Dhanasekaran K *et al*, International Journal of Computer Science and Mobile Applications,
Vol.1 Issue. 5, November- 2013, pg. 97-109 **ISSN: 2321-8363**

Insight into Information Extraction Method using Natural Language Processing Technique

Dhanasekaran K[1], Rajeswari R[2]

[1]Department of Computer Science and Engineering, Info Institute of Engineering, Anna University, India
dhana0929@gmail.com
[2] Department of Electrical and Engineering, Govt College of Technology, Anna University, India
rreee@gct.ac.in

Abstract

Text mining discovers unseen patterns from textual data sources. But these discoveries are useless unless they contribute valuable knowledge for users who make strategic decisions. Confronting this issue can lead to knowledge discovery from texts, a complicated activity that involves both discovering unseen knowledge and evaluating this potentially valuable knowledge. Information Extraction can benefit from techniques that are useful in data mining or knowledge discovery. However, we can't easily apply data mining techniques to text data for text mining because they assume a complex structure in the source content. Therefore there is a need to use new representations for text data. In many like applications, we can use more structured representations than just keywords to perform analysis to uncover unseen patterns. Early research on such an approach was based on seminal work on exploratory analysis of article titles stored in the Medline medical database. Other approaches have exploited these ideas by combining more elaborated information extraction patterns and general lexical resources such as WordNet or specific concept resources such as thesauri. Another approach, relying on Extracted patterns, uses linguistic resources such as WordNet to assist the discovery and evaluation of patterns to extract basic information from general documents. This paper presents an efficient extraction through classification model to support efficient retrieval and data processing applications. In this paper we discuss various approaches related to text mining, explore issues faced in various research.

Keywords: Data Mining; Dynamic Modelling; Natural Language Processing; Information Extraction; Text Mining

1. Introduction

A system is introduced and developed for the BioCreativeII.5 community evaluation of information extraction of proteins and protein interactions. That focuses primarily on the gene normalization task of recognizing protein mentions in text and mapping them to the appropriate database identifiers based on contextual clues. It outlines a "fuzzy" dictionary lookup approach to protein mention detection that matches regularized text to similarly regularized dictionary entries. It describes several different strategies for gene normalization that focus on species or organism mentions in the text, both globally throughout the document and locally in the immediate vicinity of a protein mention, and present the results of experimentation with a series of system variations that explore the effectiveness of the various normalization strategies, as well as the role of external knowledge sources. While the system was neither the best nor the worst performing system in the evaluation, the gene normalization strategies show promise and the system affords the opportunity to explore some of the variables affecting performance on the BCII.5 tasks. This special issue focuses on applications that innovatively use Web-scale document collections to create useful resources or applications that ultimately let end users navigate the Web more easily [Karin Verspoor, 2010].

One study explores a new domain representation method for natural language processing based on an application of possibility theory. The domain-specific information is extracted from natural language documents using a mathematical process based on Rieger's notion of semantic distances, and represented in the form of possibility distributions. The implementation is done for the distributions in the context of a possibilistic domain classifier, which is trained using the SchoolNet corpus.

This will focus on integrating the complementary notion of necessity into the possibility distributions. We believe this will enrich the representation and make applications, such as our classifier, more accurate [Richard Khoury, 2008].

Some researchers' presents the Bayesian learning approaches to adaptive PLSA modelling for solving updating problems in natural language systems. MAP PLSA was developed for corrective training while QB PLSA was designed for incremental learning. In MAP PLSA, the posterior probability was maximized, which was integrated by a prior density and a likelihood function, and came up with adaptive parameters for document modelling. QB PLSA provided the incremental mechanism of accumulating statistics for model adaptation. At each epoch, PLSA parameters were estimated, and the hyperparameters were updated for future adaptation. Storage of long historical data was avoided. Bayesian PLSA adopted Dirichlet density prior to finding the EM solution.

MAP PLSA and QB PLSA corresponded to statistical algorithms, which were different from folding-in and SVD updating for the linear algebra-based LSA model. This highlighted the contributions of incremental learning where new domain knowledge was continuously updated, or equivalently the out-of-date words/documents gradually faded away. The updating problem and the downdating problem were solved. In the experiments, QB PLSA performed adaptation more efficiently than MAP PLSA.

Both methods consistently improved posterior likelihood through the EM procedure. The PLSA model perplexity was reduced via Bayesian learning. MAP PLSA and QB PLSA outperformed folding-in and SVD updating in terms of precision, recall, mean average precision, and classification error rate. The higher the number of adaptation documents, the better the PLSA modelling that was achieved. These results were evaluated on three public datasets. In the future, we will apply these approaches for language modelling and spoken document retrieval [Jen-Tzung Chien, 2008]. Likewise many approaches has been found and implemented with some satisfactory results. Therefore we summarize the positives and negatives of the approaches in addition with future directions.

