Feature Selection for Extracting Semantically Rich Words

Young-Woo Seo Anupriya Ankolekar Katia Sycara

CMU-RI-TR-04-18

March 2004

Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

The utility of semantic knowledge, in the form of ontologies, is widely acknowledged. In particular, semantic knowledge facilitates integration, visualization, and maintenance of information from various sources. However, the majority of previous work in this field has tried to learn ontologies for relatively constrained domains. In other words, to date, there has been relatively little work on trying to construct ontologies for an open domain, where there are enormous needs for such ontologies. Moreover, there have been few studies that empirically examine the value of text learning techniques to extract a set of candidate words for concept words in a domain ontology. The goal of this work is to examine the usefulness of existing feature selection methods for the extraction of a set of good candidate words for concept words in an ontology. From the experimental results, we found that the existing word feature selection methods are quite useful for ontology learning, in that there is a good overlap between the word sets identified by feature selection methods and the words in a manually built domain ontology. Finally, from our experience of working on this paper, we enumerate the desiderata for a domain ontology learning system.

Contents

1 Introduction

Given the rampant amount of textual data these days, it is becoming increasingly important to be able to extract domain-specific semantic content from such texts. Such semantic knowledge, in the form of ontologies, can facilitate integration of information from various sources. Additionally, ontologies enable the visualisation and maintenance of knowledge.

However, in most cases, ontology building is still conducted by hand. It is time-consuming, error-prone, and labor-intensive. Moreover, manual ontology building has a critical weakness, in that the ontology usually reflects the inherent knowledge and biases of its creator, which may not be shared across people. If the ontology were created (semi-)automatically, then such biases will be significantly reduced. Therefore it would be very desirable to have a (semi or fully) automatic method for acquiring a domain ontology.

One of the early attempts at ontology learning was by Faure and Nedellec [3], who proposed applying two techniques from the field of Natural Language Processing (NLP), namely verb-subcategorization and noun-clustering for ontology learning. Kietz and his colleagues [5] developed a method for semi-automatic ontology acquisition for a corporate intranet (e.g., insurance company). They essentially used a number of heuristics to organize a concept hierarchy for the target ontology. While constructing an ontology, a human domain expert was expected to be on hand to intervene in this process by comparing the resulting ontology with a reference ontology. Navigli et al. [11] made use of techniques from Information Retrieval and Machine Learning to resolve ambiguity in the meaning of words and their semantic relationships, which is crucial to building a domain ontology. The performance of their method was evaluated with respect to a number of web pages on travel. Other techniques from Machine Learning and Information Retrieval for building ontologies have been outlined in [8].

However, the majority of this work has tried to learn ontologies for relatively constrained domains. To date, there has been relatively little work on trying to construct ontologies for an open domain. Furthermore, $tf \cdot idf$ is typically used to determine words for the domain ontology concepts. Since $tf \cdot idf$ purely reflects the frequency-based importance of words, it cannot capture dependencies, such as those between a concept in the domain and the words that correspond to that concept.

Text learning techniques, such as statistical feature selection methods, have proven to be useful in extracting more informative words from a given text for a given text learning task. However, there have been few studies that empirically examine the value of text learning techniques to extract a set of candidate words for concept words in an ontology for ontology learning.

The goal of this work is to examine the use of existing feature selection methods for the extraction of a set of good-candidate words for concept words in an ontology. In order to do this, we use a number of existing feature selection methods to identify sets of candidate concept words. These sets are then evaluated with respect to manually created domain ontologies [1].

In the next section, we present the feature selection methods used in this paper. The feature selection experiments and results in detail are in the following section. Then, we discuss the results and desiderata for a domain ontology learning system. Finally

we present possible benefits and extensions of this work.

2 Feature Selection Methods

Feature selection generally refers to the way of selecting a set of features which is more informative in executing a given machine learning task while removing irrelevant or redundant features. This process ultimately leads to the reduction of dimensionality of the original feature space, but the selected feature set should contain sufficient or more reliable information about the original data set. For the text domain, this will be formulated into the problem of identifying the most informative word features within a set of documents for a given text learning task.

