we are dealing with here, some linguistic analysis is necessary.

No pre-existing system that we tested was able to perform well on these categories. We eventually became convinced that the main subject and verb of the first sentence of the caption are particularly important in determining the category of an image.[1] These words correspond to the object in the image and to what that object is doing. For example, the most helpful words in the caption of the image shown in Figure 1 are "rescuers" and "carry". The other words are not helpful, and some, such as "victim", can even be misleading.

This paper will first describe an experiment carried out with human volunteers who viewed captions under varying conditions which we feel supports our hypothesis that consideration of syntax is necessary for optimal performance for our task. It then describes a system we developed that uses a shallow parser to extract subjects and verbs automatically, together with a novel measure of word-to-word similarity, to place images into our categories. We will show that this system outperforms seven competing systems which we have tested for this task.

2 The Task

The task discussed in this paper arose naturally in the course of our research, and only after initial attempts applying standard text categorization systems led to poor performance did we begin to consider the use of NLP techniques. The raw data from our corpus consists of news postings from a variety of Usenet newsgroups over a three year period, some of which contain an image with an associated caption. In previous research, human evaluators labeled those news documents which contain images into the categories *Disaster*, *Struggle*, *Politics*, *Crime*, and *Other*. For the experiments discussed in this paper, we started with the 296 images embedded in *Disaster* documents. We chose the *Disaster* category, approximately defined to cover natural disasters and accidents, because our previous system achieves almost perfect precision and recall for this category.

We defined four categories to apply to these images: *Workers Responding*, *Affected People*, *Wreckage*, and *Other*. The categories were defined to be mutually exclusive, and for images that seemed to fit into multiple categories, we asked human evaluators to choose the best fit based on the main focus of the image.[2] Each image was categorized by the first author of this paper and one volunteer who were shown both the image and the caption, and those with agreement were used for the experiments discussed in this paper. There was agreement for 248 images. 98 (39.5%) were classified as *Workers Responding*, 72 (29.0%) were classified as *Affected People*, 55 (22.2%) were classified as *Wreckage*, and 23 (9.3%) were classified as *Other*. The final data set was randomly divided into a training set and a test set, each containing 124 images.

3 Initial Experiments

Our original plan was to use our own classifier, which relies on bins to empirically estimate

[1] Typically, captions contain two or three sentences with the first sentence describing the image and the rest giving background information about the related story.

[2] Instructions provided to the evaluators, including definitions of our categories, can be seen at http://www.cs.columbia.edu/~sable/research/instructions.html.

term weights as described in (Sable and Church, 2001), to place images into these categories. However, we quickly found that the performance was not adequate. We then tested several alternative systems and found that they all had similar performance. Table 1 shows the results of all systems tested. The first six systems in the table comprise the publicly available Rainbow package (McCallum, 1996), and the last is our own bin-based system. The performance of the systems ranged from 54.0% to 59.7%. Choosing the largest category every time would give a baseline performance of 39.5%. While all systems beat the baseline, we did not feel that they were doing as well as possible.

System	Performance
Naive Bayes	55.6%
Rocchio/TF*IDF	54.0%
K-Nearest Neighbor	54.0%
Probabilistic Indexing	59.7%
Maximum Entropy	58.1%
Support Vector Machines	54.8%
Bins	56.5%

Table 1: The initial results were low for all systems.

In order to decide in which category to place an image, it is important to determine what is in the image and what that thing is doing. In the sample image shown in Figure 1, for example, we see *rescuers carrying*, and that focus of the image places it in the *Workers Responding* category. Words in the caption such as "disco" and "victim" refer to items in the image which are indicative of other categories such as *Wreckage* and *Affected People*, but they do not refer to the focus of the image. We formed the hypothesis that the main subject and verb of the first sentence of the caption should play a pivotal role in determining an image's category; if this is correct, it is likely that a system relying on deeper NLP techniques should be able to outperform typical systems for our task. Typical systems relying on bag of words approaches can not account for the predicate argument relationships in the captions.