2. An investigations into previous work

Anna Corazza and Giorgio Satta(2007) were considered probabilistic context-free grammars, a class of generative devices that has been successfully exploited in several applications of syntactic pattern matching, especially in statistical natural language parsing. They investigate the problem of training probabilistic context-free grammars on the basis of distributions defined over an infinite set of trees or an infinite set of sentences by minimizing the cross-entropy.

This problem has applications in cases of context-free approximation of distributions generated by more expressive statistical models. They show several interesting theoretical properties of probabilistic context-free grammars that are estimated in this way, including the previously unknown equivalence between the grammar cross-entropy with the input distribution and the so-called derivational entropy of the grammar itself.

They discuss important consequences of these results involving the standard application of the maximum-likelihood estimator on finite tree and sentence samples, as well as other finite-state models such as Hidden Markov Models and probabilistic finite automata.

PCFGs are generative devices widely used nowadays in several areas, including natural language processing, speech recognition, and computational biology. The proposed approach has applications in cases where PCFGs are used to approximate other devices that are generatively more powerful. Furthermore, under a theoretical perspective, this general setting has been used to prove some previously unknown properties of PCFGs trained over finite distributions.

Richard Khoury, Fakhri Karray, Yu Sun, Mohamed Kamel and Otman Basir(2007) described that the Modern statistical techniques used in the field of natural language processing are limited in their applications because they suffer from the loss of most of the semantic information contained in text documents. Fuzzy techniques have been proposed as a way to correct this problem through the modeling of the relationships between words while accommodating the ambiguities of natural languages. However, these techniques are currently either restricted to modeling the effects of simple words or are specialized in a single domain.

In this paper, they proposed a novel statistical-fuzzy methodology to represent the actions described in a variety of text documents by modeling the relationships between subject-verb-object triplets. The research focused in the first place on the technique used to accurately extract the triplets from the text, on the necessary equations to compute the statistics of the subject-verb and verb-object pairs, and on the formulas needed to interpolate the fuzzy membership functions from these statistics and on those needed to defuzzify the membership value of unseen triplets.

They have, however, adapted Rieger's formulas to fit this research by specializing them to noun-verb pairs and have added many new features to their system of equations, namely, those regarding the construction and handling of the fuzzy membership functions, as well as those needed to extract the triplets from the text documents.

Taken together, these sets of equations constitute a comprehensive system that allows the quantification and evaluation of the meaning of text documents, while being general enough to be applied to any domain. In the second phase, their paper proceeded to experimentally demonstrate the validity of their new methodology by applying it to the implementation of a fuzzy classifier conceived especially for this research. This classifier is trained using a section of the Brown Corpus, and its efficiency is tested with a corpus of 20 unseen documents drawn from three different domains. The experimental tests show that 15 documents have been correctly classified with an average 42% degree of confidence. Even the remaining five incorrectly classified documents exhibit a weak degree of confidence of only 11%. The positive results obtained from these experimental tests confirm the soundness of their new approach and showed that it is a promising avenue of research.

Some studies have shown that the shape and tuning of the membership functions play an important role in the behavior of the fuzzy controller. Yet, these authors decision to use a trapezoid membership function was reached intuitively and is only justified by the fact that it gives good experimental results.

Clearly, this aspect of their methodology should be analyzed in more depth, and other membership function shapes should be examined in order to determine which shape is most appropriate for their approach. Moreover, the membership functions' maximum value is not always 1. Some kind of normalization step may be needed to fix this problem. The second methodological point that should capture their attention is the triplet generation stage.

As mentioned before, the technique adopted in this paper allows a number of incorrect triplets to be included in the training and testing data. For example, a sentence such as "the house was purchased by the man" will yield the triplet "house purchase man," which is syntactically accurate. However, from a semantic point of view, the action described in that sentence would be better represented by the triplet "man purchase house." Unless a human reader manually extracts the triplets from the corpus, we can expect that such incorrect triplets, brought about by these or by other unexpected sentence structures, will always be part of the data.

Consequently, future studies can investigate the impact of these incorrect triplets on their proposed methodology. In this regard, it would be interesting to know what percentage of incorrect triplets their methodology can tolerate, and what impact these incorrect triplets can have on the accuracy of the results. This particular information is an important input into the decision to invest time and resources into a more accurate triplet extraction system.

From an implementation point of view, the triplet extraction system presents a number of clear weaknesses that could be improved upon. First and foremost is the fact that there is no formal way to cluster the nouns in categories. The categories they selected were simply those that seemed most appropriate given the training corpus. This is a clear weakness of their approach, and some work will be required in order to formalize the choice of noun categories.

Moreover, it is necessary to find a way to handle polysemy, or nouns that can be mapped to several different categories at once. One possible approach has been shown that their system is tolerant to incorrect pairs. Hence, it would be possible to consider a number of different triplets, one for each possible meaning of a noun. This would guarantee that a triplet that uses the correct sense of the noun will be present.