Feature selection methods have relied heavily on the analysis of the characteristics of a given data set through statistical or information-theoretical measures. For text-learning tasks, for example, they primarily count on the vocabulary-specific characteristics of given textual data set to identify good word features. Although the statistics itself does not care about the meaning of text, these methods have been proved to be useful for text learning tasks (e.g., classification and clustering).

In our study, we considered four methods, mutual information, χ^2 statistic, Markov blanket, and information gain. Each of these methods uses its own criterion to winnow a subset of the original feature space that seems to best capture characteristics of a given data set. Then the word features selection is done by selecting ones with the highest computation values. For text representation, we employed a multinomial model [10]. Specifically, a text document, D_i is a sequence of word events, $D_i = \{X_1, ..., X_t, ..., X_{|D_i|}\}$, drawn from a multinomial distribution of words in the identified vocabulary, V. X_t is a random variable for *t*th word in a document. Each of documents was assigned to one of the following class labels, $C = \{C_1, ..., C_j, ..., C_n\}$.

2.1 Mutual Information

The mutual information $I(X_t, C_j)$ of two random variables is the relative entropy between the joint distribution and the product distribution $P(X_t)P(C_j)$ [2].

$$\begin{aligned} I(X_t, C_j) &= H(C_j) - H(C_j|X_t) \\ &\approx log\frac{P(C_j \wedge X_t)}{P(C_j) \times P(X_t)} \end{aligned}$$

In particular, it is the reduction in the uncertainty of one random variable C_j due to knowledge of another, X_t. The less dependent X_t and C_j are, the closer $I(X_t, C_j)$ is to zero. This is commonly used in identifying word associations in Natural Language Processing.

2.2 χ^2 Statistic

The χ^2 statistic measures the lack of independence between X_t and C_j by comparing the observed co-occurrence frequencies in a 2-way contingency table with the frequen-

cies expected for independence.

$$\chi^2(X_t, C_j) = \frac{|D| \times (ad - cb)^2}{(a+c) \times (b+d) \times (a+b) \times (c+d)}$$

where a is the number of times X_t and C_j co-occur, b is the number of times X_t occurs without C_j, c is the number of times C_j occurs without X_t, d is the number of times neither C_j nor X_t occurs, and $|D|$ is the total number of documents. C_j and X_t are dependent if the difference between observed and expected frequencies is large whereas they are independent if the χ^2 statistics score is close to zero [9].

The scores derived from mutual information and χ^2 statistics should be interpreted with care. In the case of mutual information, low-frequency word features can have higher scores than more common ones whereas scores computed from the χ^2 statistic are known not to be reliable for low-frequency word features [13].

2.3 Markov Blanket

Markov blanket has been used to remove those word features from the original feature set, whose power to discriminate between classes is subsumed by other features in the set (i.e., their Markov blanket) [6]. The Markov blanket of X_t is defined by the features within the Markov boundary of X_t. In a directed acyclic graph (DAG), the Markov boundary of X_t is defined as X_t's parents, children, and other parents of its children. Therefore, removing X_t from the feature set should not make any difference if the Markov blanket of X_t is already in the feature set. Since the information X_t provides is subsumed by its Markov blanket, in some sense X_t is redundant.

Let M_t be the Markov blanket for X_t. Eliminating X_t is not harmful if the expected cross entropy δ_t between a feature X_t and its Markov blanket M_t is minimized.

$$\delta_i = P(X_i) D(P(C_j|X_i)||P(C_j|M_i))$$

where, $D(p||q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)}$ is called a relative entropy or Kullback-Leibler divergence that measures the difference between two probability distribution over the same event space [2]. As it is very hard to find a full Markov blanket for a feature, we made use of an approximate algorithm proposed in [6].

2.4 Information Gain

The last method we used for feature selection uses the information gain of each word feature. The information gain $IG(X_t)$ of a word feature X_t is defined as an expected reduction in entropy by selecting X_t.

$$IG(X_t) = -\sum_{j=1}^{k} P(c_j) log P(c_j)$$

$$+P(X_t)\sum_{j=1}^{k} P(c_j|X_t)logP(c_j|X_t)$$

$$+P(\tilde{X}_t\sum_{j=1}^{k} P(c_j|\tilde{X}_t)logP(c_j|\tilde{X}_t)$$

It is considered a global measure because it averages the reduction of uncertainty that occurs by the selection of feature X_t over all classes.

