PART-OF-SPEECH ENHANCED CONTEXT RECOGNITION

Rasmus Elsborg Madsen, Jan Larsen and Lars Kai Hansen
Department of Mathematical Modeling, Building 321
Technical University of Denmark, DK-2800 Lyngby, Denmark
Phone: +45 4525 3894
Fax: +45 4587 2599
E-mail: rem,jl,lkh@imm.dtu.dk
Web: isp.imm.dtu.dk

Abstract. Language independent 'bag-of-words' representations are surprisingly effective for text classification. In this communication our aim is to elucidate the synergy between language independent features and simple language model features. We consider term tag features estimated by a so-called part-of-speech tagger. The feature sets are combined in an early binding design with an optimized binding coefficient that allows weighting of the relative variance contributions of the participating feature sets. With the combined features documents are classified using a latent semantic indexing representation and a probabilistic neural network classifier. Three medium size data-sets are analyzed and we find consistent synergy between the term and natural language features in all three sets for a range of training set sizes. The most significant enhancement is found for small text databases where high recognition rates are possible.

INTRODUCTION

The World Wide Web is a huge, unstructured, and fast growing database, but web users are often left in frustration by the low precision and recall of today's search tools [6]. It is widely believed that machine learning techniques will play an important role in creating more efficient searches. Ambitious plans have been launched for supporting intelligent use of the web, i.e., a "semantic web" [4]. IBM's WebFountain [5] and the Stanford University semantic web platform TAP [13] are examples of machine learning methods coming into play, making human web navigation easier. Here we consider web content mining in the form of internet document classification - an information retrieval (IR) aspect of web-mining [18]. Internet documents contain text, hyper-links, meta-data, images, and other multimedia content which can be used for classification [18, 17]. This paper focuses on classification based on text part, i.e., text categorization. Text categorization is the process of creating a supervised automatic text classifier, by means of machine learning techniques. The classifier labels documents from the corpus $\mathcal{D} = [d_1, \cdot, d_j, \cdot, d_{|\mathcal{D}|}]$ into a set of classes $\mathcal{C} = [c_1, \cdot, c_k, \cdot, c_{|\mathcal{C}|}]$, based on an initial set of labeled documents.

Generic text categorization systems are based on the bag-of-words representation, which is surprisingly effective for the task. In the bag-of-words representation we summarize documents by their term histograms. The main motivation for this reduction (removing the semantics) is that it is easily automated and needs minimal user intervention beyond filtering of the term list. The term list typically contains in the range of $10^3 - 10^5$ terms, hence further reduction is necessary for most pattern recognition devices. Latent semantic indexing (LSI) [12, 11] aka principal component analysis is often used to construct low dimensional representations. LSI is furthermore believed to reduce synonymy and polysemy problems [11, 19]. Synonymy is when multiple words have the same meaning and polysemy is when a single word have multiple meanings. Although LSI and other more elaborate vector space models have been successful in text classification in small and medium size databases, see e.g., [17, 14], it is still not at human level text classification performance. When training classifiers on relatively small databases generalizability is a key issue. How well does a model adapted on one set of data predict the labels of another test data set? Generalizability is in general a function of the number of training cases and of the effective model dimension.

In this communication our aim is to understand the role of natural language features for classification. Specifically, we are interested in the role of term characteristics as derived by natural language processing (NLP). We have chosen the so-called QTAG [20] part-of-speech (POS)-tagger to estimate term characteristics. Synergy of bag-of-words features and POS-features will be evaluated by the effects their combination has on document classification rates. We will use 'early binding' combining the feature sets prior to LSI projection.

NLP features have been used for document classification in a number of studies. In the so-called WordNet system [9] synonymy features were used to expand term-lists for each text category. This strategy enhanced the accuracy of the text classifier significantly. Limited improvements were obtained by invoking semantic features from WordNet's lexical database [15]. In [3] and [2] enhanced classification ability was reported by the use of POS-tagged terms to avoid the confusion from polysemy. In [1] a POS-tagger was used to extract more than $3.0 \cdot 10^6$ compound terms in a database. A classifier based on the extended term list showed improved classification rates.