On the other hand, this solution will also cause a major increase in the number of false triplets. It is understood that their system is resistant to a small number of false triplets; however, the impact of a number of false triplets of the magnitude they just proposed remains to be studied.

Another practical improvement that future work should seek is to achieve the goal for the classifier, which is to expand it so that it can cover the entire Brown Corpus, and thus be able to classify texts covering all domains. Although the system's framework is complete, a lot of work remains to be done in order to integrate in it the entire corpus. To their knowledge, a fuzzy classifier on such a scale has never been devised. Alternatively, another corpus could be envisioned as training data, in lieu of the Brown Corpus. Indeed, the Brown Corpus is already 40 years old, and even though variations in English syntax and text structure are eliminated when the text is converted into triplets, new and potentially crucial verbs and nouns that have been added to the English lexicon these past decades are missing from it.

Moreover, the Brown Corpus consists of only 1 million words; a size that was quite respectable at the time of its creation but has become somewhat limited by today's standards, particularly in view of the requirements of their work. Indeed, Rieger's approach to extract a word's meaning from regularities in its usage, on which their methodology is based, requires the analysis of a great number of texts in order to obtain the best results. For this reason, it is imperative to examine the possibility of replacing the Brown Corpus with a larger and more up-to-date training corpus.

In addition to expanding the system horizontally by adding new domains, it should also examine the possibility of expanding it vertically, by adding subdomains to the domains already implemented. This had transformed their classifier into a hierarchical classifier, capable of making an initial classification and refining it subsequently. With this improvement, the classifier could, for example, begin by analyzing a test document and classify it as a belonging to the business domain. Then it would subclassify the document as a text dealing with a corporate takeover, and further subclassify it as describing either a friendly or a hostile takeover.

The implementation of this hierarchical structure would be quite different from the single-level system architecture presented in their paper and would raise a number of new concerns. To illustrate, one such concern relates to the statistics of the new children categories, which could be computed either in relation to each other or in relation to their parent category only. Each scheme has its own advantages and drawbacks, and further efforts are needed to figure out the most appropriate scheme for their approach.

Another concern has to do with how far down in the children categories the system should try to classify a test document, and what its cutoff conditions to end the classification should be. Most importantly, the level of classification accuracy that their methodology can allow, and how precise the children categories can get, also becomes a concern. Finally, it is likely that this new technique has the potential to be applied positively in all fields of NLP, from text classification to semantic search engines to text labeling and summarization.

Although the basic principles presented in this paper do bear scrutiny, the work has so far focused on one specific application. Future research will need to explore new experimental directions in order to make the approach less application-dependent and prove its generality.

Jeremy Morris, and Eric Fosler-Lussier(2008) have presented that the conditional random fields (CRFs) are a statistical framework that has recently gained in popularity in both the automatic speech recognition (ASR) and natural language processing communities because of the different nature of assumptions that are made in predicting sequences of labels compared to the more traditional hidden Markov model (HMM).

In the ASR community, CRFs have been employed in a method similar to that of HMMs, using the sufficient statistics of input data to compute the probability of label sequences given acoustic input. These authors explore the application of CRFs to combine local posterior estimates provided by multilayer perceptrons (MLPs) corresponding to the frame-level prediction of phone classes and phonological attribute classes. They compare phonetic recognition using CRFs to an HMM system trained on the same input features and show that the monophone label CRF is able to achieve superior performance to a monophone-based HMM and performance comparable to a 16 Gaussian mixture triphone-based HMM; in both of these cases, the CRF obtains these results with far fewer free parameters.

The CRF is also able to better combine these posterior estimators, achieving a substantial increase in performance over an HMM-based triphone system by mixing the two highly correlated sets of phone class and phonetic attribute class posteriors.

This paper presented their pilot study into feature-based phone recognition using the model of CRFs. They have also shown that the CRF model can achieve these results with not only a much smaller context, but also with a much smaller set of parameters to model the space with.

Additionally they have shown that features that are highly correlated (such as phonological features and phone classes) can be added to a CRF system in a straightforward manner and give significant improvements in phone recognition performance. In their experiments, these improvements come not at the expense of one set of phones over another set, but instead by raising the overall performance of almost all of the phones in the test set. While adding features to a comparable HMM system does improve correct labeling, it comes at the expense of many spurious insertions that affect overall accuracy.

In contrast, the CRF model shows improvement in overall recognition accuracy, with an increase in correct labels and a reduction in insertions, deletions, and substitutions. In achieving greater accuracy, the system produces fewer insertions and in some sense trades a lower overall correctness for a higher accuracy. In an HMM system, insertions can be tuned through the use of a phone insertion penalty parameter, something that the CRF as implemented here currently lacks.