3 Experiments

Our objective is to compare the set of word features identified by statistics-based feature selection methods with the concept words in an ontology. Here, we model the domain of an ontology as the target class for the feature selection methods. To achieve this objective, we measured the overlap between words that occurred in the ontology and their ranking in descending order by the four feature selection methods: mutual information, χ^2 statistics, markov blanket and information gain. The overlap was measured as follows: we used the automatic feature selection methods to first rank all the words in the vocabulary set of a category. Then, we noted the rank for those words that also appeared in the ontology. This was done for four different categories, as explained in the next section.

Assuming that the target class label is the core concept in a domain ontology, words (i.e., concepts) in the ontology have a natural rank based on their distances from the core concept. However we do not consider their distance as ranks for comparison because this does not address our objective which is to investigate how useful the existing methods for feature selection are for our task of ontology building.

3.1 Text Data set

We used the publicly available Reuters-21578 dataset [7], which consists of world news stories from 1987 and has become a benchmark in text learning evaluations. While it is more difficult to create ontologies for such data, we feel, corresponds much better to real-world contexts for ontology learning.

The data set has been labelled by human with respect to a list of categories. These categories have been grouped into super-categories of people, topics, places, organizations etc. The category distribution is skewed: the most common category has a training-set frequency of 2,877, but 82% of the categories have less than 100 instances and 33% of the categories have less than 10 instances.

We use the "ModLewis" split of Reuters-21578, which contains 13,625 training documents and 6,188 testing documents leaving 1,765 unused documents. There are 135 overlapping topic categories, but we used only those 4 for which there exists a relatively large set of documents across the training set: `cocoa`, `copper`, `cotton`, and `nat-gas` (natural-gas). We limited ourselves to four categories for this paper because manual building corresponding ontologies is time-consuming task [1].

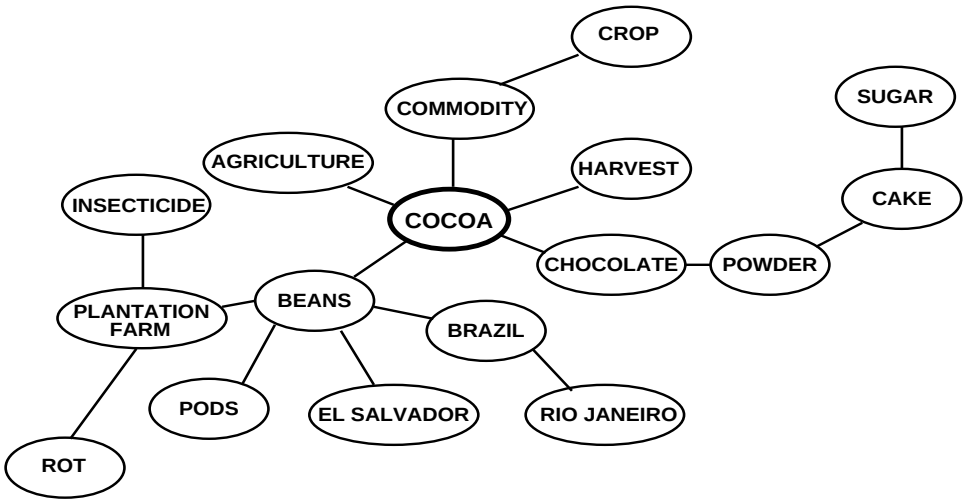

Figure 1: Part of ontology manually constructed for the `cocoa`category.

3.2 Ontologies

In order to evaluate the results of feature selection for the four different categories, we constructed manually ontologies for each of the categories `cocoa`, `copper`, `cotton` and `nat-gas`.

This was done as follows: first, for each category, a lexicon of all the distinct words that appeared in documents within that category after eliminating stop words [1] was compiled. Next, the lexicon associated with the category was examined manually and a subset of the words in the lexicon picked as being most closely associated semantically with the category. This list was then used to create an ontology.