METHODS

The documents are arranged in a term document matrix $\mathbf{X}$, where $X_{i,j}$ is the number of times term i occur in document j. The dimensionality of $\mathbf{X}$ is reduced by filtering and stemming. Stemming refers to a process in which words with different endings are merged, e.g., 'trained' and 'training' are merged into the common stem 'train'. About 500 common non-discriminative stop-words, i.e. (a, i, and, an, as, at), are removed by filtering. In addition high and low frequency words are also removed from the term list. The term-document matrix can be normalized in various ways. In [10] experiments with different term weighting schemes are carried out. The term frequency / inverse document frequency (TFIDF) weighting is consistently good among term weighting methods purposed, and is the method generally used. After TFIDF normalization the resulting elements in $\mathbf{X}$ becomes

$$X_{i,j}^{\text{tfidf}} = X_{i,j}^{\text{tf}} \log \frac{|\mathcal{D}|}{DF_i} \qquad (1)$$

where DF_i is the document frequency of term i and $X_{i,j}^{\text{tf}}$ is the log normalized term frequency.

$$X_{i,j}^{\text{tf}} = \begin{cases} 1 + \log(X_{i,j}) & \text{if} X_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

The length of the documents is often a good prior for predicting the content within a little corpora. While document length might be a solid variable within the corpora, it is likely that this is not generally a valid parameter. The length of the documents is usually normalized to prevent the influence the document length might have. The Frobenius norm is used to length normalize the term document matrix to one.

$$X_{i,j}^{\text{n2tfidf}} = \frac{X_{i,j}^{\text{tfidf}}}{\sqrt{|\mathcal{T}|^{-1} \sum_{i'=1}^{|\mathcal{T}|} X_{i',j}^{\text{tfidf}^2}}} \qquad (3)$$

We use POS-tags in a design similar to the bag-of-words representation. A tag-document matrix $\mathbf{Y}$ is generated, where Y_{gj} is the number of times tag g occur in document j. The POS-tagger analyzes all sentences in the documents and words part-of-speech function is determined, i.e. noun, verb, adverb, number, punctuation, etc. The POS-tagger distinguishes between 90 different tags. The tagging accuracy of QTAG is approximately 97% [22]. The tag document matrix is normalized as the term document matrix.

Feature set combination is often referred to as 'binding' in analogy with the ability of human brain to bind multiple features for enhanced pattern recognition. Binding can be achieved at different levels. In 'early binding' features are combined in the pre-processing steps. Early binding of feature sets with different statistics and based on variance decomposition requires determination of the relative weights of the participating feature sets. One possibility would be to use variance decomposition based on factor analysis which is insensitive to relative scaling of variables. For simplicity, we have chosen to introduce a single binding coefficient α which can be tuned for each corpus separately,

$$\mathbf{Z} = \left[\begin{array}{c} \alpha\mathbf{X} \\ (1-\alpha)\mathbf{Y} \end{array} \right] \tag{4}$$

If $\alpha \approx 0$ variance is dominated by tag features while when $\alpha \approx 1$ term features dominate.

The combined matrix $\mathbf{Z}$ is reduced to a feature-document matrix using LSI. The reduced dimension features are found by projecting the matrix $\mathbf{Z}$, onto a set of orthogonal basis vectors found by singular value decomposition $\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathbf{T}}$. A wide variety of classification algorithms have been applied to the text mining problem, see e.g., [18]. We have extensive experience with probabilistic neural network classifiers and a well tested ANN toolbox is available[24]. The toolbox adapts the network weights and tunes complexity by adaptive regularization using the Bayesian ML-II framework, hence, requires minimal user intervention [25]. According to [23], this baseline method is among the best for text classification.

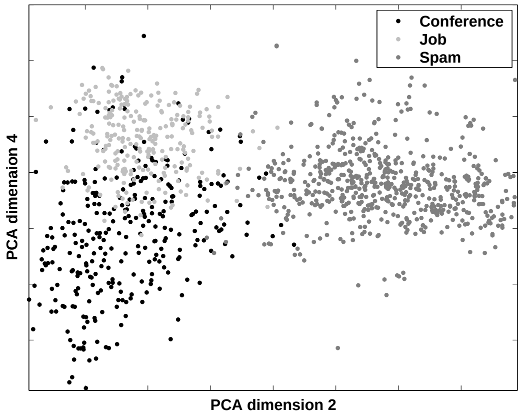

Figure 1: Illustration of the document distribution in feature space. Here we show the email corpus projected onto the 2nd and 4th principal directions. In this projection the ‘spam’ class is well separated while the two other classes in the set (‘conferences’ and ’jobs’) show some overlap.