Ruhi Sarikaya,,Mohamed Afify, Yonggang Deng, Hakan Erdogan, and Yuqing Gao(2008) have proposed language modeling for an inflected language such as Arabic poses new challenges for speech recognition and machine translation due to its rich morphology. Rich morphology results in large increases in out-of-vocabulary (OOV) rate and poor language model parameter estimation in the absence of large quantities of data. In this study, they present a joint morphological-lexical language model (JMLLM) that takes advantage of Arabic morphology. JMLLM combines morphological segments with the underlying lexical items and

additional available information sources with regards to morphological segments and lexical items in a single joint model. Two implementations of the JMLLM were proposed. One called *JMLLM-leaf* that loosely integrates the parse information, while the other tightly integrates the parse information and is referred to as *JMLLM-tree.*

Joint representation and modeling of morphological and lexical items reduces the OOV rate and provides smooth probability estimates while keeping the predictive power of whole words. Speech recognition and machine translation experiments in dialectal-Arabic show improvements over word and morpheme based trigram language models. They also show that as the tightness of integration between different information sources increases, both speech recognition and machine translation performances improve.

Abhinav Sethy, Panayiotis G. Georgiou,Bhuvana Ramabhadran, and Shrikanth Narayanan(2009) have described that the performance of statistical n-gram language models depends heavily on the amount of training text material and the degree to which the training text matches the domain of interest.

The language modeling community is showing a growing interest in using large collections of text (obtainable, for example, from a diverse set of resources on the Internet) to supplement sparse in-domain resources. However, in most cases the style and content of the text harvested from the web differs significantly from the specific nature of these domains. They present a relative entropy based method to select subsets of sentences whose n-gram distribution matches the domain of interest.

They present results on language model adaptation using two speech recognition tasks: a medium vocabulary medical domain doctor-patient dialog system and a large vocabulary transcription system for European Parliamentary Plenary Speeches (EPPS). They show that the proposed subset selection scheme leads to performance improvements over state of the art speech recognition systems in terms of both speech recognition word error rate (WER) and language model perplexity (PPL).

A. ***Summary of contributions***

Their novel scheme is used for selecting *relevant* subsets of sentences from large collections of text acquired from the web. Their results indicate that with this scheme, it can be possible to identify significantly smaller sets of sentences such that the models built from the selected data have a substantially sparser representation and yet perform better (in terms of both perplexity and WER) than models built from the entire corpus. On their medical domain task which had sparse in-domain data (200 K words), they were able to achieve around 4% relative improvement in WER with a factor of 7 reduction in language model parameters while selecting a set of sentences 1/10th the size of the original corpus.

For the TC-STAR task where the in-domain resources were much larger (50Mwords), they achieved 6%relativeWERimprovement by using just 1/3rd of the data. Although most of their results in their paper were on data acquired from the web, the proposed method can easily be used for adaptation of domain specific models from other large generic corpora.

B. ***Scope of their work***

The research effort presented in their paper is directed towards selecting relevant domain specific data from large collections of generic text. They make no assumptions on how the data were collected or what specific web crawling and querying techniques are used. The methods they have developed can be seen as supplementing the research efforts by the machine translation community on identifying web resources or using web counts for language modeling.

They also believe that this work can augment topic based LM adaptation techniques. Topic based LM adaptation schemes typically use LSA or variants to automatically split the available training text across multiple topics. This allows for better modeling of each individual topic in the in-domain collection. The tradeoff is that since the available text is split across topics, each individual model is trained on less data. They believe that this problem can be addressed by selecting data for each topic from large generic corpora using the proposed data selection algorithm.

C. ***Directions for future work***

The effect of varying data granularity has not been studied in this work. They have used sentence level selection, but the selection process can also be naturally extended to groups of sentences, fixed number of words, paragraphs or even entire documents. Selection of data in smaller chunks has the potential to select data better suited to the task but may result in over-fitting to the existing in-domain distribution. In such a case the adaptation model will provide little extra information to the existing model.

They plan to study the effect of this tradeoff between data novelty and match to in-domain model on the LM performance, for different levels of selection granularity. They are also looking into extending the algorithm to work directly on collections of n-gram counts. One motivation for research in this direction is that Google has released aggregate unigram to 5-gram counts for their web snapshot.

The proposed method can be combined with rank-and-select schemes. These authors are exploring the use of ranking to reorder the data such that the sequential selection process gives better results with fewer numbers of randomized searches.

The current framework relies on multiple traversals of data in random sequences to identify the relevant subset. An online single-pass version of the algorithm would be of interest in cases where the text data is available as a continuous stream (one such source is RSS feeds from blogs and news sites). If updates from the stream sources are frequent, iterating through the entire text collection is not feasible.

One of the ideas they are investigating to make the selection process single-pass is to use multiple instances of the algorithm with different initial in-domain models generated by bagging. Voting across these multiple instances can be then used to select data. They are also investigating how to select sentences with a probability proportional to the relative entropy gain instead of the threshold based approach currently being used.

Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen,Susumu Horiguchi and Quang-Thuy Ha(2011) have introduced a hidden topic-based framework to build classification and matching/ranking models for short and sparse text/Web data by taking advantage of hidden topics from large-scale external data collections (e.g., search result snippets, product descriptions, book/movie summaries, and advertising messages) on the Web. The framework focuses on solving two main challenges posed by these kinds of documents: 1) data sparseness and 2) synonyms/homonyms.

The former leads to the lack of shared words and contexts among documents while the latter are big linguistic obstacles in natural language processing (NLP) and information retrieval (IR). The underlying idea of the framework is that common hidden topics discovered from large external data sets (universal data sets), when included, can make short documents less sparse and more topic-oriented. This is possible by performing topic inference for them with a rich source of global information about words/terms and concepts/topics coming from universal data sets.

Therefore, hidden topics from universal data sets help handle unseen data better. They carefully evaluated the framework by carrying out two experiments for two important online applications (Web search result classification and matching/ranking for contextual advertising) with large-scale universal data sets and achieved significant results.

The integration of hidden topics helps uncover and highlight underlying themes of the short and sparse documents, helping us overcome difficulties like synonyms, hyponyms, and vocabulary mismatch, noisy words for better classification, clustering, matching, and ranking. In addition to sparseness and ambiguity reduction, a classifier or matcher built on top of this framework can handle future data better as it inherits a lot of unknown words from the universal data set.

The proposed framework is general and flexible to be applied to different languages and application domains. The experiments for two evaluation tasks have shown how the framework can overcome data sparseness and ambiguity in order to enhance classification, matching, and ranking performance.

Their future studies will be on improving the framework in a number of ways: how to estimate and adjust the number of hidden topics automatically; find more finegrained topic analysis, e.g., hierarchical or nested topics, to meet more sophisticated data and applications; pay more attention to the consistency between the universal data set and the data we need to work with; and incorporate keyword bid information into ad ranking to achieve a full solution to matching and ranking for online contextual advertising.

Zhiwei Lin,Hui Wang and Sally McClean(2012) introduced "A Multidimensional Sequence Approach to Measuring Tree Similarity". It describes the Tree is one of the most common and well-studied data structures in computer science. Measuring the similarity of such structures is key to analyzing this type of data.

However, measuring tree similarity is not trivial due to the inherent complexity of trees and the ensuing large search space. Tree kernel, a state of the art similarity measurement of trees, represents trees as vectors in a feature space and measures similarity in this space. When different features are used, different algorithms are required.

Tree edit distance is another widely used similarity measurement of trees. It measures similarity through edit operations needed to transform one tree to another. Without any restrictions on edit operations, the computation cost is too high to be applicable to large volume of data.

To improve efficiency of tree edit distance, some approximations were introduced into tree edit distance. However, the effectiveness can be compromised. In their paper, a novel approach to measuring tree similarity is presented. Trees are represented as multidimensional sequences and their similarity is measured on the basis of their sequence representations. Multidimensional sequences have their sequential dimensions and spatial dimensions. They measure the sequential similarity by the all common subsequences sequence similarity measurement or the longest common subsequence measurement, and measure the spatial similarity by dynamic time warping. Then they combine them to give a measure of tree similarity.

A brute force algorithm to calculate the similarity will have high computational cost. In the spirit of dynamic programming two efficient algorithms are designed for calculating the similarity, which have quadratic time complexity. The new measurements are evaluated in terms of classification accuracy in two popular classifiers (k-nearest neighbor and support vector machine) and in terms of search effectiveness and efficiency in k-nearest neighbor similarity search, using three different data sets from natural language processing and information retrieval. Experimental results show that the new measurements outperform the benchmark measures consistently and significantly.In the future work, they will study other combinations of existing similarity measures for trees. They also have planned to investigate how this approach can be applied to graphs and images.

Rile Hu, Chengqing Zong, and Bo Xu(2006) introduced "An Approach to Automatic Acquisition of Translation Templates Based on Phrase Structure Extraction and Alignment". The new approach for automatically acquiring translation templates from unannotated bilingual spoken language corpora adopted the two basic algorithms: a grammar induction algorithm, and an alignment algorithm using bracketing

transduction grammar. The approach is unsupervised, statistical, and data-driven, and employs no parsing procedure.

The acquisition procedure consists of two steps. First, a grammar induction algorithm extracts semantic groups and phrase structure groups from both the source language and the target language. Second, an alignment algorithm based on bracketing transduction grammar aligns the phrase structure groups using BTG. The aligned phrase structure groups are post-processed, yielding translation templates. Their method needs fewer resources than the method which is called as "parse-parse-match. Preliminary experimental results show that the algorithm is effective.