4 Experiments with Humans

To test our hypothesis, we randomly divided our data set of 248 images into four equally sized subsets and recruited four volunteers to view text associated with our images under four conditions. Each volunteer was a native speaker of English, and none had any connection to this or any related research. The four conditions were:

- Sent: The full first sentence of the caption.
- Rand: The words from the first sentence of the caption in random order.
- IDF: The top two words, not including proper nouns, from the first sentence of the caption, according to TF*IDF weights (Salton and Buckley, 1988; Salton, 1989).
- S-V: The two words, manually extracted, best representing the main subject and verb. If the subject was a proper noun, only the token "NAME" was provided.

Sent	Philippine rescuers carry a fire victim March 19 who perished in a blaze at a Manila disco.
Rand	at perished disco who Manila a a in 19 carry Philippine blaze victim a rescuers March fire
IDF	disco rescuers
S-V	subject = "rescuers", verb = "carry"

Table 2: The subject and verb make it clear that the category for the sample image is *Workers Responding*. Other words such as "disco" and "victim" are not helpful and can be misleading.

Table 2 illustrates the four conditions for the sample image shown in Figure 1. As was the case with many images, the subject and verb alone ("rescuers carry") are enough to confidently predict the category of the image. The top two TF*IDF words might be enough, since "rescuers" happened to be one of them, but "disco" is not helpful. If "victim" had happened to show up instead of "rescuers", this condition would have been misleading. Viewing all the words in

random order is confusing; there are mixed signals here, and unless you take the time to unscramble the words and regain some syntactic clues, you are forced to guess.

A web interface was set up which allows volunteers to predict each categories of images. Each volunteer was tested with a different condition for each of the four subsets of our data, and each subset was presented to our four volunteers with the four different conditions. In this way, a prediction was recorded for every image under each condition once, every volunteer was tested under all conditions, and no volunteer was presented with the same image twice.

Volunteer	Sent	Rand	IDF	S-V
#1	95.2%	83.9%	50.0%	64.5%
#2	95.2%	75.8%	46.8%	74.2%
#3	83.9%	62.9%	56.5%	64.5%
#4	90.3%	75.8%	61.3%	83.9%
Avg	91.1%	74.6%	53.6%	71.8%

Table 3: Subject and verb alone performed almost as well as all words in random order, and much better than the top two TF*IDF words.

Table 3 shows the performance of each volunteer under each condition as well as the overall performance for each condition. All volunteers were reasonably consistent. In summary, Sent $\gg$ Rand $>$ S-V $\gg$ IDF. That is, (1) more words (Sent, Rand) are better than fewer words (S-V, IDF), and (2) NLP helps (Sent is better than Rand and S-V is better than IDF). The NLP effect is remarkably strong and almost compensates for the other effect; i.e. Rand is only slightly better than S-V (for most volunteers).

Condition	Average Time
Rand	68.1
Sent	34.3
IDF	22.7
S-V	20.3

Table 4: Volunteers spent the most time making decisions when presented all words in random order.

In addition to measuring performance, our interface also keeps track of how long each decision takes. Table 4 shows the average time of decisions in seconds under each of the four conditions. As can be seen, volunteers took the longest, by far, to make decisions with the Rand condition. Comparatively, with the S-V condition, they took less than one third of the time.

Examination of these results led us to the conclusion that syntax clearly matters for this task. All volunteers performed much better when shown the full first sentence with words in their original order than when the same words were shown in random order, and the task took approximately half the time. Therefore, any bag of words approach is likely limited by a significantly lower upper bound than one which uses NLP techniques. In particular, the main subject and verb from the sentence were important. Given only these two words, volunteers performed almost as well as when they had all the words in random order, and much better than when they were given the top two words according to TF*IDF weights, a very common measure of word importance in IR literature.

5 Using Only Subjects and Verbs with Standard Systems

System	Performance	
	Sent	S-V
Naive Bayes	55.6%	54.8%
Rocchio/TF*IDF	54.0%	54.0%
K-Nearest Neighbor	54.0%	54.8%
Probabilistic Indexing	59.7%	54.0%
Maximum Entropy	58.1%	53.2%
Support Vector Machines	54.8%	54.0%
Bins	56.5%	53.2%

Table 5: Systems performed almost as well using single word subjects and verbs as they did when provided with the entire first sentence.