DATA

We measure the synergy of term and POS features in three corpora: Email [21], Multimedia [16] [17] and webKB [7]. The data-sets have been split into training and test sets and have been re-sampled for statistical verification of the results. 10 splits were used in all experiments. The email set consists of texts from 1431 emails manually classified in three categories: Conference (370), job (272) and spam (789). The multimedia corpus consists of texts and images from 1200 web pages. Only the text part is considered here. The categories are: Sports (400), aviation (400) and paintball (400). The WebKB contains 8282 web-pages from various universities computer science departments. We use a subset of the corpus extracted in [8], and used in [14] and [19], containing 2240 pages. The categories are: Project (353), faculty (483), course (553) and student (851). All 'html' tags were removed from the corpus in this investigation. The multimedia data has a relatively small vocabulary with only 3500 terms after preprocessing. The email data has 9500 terms, and the WebKB data has 13000 terms after preprocessing. The POS-tag features represent a space which is smaller than the term space by a factor of 40-140 for the three data-sets.

RESULTS

Preliminary experiments indicated that a reduced feature space of $K = 48$ projections and a neural network classifier with five hidden units were sufficient for the task. These parameters have been estimated, using cross-validation re-sampling of the training data, see e.g. [26] (data not shown). The complexity of the combined system is optimized by adaptive regularization ('weight decay') for each corpus separately by the neural network training procedure which is based on Bayesian ML-II methods [24].

In previous studies on the three corpora it has been shown that the email and multimedia data set are relatively well classified with term features alone, while the WebKB data set is relatively hard to classify. In figure 1 we show a 2D projection of the email set indicating that the classes are indeed well separated.

We performed three types of experiments. Using the POS-tags alone, using the terms alone and using the combined feature set. We split the corpora in 20% for training and 80% for testing (the role of the split ratio is discussed below). The POS-tags features alone (i.e., using only the relative frequencies of word category) are surprisingly potent: We found that 89.7% of the multimedia data-set is classified correctly using 90 POS-tag features. This should be compared to 96.6% classification accuracy obtained with the almost 3500 term features. For the email data, using the POS-tag and term features separately resulted in accuracies of 74.6% and 94.2% respectively. The WebKB data is somewhat harder to classify. Here the POS-tag and term features lead to accuracies of 57.2% and 76.1% respectively.

The potential synergy of terms and POS-tags is illustrated in figure 2. The figure shows the performance correlation between the classifiers trained on the individual feature sets. The bars labelled 'independent' indicate the rates of events where the two classifiers are both correct as well as events where one is correct and one is incorrect obtained from their basic performance and assuming independence of their decisions. In bars labelled 'real' we show the actually observed rates. Note that there is a high potential synergy, since the observed performances are close to those predicted by independence. We

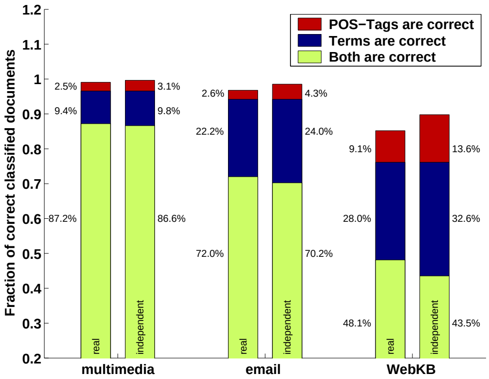

Figure 2: Fraction of correct classified documents for the POS-tag and term representations. The bars labelled 'real' indicate observed rates of events where the two feature sets lead to correct decision and one correct/one incorrect respectively. This is compared with rates estimated from assumed independence of errors (bars labelled 'independent'). The figure indicates that the errors made by classifiers based on POS-tags and the term features sets are relatively independent, hence, that there is a potential synergy to be gained from binding the feature sets.