However, they were faced many difficult tasks, including the improvement of grammar induction and alignment. In the future, they will introduce more information such as some dictionary information (including a synonym dictionary) and some additional preprocessing.

Der-Chiang Li and Chiao-Wen Liu(2012) investigated on "Extending Attribute Information for Small Data Set Classification". It described the data quantity as the main issue in the small data set problem, because usually insufficient data will not lead to a robust classification performance. How to extract more effective information from a small data set is thus of considerable interest. Their paper included a new attribute construction approach which converts the original data attributes into a higher dimensional feature space to extract more attribute information by a similarity-based algorithm using the classification-oriented fuzzy membership function. They computed the classification- oriented membership degrees to construct new attributes; so-called class-possibility attributes, and also developed an attribute construction procedure to construct new attributes, so-called synthetic attributes, to increase the amount of information for small data set analysis. Seven data sets with different attribute sizes are employed to examine the performance of the proposed method.

The results show that the proposed method has a superior classification performance when compared to principal component analysis (PCA), kernel principal component analysis (KPCA), and kernel independent component analysis (KICA) with a Gaussian kernel in the support vector machine (SVM) classifier.

Learning from small data sets is fundamentally difficult, and many data preprocessing methods have been proposed to improve the analysis performance, such as the PCA, KPCA, and KICA. While KPCA and KICA are widely used nonlinear feature extraction methods, there are still weaknesses in using general purpose kernels to extend data into higher dimensional spaces, because one cannot be sure that the transformed space is suitable for a specific research purpose, such as classification.

It is observed that when the data size is small, there is a high level of uncertainty and the insufficient information often results in less robust analysis. Therefore the authors proposed data attribute construction procedure to extend the data information of small data sets to improve classification performance.

The result suggests that increasing the amount of new purpose-related data information is an effective way to improve the analysis performance for small data sets. In future research, two research directions have been considered. One is combining the use of virtual samples with the proposed method to further enhance the analytical performance. The other is to develop a category-oriented method to deal with the category attributes for extending the categorical attributes' information.

Jerome R. Bellegarda,(2010) have proposed "Part-of-Speech Tagging by Latent Analogy" which presented the POS tagging as a critical first step in various speech and language processing tasks. High-accuracy taggers (e.g., based on conditional random fields) rely on well chosen feature functions to ensure that important characteristics of the empirical training distribution are reflected in the trained model.

This makes them vulnerable to any discrepancy between training and tagging corpora, and, in particular, accuracy is adversely affected by the presence of out-of-vocabulary words. Their paper explores an alternative tagging strategy based on the principle of *latent analogy*, which was originally introduced in the context of a speech synthesis application. In this approach, locally optimal tag subsequences emerge automatically from an appropriate representation of global sentence-level information.

This solution eliminates the need for feature engineering, while exploiting a broader context more conducive to word sense disambiguation. Empirical evidence suggests that, in practice, tagging by latent analogy is essentially competitive with conventional Markovian techniques, while benefiting from substantially less onerous training costs. This opens up the possibility that integration with such techniques may lead to further improvements in tagging accuracy.

In an attempt at cross-fertilization between the areas of (speech-centric) grapheme-to-phoneme conversion and (language-centric) POS tagging, these authors have explored this alternative tagging design based on the emerging principle of latent analogy. This strategy focuses on two loosely coupled subproblems:

1) extract from the training corpus those sentences which are the most germane in a global sense, and 2) exploit the evidence thus gathered to assemble the POS sequence based on local constraints. They address 1) by leveraging the latent topicality of every sentence, as uncovered by a global LSM analysis of the entire training corpus.

Each input surface form thus leads to its own customized neighborhood, comprising those training sentences which are most related to it. POS tagging then follows via locally optimal sequence alignment and maximum likelihood position scoring, in which the influence of the entire neighborhood is implicitly and automatically taken into account. This method was observed to be effective on two different corpora: a subset of the Penn Treebank suitable for conducting benchmark comparisons, and a corpus of more diverse material collected in the context of a concatenative speech synthesis task. In practice, tagging by latent analogy does not quite achieve the same level of performance as maximum entropy tagging, but benefits from substantially less onerous training costs.

This bodes well for its general deployability across a wide range of applications. In addition, it exhibits attractive complementarity with techniques such as CRF, opening up the possibility of integrated solutions which could lead to further improvements in tagging accuracy. Future efforts will concentrate on characterizing the influence of the parameters and on the vector space of sentence anchors, and the corresponding tradeoff between modeling power and generalization properties.

It is also of interest to carefully study the impact of the parameters and outcome of maximum-likelihood estimation. In addition, their current position scoring strategy could benefit from several refinements, including the inclusion of higher level units (such as common lexical compounds), as well as the integration of confidence measures into the final score. Finally, they intend to investigate how to best combine latent analogy and conventional tagging in order to further increase POS accuracy.