We next decided to test how the standard text categorization systems we had previously tested would fair if only subjects and verbs were provided. At this point, we were still using manually extracted words. Table 5 shows how the results using only subjects and verbs compared to results using the entire first sentence of the caption (the first column of results is the same

as that from Table 1). As can be seen, the performance was slightly worse for five of the seven systems, slightly better for one, and the same for another. As with humans, results were almost as high using just two specifically chosen words as when all words in the sentence (not accounting for syntax) were used.

6 NLP Based System

With the results of our experiment with humans in mind, we set out to create a fully automatic text categorization system that takes advantage of our findings. First, our system tries to extract the single words best representing the main subject and verb from the first sentence of each caption in our training set, and these comprise lists of subjects and verbs which are representative of our categories. Next, for each test image, the subject and verb from its caption are extracted, and these are compared to those from the training set using a measure of word-to-word similarity. A score is generated for every category based on these similarities, and the category with the highest score is predicted.

6.1 Extracting Subjects and Verbs

Subjects and verbs are automatically extracted using a three step process. First, Church's statistical part-of-speech tagger, POS (Church, 1988), assigns a grammatical category to every word in each caption. Second, the shallow parser CASS (Abney, 1997) parses each tagged caption. Third, a final script operates on the output of CASS, extracting the heads of the appropriate noun phrase and verb phrase to obtain the single words assumed to best represent the subject and verb of the sentence. (If CASS considers the head of the noun phrase to be a name, the token "NAME" is used instead). On our test set, this process leads to an accuracy of 83.9% for subjects and 80.6% for verbs, according to our manually extracted words. WordNet (Fellbaum, 1998) is used to convert each extracted subject and verb to its morphological base-word.

6.2 Word Similarity

In order to compare subjects and verbs extracted from test captions to those from the training set, we examined a large "extended" corpus consisting of thousands of news articles and captions taken from the same newsgroups as the images discussed in this paper. Using the same method of extraction as discussed in Section 6.1, the single words best representing the subjects and verbs were extracted from every sentence of every article and caption in the extended corpus. When dealing with text categorization, the creation of the corpus is generally one of the most time consuming tasks, since documents usually need to be manually labeled for the training set, but for the purposes of word similarity as we are doing it, this extended corpus is unlabeled and easily obtainable.

Based on these extracted subject/verb pairs, we defined the similarity between two subjects to be the percentage of verbs they share in common, and the similarity between two verbs to be the percentage of subjects they share in common. The idea was that two subjects should be considered similar if they often partake in similar actions, and that two verbs should be considered similar if they represent actions that are often executed by similar entities. This is not necessarily a good measure of word similarity for other tasks, but we thought it might work well for this domain. For example, let's say that the word "fireman" never appears in the training set of the corpus, but words such as "policeman" and "volunteer" do, in captions from images belonging to the category *Workers Responding*; these subjects likely share a higher percentage of verbs in common than most randomly selected pairs of subjects, and would therefore have a relatively high similarity. In addition, for our current domain, they are representative of the same category (*Workers Responding*). By our definitions, the similarity between any subject or verb and itself comes out to be one, and the similarity between any two non-identical words is generally much less.

We also defined the similarity between a subject and a verb to be twice the number of times they appear together divided by the total number of times each appears. Therefore, if the subject/verb pair always appears together, the similarity between the two words would be one, and

otherwise it would be less. The idea is that subjects which are likely to perform actions seen as representative of a category should in and of themselves be considered representative of the category. The same is true for verbs which represent actions that are likely to be performed by subjects that are representative of a category. For example, let's say that the word "fireman" never appears in the training set of the corpus, but verbs such as "help" and "rescue" do, in captions from images belonging to the category *Workers Responding*. Since a "fireman" is more likely to "help" and "rescue" than perform other activities, it will contribute more to the *Workers Responding* category than to others.