An important choice that was made concerned how exactly a portion of an ontology is used to provide word features for evaluation of the feature selection methods. A straightforward idea is to use the ontological concept labels as word features. Since the concept labels are primarily words contained in the text collection, they can be easily compared with word features identified by feature selection methods. Our text representation model assume unigram (single word)word features. Therefore it cannot capture bigram (two-word) concept labels. However, since the majority of concept labels are unigrams, this does not pose a severe restriction.

A part of the ontology constructed for `cocoa` is shown in Figure 1. The ovals represent concepts and the lines connecting them represent properties or relations between concepts. This is a highly simplified version of the actual ontology. In particular, we do not consider the different kinds of relations connecting the concepts. Furthermore, we are primarily making use of the ontological structure, as opposed to the deductive capabilities a full ontology enables.

[1] A list of stop words is defined as common functional words such as “and”, “of”, “or”, “the”, etc., which are irrelevant to represent the content of text. [12]

Ontology	MI	χ^2	MB	IG
cocoa	patterson	tonnes	temporao	lt
chocolate	vengeance	sugar	alleviating	qtr
powder	eshleman	production	carnival	stock
crushed	melicias	mln	dificulties	bank
butter	muscles	export	comissaria	company
bean	roldan	prices	covertible	shares
agriculture	flex	imports	liquor	vs
commodity	flaws	cocoa	arroba	unproc
harvest	nastro	pct	origins	type
cake	maccia	week	bahian	dlrs
sugar	demico	total	undeclared	cts
crop	barros	traders	pollinates	net
plantation	wrangles	report	doubly	dollar
pods	practised	oil	swollen	dividend
⋮	⋮	⋮	⋮	⋮

Table 1: Word features extracted by the various feature selection methods for the category `cocoa`. (MI: mutual information; MB: Markov blanket; χ^2: χ^2 statistics; IG: information gain)

3.3 Experimental Results

The feature selection experiment for the Reuters-21578 data set resulted in the kind of data shown in Table 1. The table shows typical word features extracted for the `cocoa` category by the four feature selection methods. The various feature selection methods selected very different word features and their individual biases are clearly visible.

Figures 2 to 5 present the results of the feature selection methods for the four categories `cocoa`, `cotton`, `copper` and `nat-gas`, showing the number of words identified that were in the ontology. The horizontal axis indicates the rank of words that were present in the category lexicon, and the vertical axis represents the cumulative number of words that occurred in the ontology. Thus, the graphs are to be read as follows. For any rank x on the horizontal axis, the graph shows the number of words in the ontology that occurred in the top x number of features as determined by the four feature selection methods.

As can be seen from the graphs, the performance of the feature selection is relatively good, as in, a large number of words in the ontology show up relatively high in the ranking by both automatic feature selection methods. The results are best for the `copper` category, then the `cocoa` category, `cotton` category and finally the `cocoa` category. Generally, in order to find half of the words in the ontology, one needs only go through about 200 words in the word list identified by feature selection methods.

The results for the `nat-gas` category were not as good as that of either of the other three categories. There are a number of possible reasons that can account for the bad performance. Firstly, the `nat-gas` category was a difficult case, because the lexicon for that category had a significant overlap with the lexicons for other categories