next turned to the combined feature set. In figure 3 we illustrate the role of the binding coefficient α, c.f., (4). The classification test set error rates (an unbiased estimate of the generalization error defined as the probability of misclassification of a random test datum) were obtained by ten-fold cross-validation. We observed significant synergy: The performance of the term features ($\alpha = 1$) is indeed improved by adding POS-tag feature information. The effect is relatively high for the multimedia data-set (reducing the error by almost 30%), while the effect is smaller for the harder WebKB set (the error is reduced by about 8%). The synergistic advantage is likely to depend on the size of the database, to further investigate this we estimated 'learning curves' for the the combined system by changing the split ratio allowing for variable training set sizes. The results are provided in figure 4. In these ten-fold cross-validation experiments we used the 'optimal' binding coefficients found in figure 3. In these relatively limited data sets there is a positive, albeit diminishing, synergy to be obtained for all training set sizes.

CONCLUSION

Natural language features in the form of part-of-speech (POS) tags were introduced to supplement bag-of-words features. We propose simple statistical POS-tag features: The frequency of different term types. By early binding of POS-tags and term features we find a synergistic effect for a range of binding coefficients and for all training sets sizes studied. The results were consistent for three different corpora posing variable classification difficulties. As the POS-tag features are relatively automatic and computationally 'inex-

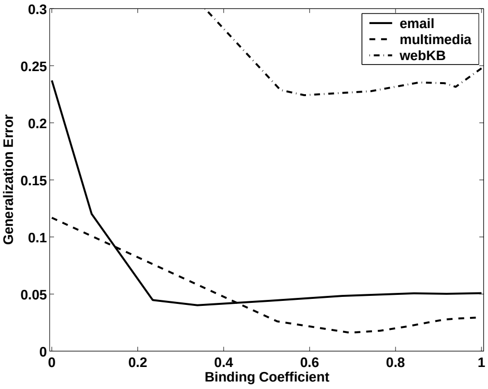

Figure 3: Misclassification error obtained by binding POS-tags and Terms with variable a binding coefficient. $\alpha = \mathbf{1}$ corresponds to tag features only. Optimal binding results in reduction of the error rate by 30%, 22% and 8% in the email, multimedia and WebKB corpora respectively Results obtained by ten-fold cross-validation using a 20/80 train/test set split ratio.

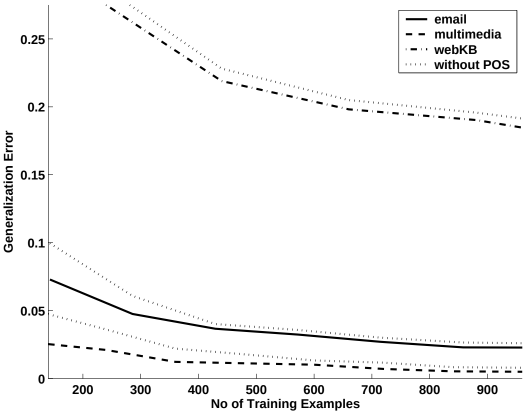

Figure 4: Learning curves with and without binding. The binding of natural language and conventional term features improves performance for all the training set sizes investigated.

pensive' to estimate we recommend that these feature be included in future text/contex classification applications.

REFERENCES

[1] A. Aizawa, "Linguistic Techniques to Improve the Performance of Automatic Text Categorization," in **Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium**, Tokyo, JP, 2001, pp. 307–314.

[2] R. Basili and A. Moschitti, "A robust model for intelligent text classification,"

in **Proceedings of ICTAI-01, 13th IEEE International Conference on Tools with Artificial Intelligence**, Dallas, US: IEEE Computer Society Press, Los Alamitos, US, 2001, pp. 265–272.

[3] R. Basili, A. Moschitti and M. Pazienza, “NLP-driven IR: Evaluating Performances over a Text Classification task,” in B. Nebel (ed.), **Proceeding of IJCAI-01, 17th International Joint Conference on Artificial Intelligence**, Seattle, US, 2001, pp. 1286–1291.

[4] T. Berners-Lee, J. Hendler and O. Lassila, “The Semantic Web,” **Scientific American**, 2001.

[5] I. A. R. Center, “The WebFountain,” http://www.almaden.ibm.com /webfountain/publications/.

[6] S. Chakrabarti, “Data mining for hypertext: a tutorial survey,” **SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM**, vol. 1, pp. 1–11, 2000.