6.3 Choosing a Category

To choose a category for some specified image, the single word subject and verb from the first sentence of its caption are extracted, all relevant similarities are added together for each category, and the category with the highest score is then predicted. More formally, let C be the set of categories, and for some specific category c, let S_c and V_c be the set of subjects and verbs extracted from training instances of c, respectively. For a particular test image d, let s_d and v_d be the single word subject and verb extracted, respectively. For any two words w_1 and w_2, regardless of whether they are subjects or verbs, let Sim_{w_1,w_2} be the similarity between the two words as defined in the previous subsection (any similarity involving a "NAME" token is defined to be 0). Let $T(c|d)$ be the total score for a category c given a test document d. Then:

$$T(c|d) = \left(\begin{array}{c} \sum_{s_c \in S_c} [Sim_{s_d,s_c} + Sim_{v_d,s_c}] \\ + \sum_{v_c \in V_c} [Sim_{s_d,v_c} + Sim_{v_d,v_c}] \end{array} \right)$$

For a document d which does not have a "NAME" token extracted as the subject, the chosen category is simply:

$$\underset{c \in C}{argmax}[T(c|d)]$$

In order to take "NAME" tokens into account when they are extracted (this occurs in 16 of the 124 test cases), we decided to multiply the score for each category by the a-priori probability of the category, based on the training set, given that a "NAME" token is extracted. For example, in the training set, 56.0% of the "NAME" tokens come from the *Affected People* category, whereas only 33.9% of the training images belong to this category overall, so the final score for the *Affect People* category is multiplied by 0.56 if a "NAME" token is extracted to account for the new skew. More formally, let $P(N|c)$ be the estimated probability of a "NAME" token given a category, based on the training set. Then the category chosen for a document d that has a "NAME" token extracted is:

$$\underset{c \in C}{argmax}[T(c|d) \times P(N|c)]$$

7 Results and Evaluation

System	Performance	
	Sent	S-V
Naive Bayes	55.6%	54.8%
Rocchio/TF*IDF	54.0%	54.0%
K-Nearest Neighbor	54.0%	54.8%
Probabilistic Indexing	59.7%	54.0%
Maximum Entropy	58.1%	53.2%
Support Vector Machines	54.8%	54.0%
Bins	56.5%	53.2%
NLP Based system	—	65.3%

Table 6: Our NLP based system outperforms seven standard systems by a considerable margin.

The final line of Table 6, which is otherwise the same as Table 5, shows the performance of our NLP based system described in the previous section. As can be seen, the system's accuracy of 65.3% is at least 10% higher than the seven standard systems achieved when using only subjects and verbs (manually extracted for the standard systems), and it is at least 5% higher than the seven standard systems achieved when given the entire first sentence of each caption. Looking back at Section 4, we see that humans given only the subject and verb of each sentence achieved, on average, a 71.8% accuracy. We consider this a reasonable upper bound for the accuracy that a system such as ours might achieve.

8 Related Work

There has long been a lot of interest in combining NLP and IR. Some of the recent work by various researchers in this area is summarized in (Strzalkowski et al., 1998) and (Strzalkowski, 1999), and as can be seen, the results have been mixed, at best. Recently, there has been some success using NLP to aid in the retrieval of images. Smeaton and Quigley (1996) showed some improvement using WordNet to compute noun to noun similarities which were then used to compare queries with captions. Elworthy (2000) showed improvement using an NLP technique he calls "phrase matching" which first converts queries and captions to "dependency structures". In both of these cases, the researchers manually constructed captions for their images, and in the case of Smeaton and Quigley, they manually disambiguated all words.

Working on domain-specific text categorization tasks involving full length news articles, Riloff has created the system AutoSlog-TS (Riloff and Lorenzen, 1999) which relies on NLP techniques to fully-automatically create dictionaries of "augmented relevancy signatures" which can then be used to improve results for binary text categorization tasks. She found that her system, which labels a document in a category if any augmented relevancy signature associated with the category is found in the document, performs about as well with automatically constructed dictionaries as it does with hand constructed dictionaries and much better than when no dictionary is used at all. No comparison was made to other standard text categorization techniques.

With IR tasks such as query expansion and word sense disambiguation in mind, there have been previous attempts at measuring word-to-word similarity. The research discussed in (Sussna, 1993), (Resnik, 1999), and (Richardson et al., 1994) concerns using WordNet link structure to determine semantic similarity between nouns. Our task also requires us to compute similarity between verbs with each other and verbs with nouns, so the techniques discussed in these papers do not apply. Our approach is simpler, and not necessarily appropriate for general tasks, but it serves our intended purpose well and leads to positive results in our experiments.