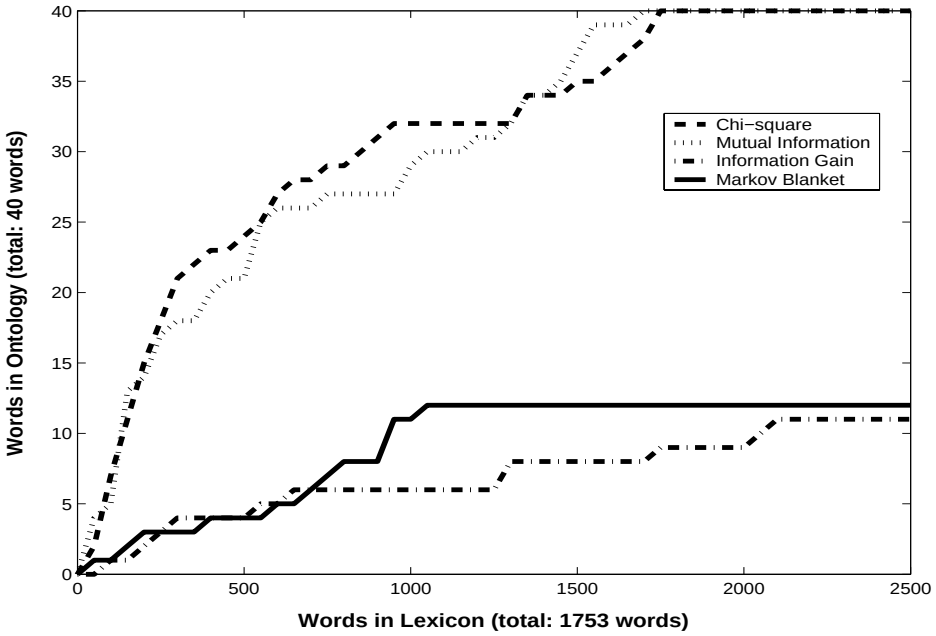

Figure 2: The four feature selection methods for the `cocoa` category.

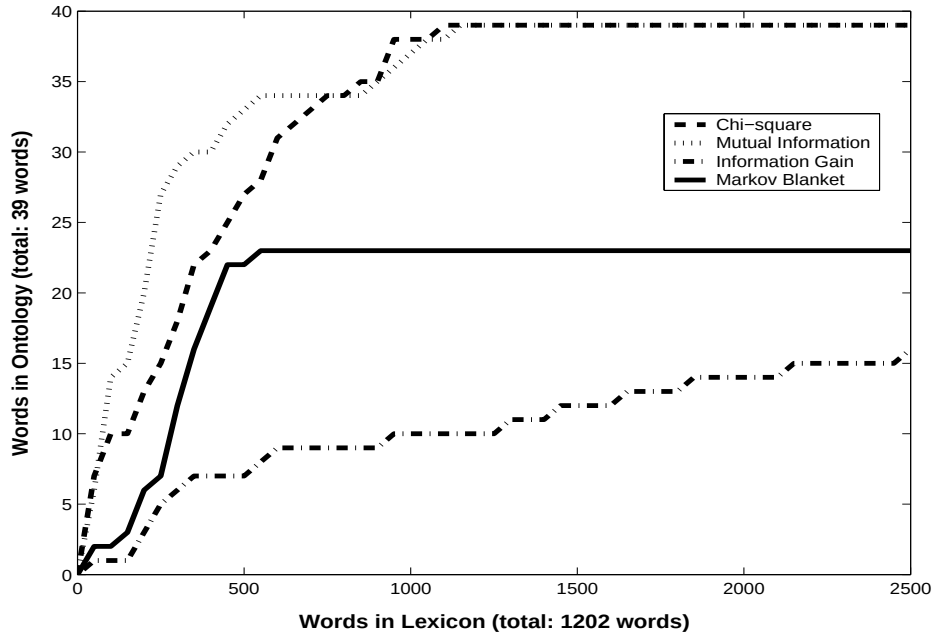

Figure 3: The four feature selection methods for the `copper` category.