[7] CMU-WebKB, “The 4 Universities Data Set,” http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/, 1997.

[8] CMU-WebKB-2240, “A subset of the WebKB,” http://www.imm.dtu.dk/∼rem/, 1999.

[9] M. De Buenaga Rodríguez, J. M. Gómez-Hidalgo and B. Díaz-Agudo, “Using WordNet to Complement Training Information in Text Categorization,” in R. Milkov, N. Nicolov and N. Nikolov (eds.), **Proceedings of RANLP-97, 2nd International Conference on Recent Advances in Natural Language Processing**, Tzigov Chark, BL, 1997.

[10] F. Debole and F. Sebastiani, “Supervised term weighting for automated text categorization,” in **Proceedings of SAC-03, 18th ACM Symposium on Applied Computing**, Melbourne, US: ACM Press, New York, US, 2003, pp. 784–788.

[11] S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman, “Indexing by Latent Semantic Analysis,” **Journal of the American Society of Information Science**, vol. 41, pp. 391–407, 1990.

[12] G. Furnas, S. Deerwester, S. Dumais, T. Landauer, R. Harshman, L. Streeter and K. Lochbaum, “Information retrieval using a singular value decomposition model of latent semantic structure,” in **The 11th International Conference on Research and Development in Information Retrieval**, Grenoble, France: ACM Press, 1988, pp. 465–480.

[13] R. Guha and R. McCool, “TAP: A Semantic Web Platform,” **Computer Networks: The International Journal of Computer and Telecommunications Networking**, vol. 42, pp. 557–577, 2003.

[14] L. Hansen, S. Sigurdsson, T. Kolenda, F. Nielsen, U. Kjems and J. Larsen, “Modeling text with generalizable Gaussian mixtures,” in **International Conference on Acoustics, Speech and Signal Processing**, IEEE, 2000, pp. 3494–3497.

[15] A. Kehagias, V. Petridis, V. G. Kaburlasos and P. Fragkou, “A Comparison of Word- and Sense-based Text Categorization Using Several Classification Algorithms,” **Journal of Intelligent Information Systems**, vol. 21, no. 3, pp. 227–247, 2003.

[16] T. Kolenda, “Multimedia Dataset,” http://mole.imm.dtu.dk/faq/MMdata/, 2002.

[17] T. Kolenda, L. Hansen, J. Larsen and O. Winther, "Independent component analysis for understanding multimedia content," in S. B. J. L. H. Bourlard, T. Adali and S. Douglas (eds.), **Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII**, Piscataway, New Jersey: IEEE Press, 2002, pp. 757–766.

[18] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," in **SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM**, ACM Press, 2000, pp. 1–15.

[19] J. Larsen, L. Hansen, A. Have, T. Christiansen and T. Kolenda, "Webmining: learning from the world wide web," **Computational Statistics and Data Analysis**, vol. 38, pp. 517–532, 2002.

[20] O. Mason, "Probabilistic Part-of-Speech Tagger," http://web.bham.ac.uk/o.mason/software/tagger/, 2003.

[21] F. Nielsen, "Email Data-Set," http://www.imm.dtu.dk/∼rem/, 2001.

[22] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees." in **International Conference on New Methods in Language Processing**, Manchester, UK, 1994.

[23] F. Sebastiani, "Machine learning in automated text categorization," **ACM Computing Surveys**, vol. 34, pp. 1–47, 2002.

[24] S. Sigurdsson, "The DTU: Artificial Neural Network Toolbox," http://mole.imm.dtu.dk/toolbox/ann/, 2002.

[25] S. Sigurdsson, J. Larsen and L. Hansen, "On Comparison of Adaptive Regularization Methods," in B. Widrow, L. Guan, K. Paliwa, T. Adali, J. Larsen, E. Wilson and S. Douglas (eds.), **Proceedings of the IEEE Workshop on Neural Networks for Signal Processing**, 2000, pp. 221–230.

[26] S. Strother, J. Anderson, L. Hansen, U. Kjems, R. Kustra, J. Siditis, S. Frutiger, S. Muley, S. LaConte and D. Rottenberg, "The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework," **Neuroimage**, vol. 15, pp. 747–771, 2002.