So-Young Park,Jeunghyun Byun, Hae-Chang Rim, Do-Gil Lee, and Heuiseok Lim(2010) have introduced an approach which include a natural language based interface model to enable a user to articulate a request without having any specific knowledge about a mobile device.

In consideration of the very limited computing and memory capacity of the mobile device and to keep the development cost low, their proposed model does not depend on typical natural language techniques, but on a ranking technique, that is simplified based on the mathematical derivation process with the following assumptions. One assumption is that a device control command consists of a function and its parameters.

The other assumption is that the parameter is represented as few predictable patterns, whereas the function can be represented as various sentence patterns. To deal with these various sentence patterns, the proposed model selects the top ranked command candidate with the highest score after generating all possible candidates with their scores. Furthermore, the ranking score function is designed to achieve a high discriminative capability by the simulation of the process of generating every candidate. Experimental results show that the proposed model with 2.9 megabytes performs at 96.27% accuracy, which is slightly lower than 97.06% of the baseline model with 135.2 megabytes.

The proposed model consists of a parameter recognition phase and a function recognition phase. The desirable characteristics of the proposed model are outlined as follows: First, the proposed model can considerably reduce the complexity of the transformation problem, by recognizing the unambiguous parameters before recognizing the function which can be represented as various sentence patterns.

Second, the proposed model is small enough to be easily embedded in the mobile device. Instead of depending on the typical natural language processing techniques, the proposed model utilizes the simple ranking score function, the automata, and some rules obtained by analyzing the characteristics of the restricted domain.

Third, the proposed model provides a fast response time, because the proposed model can significantly reduce the search time by excluding the heavy resource. Furthermore, the proposed model employs the trie structure to quickly search for a large number of names.

Fourth, the proposed model can achieve a high accuracy, even with limited resources; because it's ranking score function is designed to consider the generation history of each candidate. Experimental results show that the final accuracy of the proposed model is 96.27%.

For future works, they want to integrate the proposed model with a speech recognition engine, and devise the method of handling noise or errors propagated from the speech recognition engine. Finally, they want to apply the proposed model to other devices, such as a television content search system.

4. Conclusion

In Mining the Web to Create Specialized Glossaries, one approach addresses automatic compilation of domain-specific glossaries from the Web. Terms have different meanings depending on the domain. For example, in the financial domain, "security" refers to investment instruments, whereas in the information technology domain, it refers to the protection of a computer system's integrity. The system includes several components, including modules for term extraction, glossary creation based on Web-mining techniques, and glossary filtering and validation.

The presence of irrelevant features in training data is a significant obstacle for many machine learning tasks. One approach to this problem is to extract appropriate features and, often, one selects a feature extraction method based on the inference algorithm. In a general framework for feature extraction based on Partial Least Squares, one can select a user-defined criterion to compute projection directions. The framework draws together a number of existing results and provides additional insights into several popular feature extraction methods. Key advantages of these approaches include simple implementation and a training time which scales linearly in the number of examples. Furthermore, one can project a new test example using only k kernel evaluations, where k is the output dimensionality. Computational results on several real-world data sets show that SMA and SMC extract features which are as predictive as those found using other popular feature extraction methods. Additionally, on large text retrieval and face detection data sets, they produce features which match the performance of the original ones in conjunction with a Support Vector Machine.

The two most important tasks in information extraction from the Web are webpage structure understanding and natural language sentences processing. However, little work has been done toward an integrated statistical model for understanding webpage structures and processing natural language sentences within the HTML elements.

In "Learning to Tag and Tagging to Learn: A Case Study on Wikipedia," examine named entity recognition (NER), a method for automatically assigning semantic categories to named entities as they appear in running text. Traditionally, NER categories include people, places, organizations, artifacts, and quantities, among others. One of the biggest challenges to NER is that texts in different domains and from different sources use different surface patterns to express the same concept. So, porting a system to a completely new text source requires domain adaptation. In this work, the authors perform domain adaptation and use NER to enrich Wikipedia metadata.

Another article addresses a complementary topic. "Linking Documents to Encyclopedic Knowledge," this paper uses Wikipedia to perform automated keyword extraction and word sense disambiguation. A typical task to which they apply their method is text "wiki- fication"—automatically identifying words and phrases in a document and linking each to the appropriate Wikipedia entry. Keyword extraction is needed to address ambiguous phrases. One such example is "New York Times," which should link to the newspaper's entry, not the city's entry.

On the other hand, word sense disambiguation is needed to decide which target pages with the same entry name to link to. For example, Wikipedia has more than 20 different entries labeled "Chicago," whose topics include not only the eponymous city, musical, movie, and rock band but also a poker game, an asteroid, a ship, and a typeface. The authors address the challenge of finding the right target page. So it is interesting to see that much work is done to address issues in natural language processing using data mining techniques. This paper gives the future researchers the way to find an approach to address various issues in the related area of information extraction, data mining, text mining, web mining, natural language processing, and semantic web.