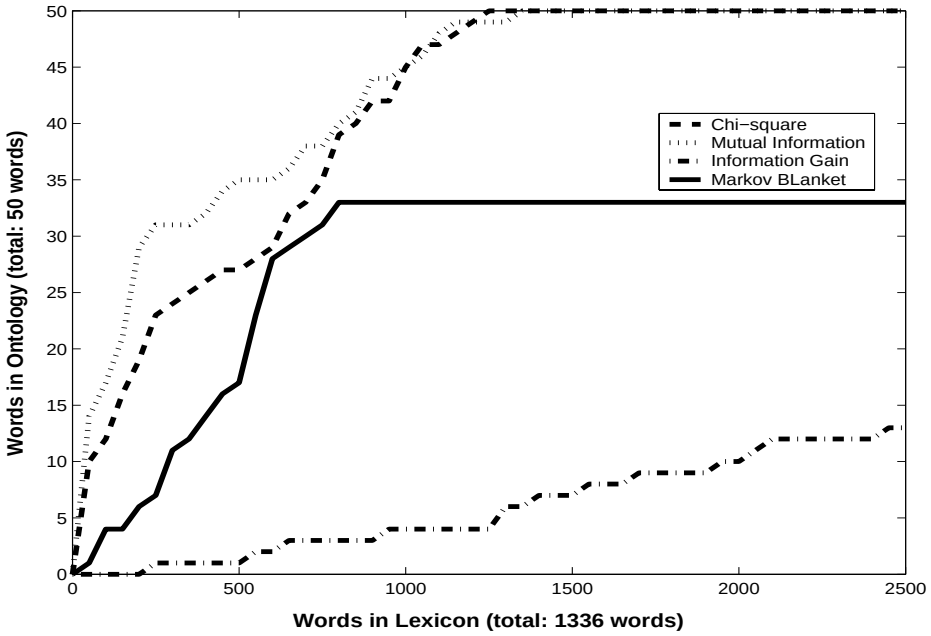

Figure 4: The four feature selection methods for the `cotton` category.

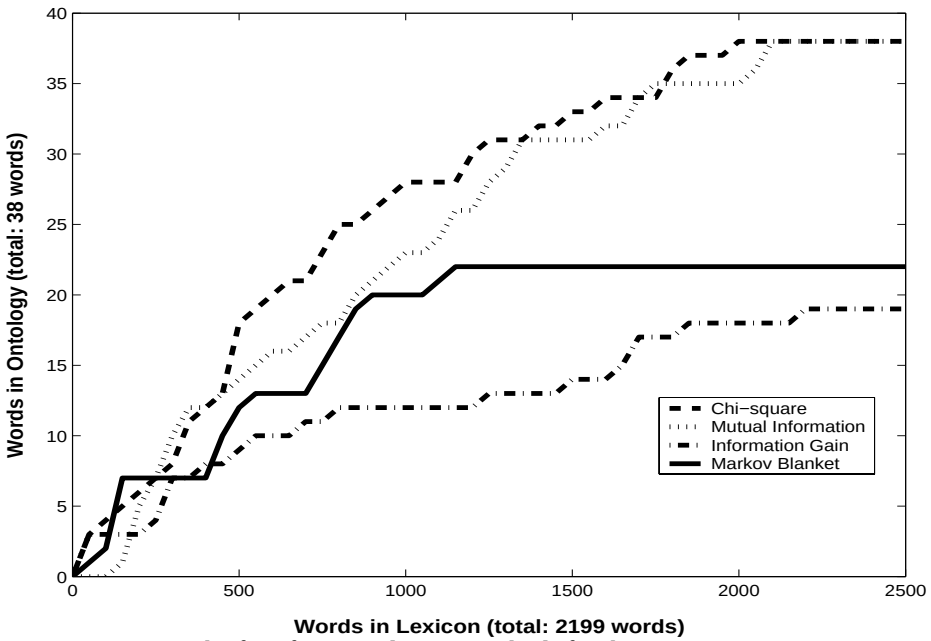

Figure 5: The four feature selection methods for the `nat-gas` category.

in the corpus, such as `oil` and `pet-chem` (pertrochemicals). This lead to the terms included in the ontology having less discriminating power. Also, the manually constructed ontology for natural gas itself is suspect, as its construction required domain knowledge which was not present.

Of the four feature selection methods, information gain has the worst performance in identifying word features in the ontology. This is perhaps unsurprising, as it is the only global measure of the four and was included primarily as a baseline.

The Markov blanket feature selection method performs better than information gain. However, it is also limited in that it can only identify a relatively small set of relevant word features. The curves for Markov Blanket feature selection flatten relatively quickly at the size of the overlap of the Markov Blanket for that category and the words in the ontology.

The mutual information and χ^2 methods are far superior to the other two feature selection methods. Between the two, there is no clear winner as far as selecting several word features in the ontology goes. These two methods are recognized in information retrieval for their ability to capture dependencies between word features and the class label. Therefore, they were able to identify words with high semantic content for the class, as required for ontologies. Since they require a class label to be effective, though, they are only useful for building domain-specific ontologies.