Other commonly used metrics to measure word-to-word similarity for use with NLP applications include the Jaccard Coefficient and the Dice Coefficient (Radecki, 1982; van Rijsbergen, 1979; Smadja et al., 1996). These measures are related to the ratio of the frequency with which two words appear together (i.e. near each other) in text to the frequencies of the two words independently. While simple and general, they do not apply well to our specific task and domain. For example, "rescuers" and "victim" might often appear together in text, as they do in the caption of the sample image in Figure 1, but for our current categorization task, as subjects they would be indicative of two different categories (*Workers Responding* versus *Affected People*), and we do not want them to be considered similar. On the other hand, words such as "firefighter" and "fireman" may hardly ever appear together, since an author will likely use one or the other consistently, but for our task, they should be considered very similar. Our method of measuring word-to-word similarity takes these problems into account.

9 Conclusions

We have shown that NLP is important for a particular text categorization task. We believe that this importance depends on both the task and domain. NLP becomes helpful when we are dealing with tasks that rely on focus, perspective, point of view, etc. Admittedly, most of the standard IR test collections are not like this, and bag of words approaches work well for them. However, we believe that tasks such as the one described in this paper, which arose naturally in the course of our research, will continue to appear, and when they do, approaches similar to ours will be useful.

The categories discussed in this paper are not nominal categories determined by the presence or absence of any specific object in an image. These categories deal with predicate argument relationships that can only be determined using

linguistic analysis. Looking back, once again, at Figure 1, we see that the subject and verb of the first sentence of the caption refer to the object of focus in the image and the action taking place. The phrase "rescuers carry" is a clear indication of the *Workers Responding* category, whereas other words that might have high IDF weights, such as "disco" and "victims", would not be helpful and may even be misleading to any system using a bag of words approach. We have verified the importance of NLP for our task by presenting evidence from experiments with human subjects, and we have described a new NLP based system which considerably outperforms seven standard systems. This is a positive result which shows promise for combining NLP and IR in the future, at least for certain tasks.

References

S. Abney. 1997. The scol manual (version 0.1b).

K. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-88).*

D. Elworthy. 2000. Retrieval from captioned image databases using natural language processing. In *Proceedings of the 9th International Conference on Information Knowledge and Management (CIKM-00).*

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press.

A. McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification, and clustering. http://www.cs.cmu.edu/~mccallum/bow.

T. Radecki. 1982. Similarity measures for boolean search request formulations. *Journal of the American Society for Information Science*, 33(1):8–17.

P. Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, **11**:95–130.

R. Richardson, A. F. Smeaton, and J. Murphy. 1994. Using WordNet as a knowledge base for measuring semantic similarity between words. Technical Report CA-1294, Dublin City University.

E. Riloff and J. Lorenzen. 1999. Extraction-based text categorization: Generating domain specific role relationships automatically. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, chapter 7. Kluwer Academic Publishers.

C. Sable and K. W. Church. 2001. Using bins to empirically estimate term weights for text categorization. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01).*

G. Salton and C. Buckley. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, **24**(5):513–523.

G. Salton. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley.

F. Z. Smadja, K. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

A. F. Smeaton and I. Quigley. 1996. Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-96).*

A. F. Smeaton. 1999. Using NLP or NLP resources for information retrieval tasks. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, chapter 4. Kluwer Academic Publishers.

T. Strzalkowski, F. Lin, and J. Perez-Carballo. 1998. Natural language information retrieval: TREC-6 report. In *The Sixth Text Retrieval Conference (TREC-6).* NIST Special Publication 500-240.

T. Strzalkowski, editor. 1999. *Natural Language Information Retrieval.* Kluwer Academic Publishers.

M. Sussna. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the 2nd International Conference on Information Knowledge and Management (CIKM-93).*

C. J. van Rijsbergen. 1979. *Information Retrieval.* Butterworths, London, 2nd edition.

E. M. Voorhees. 1993. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-93).*