References

[1] Anna Corazza and Giorgio Satta(2007), "Probabilistic Context-Free Grammars Estimated from Infinite Distributions,"*IEEE Transactions on Pattern Analysis and Machine Intelligence*, VOL. 29, NO. 8.

[2] Richard Khoury, Fakhri Karray, Yu Sun, Mohamed Kamel and Otman Basir(2007)," Semantic Understanding of General Linguistic Items by Means of Fuzzy Set Theory," *IEEE Transactions on Fuzzy Systems*, VOL. 15, NO. 5.

[3] Jeremy Morris, and Eric Fosler-Lussier(2008), "Conditional Random Fields for Integrating Local Discriminative Classifiers," *IEEE Transactions on Audio, Speech, and Language Processing*, VOL. 16, NO. 3.

[4] Ruhi Sarikaya,Mohamed Afify, Yonggang Deng, Hakan Erdogan, and Yuqing Gao(2008),"Joint Morphological-Lexical Language Modeling for Processing Morphologically Rich Languages With Application to Dialectal Arabic," *IEEE Transactions on Audio, Speech, and Language Processing*, VOL. 16, NO. 7.

[5] Abhinav Sethy, Panayiotis G. Georgiou,Bhuvana Ramabhadran, and Shrikanth Narayanan(2009),"An Iterative Relative Entropy Minimization-Based Data Selection Approach for n-Gram Model Adaptation," *IEEE Transactions on Audio, Speech, and Language Processing*, VOL. 17, NO. 1.

[6] Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen,Susumu Horiguchi and Quang-Thuy Ha(2011),"A Hidden Topic-Based Framework toward Building Applications with Short Web Documents," *IEEE Transactions on Knowledge and Data Engineering*, VOL. 23, NO. 7.

[7] Zhiwei Lin,Hui Wang and Sally McClean(2012), "A Multidimensional Sequence Approach to Measuring Tree Similarity,"*IEEE Transactions On Knowledge and Data Engineering*, VOL. 24, NO. 2.

[8] Rile Hu, Chengqing Zong, and Bo Xu(2006), "An Approach to Automatic Acquisition of Translation Templates Based on Phrase Structure Extraction and Alignment," *IEEE Transactions on Audio, Speech, And Language Processing* , VOL. 14, NO. 5.

[9] Der-Chiang Li and Chiao-Wen Liu(2012), "Extending Attribute Information for Small Data Set Classification," *IEEE Transactions On Knowledge and Data Engineering*, VOL. 24, NO. 3,pp.452-464.

[10] Jerome R. Bellegarda(2010), "Part-of-Speech Tagging by Latent Analogy," *IEEE Journal of Selected Topics in Signal Processing*, VOL. 4, NO. 6.

[11] So-Young Park, Member, IEEE, Jeunghyun Byun, Hae-Chang Rim, Do-Gil Lee, and Heuiseok Lim(2010)," Natural Language-based User Interface for Mobile Devices with Limited Resources," *IEEE Transactions on Consumer Electronics*, Vol. 56, No. 4.

[12] Karin Verspoor, Christophe Roeder, Helen L. Johnson, K. Bretonnel Cohen, William A. Baumgartner Jr., and Lawrence E. Hunter(2010)," Exploring Species-Based Strategies for Gene Normalization," *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, VOL. 7, NO. 3.

[13] Richard Khoury, Fakhreddine Karray and Mohamed S. Kamel(2008)," Domain Representation Using Possibility Theory: An Exploratory Study," *IEEE Transactions on Fuzzy Systems*, VOL. 16, NO. 6.

[14] Jen-Tzung Chien and Meng-Sung Wu(2008)," Adaptive Bayesian Latent Semantic Analysis," *IEEE Transactions On Audio, Speech, And Language Processing*, VOL. 16, NO. 1.

K. Dhanasekaran received B.E degree in Computer Science from Mahendra Engineering College affiliated to Anna University Chennai in 2006.He received M.E degree in computer science from K.S.R College of Engineering affiliated to the Anna University of Technology Coimbatore in June 2009.Currently he is a research scholar in Info Institute of Engineering affiliated to Anna University of Technology,Coimbatore,Tamilnadu,India.His current research interests include semantic Web, ontologies, machine learning,Data mining, semantic Web services, and information search and retrieval. He is a member of the MISTE.

R. Rajeswari received B.E degree from Thiagarayar Engineering College in 1995.She received M.E degree from Thiagarayar Engineering College in 1998.She received her Ph.D in Power System Engineering from Anna University, Chennai, India in 2009.She is currently an Assistant Professor of Electrical Engineering Department at Government College of Technology, Coimbatore, India. Her research areas include Power System Engineering, Power System Protection. She is a member of ISTE.