4 Discussion

In order to place our contributions within the broader context of domain ontology learning, we need to first understand what domain ontology learning refers to. A domain ontology identifies and defines a set of relevant concepts that characterize a given domain. It contains a set of generic concepts together with their definitions and interrelationships. Domain ontology learning refers to the acquisition of such domain ontologies from given information about the domain. In particular, the kind of information we are dealing with here is a collection of textual documents.

From our experience of working on this paper, we believe the following are important components of domain ontology learning:

- A component for any domain ontology learning method is the identification of a set of candidate words for concept words in the domain ontology and their interrelationships. The contributions described in this paper are located in this step. We used existing feature selection methods for the extraction of concept words for various domain ontologies. Other Natural Language Processing (NLP) techniques could also be useful for this step.
- Another component is some kind of reference ontology, like WordNet, which specifies basic relationship between concepts. The WordNet [4], for example, specifies taxonomical relationships between words, which is useful to provide a hierarchical structure for identified concept words.
- Since any domain ontology learning method involves a certain degree of inaccuracy, ideally, a method should provide a confidence degree for each component of the ontology, concepts and relationships.

- The creation of a domain ontology is the initial step in the life cycle of the ontology. As the data collection changes dynamically, the ontology will need to be constantly updated to accurately reflect the content of the collection.
- We expect that domain ontology learning cannot be achieved fully automatically. Some measure of human intervention will prove to be necessary because building a domain ontology requires the context of data set and the context of usage of the ontology.

5 Conclusions and Future Work

We set out with the hypothesis that a word feature set identified by the existing feature selection methods would be useful for domain ontology learning. In this paper, we presented an experiment to evaluate the goodness of fit of word feature sets identified by feature selection methods for ontology learning.

The experimental results on feature selection showed that there was a good overlap between the word feature set identified by existing feature selection methods and the word feature set derived from a domain ontology. With the sole exception of the `nat-gas` category, half of the words in the category ontologies were placed approximately in the top 200 word features identified by the statistics-based feature selection. The mutual information and χ^2 statistics were particularly good at identifying candidate concept words. This indicates that the existing feature selection methods could be useful for identifying a set of candidate words for a domain ontology. In summary, the experimental results seem to support our hypothesis on the usefulness of the existing feature selection methods for ontology learning.

Our future work will include the development of other components of domain ontology learning as outlined in the Discussion section. In particular, we will explore NLP techniques to identify interrelationships between identified concept words. Furthermore, we will investigate the maintenance of domain ontologies for dynamic streams of data.

Acknowledgments

The research was funded by the Defense Advanced Research Projects Agency as part of the DARPA Agent Markup Language (DAML) program under the Air Force Research Laboratory contract F30601-00-2-0592 to Carnegie Mellon University.

References

[1] A. Ankolekar, Y.-W. Seo, and K. Sycara. Investigating semantic knowledge for text learning. In *Proceedings of ACM SIGIR Workshop on Semantic Web*, 2003.

[2] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[3] D. Faure and C. Nedellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *LREC Workshop on adapting lexical and corpus resources to sublanguages and applications*, 1998.

[4] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[5] J.-U. Kietz, A. Maedche, and R. Volz. A method for semi-automatic ontology acquisition from a corporate intranet. In *Proceedings of EKAW-2000 Workshop on Ontologies and Text*, 2000.

[6] D. Koller and M. Sahami. Toward optimal feature selection. In *Proceedings of International Conference on Machine Learning (ICML-96)*, pages 284–292, 1996.

[7] D. Lewis. The reuters-21578 data set, 1987. http://www.daviddlewis.com/resources/testcollections/-reuters21578/.

[8] A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, pages 72–79, March/April 2001.

[9] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[10] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI 98 Workshop on Learning for Text Categorization*, pages 41–48, 1998.

[11] R. Navigli, P. Velardi, and A. Gangemi. Ontology learnig and its application to automated terminology translation. *IEEE Intelligent Systems*, pages 22–31, 2003.

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

[13] Y. Yang and J. O. Pederson. A comparative study on feature selection in text categorisation. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 412–420. Morgan Kaufmann Publishers, 1997.