


Pumping Documents Through a Domain and Genre Classification Pipeline

Udo Hahn Joachim Wermter

 Text Knowledge Engineering Lab
Freiburg University
Werthmannplatz 1
D-79098 Freiburg, Germany
<http://www.coling.uni-freiburg.de/>

Abstract

We propose a simple, yet effective, pipeline architecture for document classification. The task we intend to solve is to classify large and content-wise heterogeneous document streams on a layered nine-category system, which distinguishes medical from non-medical texts and sorts medical texts into various subgenres. While the document classification problem is often dealt with using computationally powerful and, hence, costly classifiers (e.g., Bayesian ones), we have gathered empirical evidence that a much simpler approach based on n -gram-statistics achieves a comparable level of classification performance.

1. Introduction

The task of text classification is to assign documents to a set of pre-defined subject categories. The form and sophistication of such a category system may vary considerably. There are highly differentiated and comprehensive classification codes, often developed by human experts over many years, such as the disease classification ICD 10 (ICD-10, 1992) in medicine. The distinction between text genres (e.g., fiction vs. newspaper articles) or broad topic areas within a text genre (e.g., newspaper articles dealing with politics or economy) are examples of higher-level, less sophisticated category systems.

The problem we want to solve is to distinguish, from a large, continuous and content-wise heterogeneous text stream, medical from non-medical documents, and to assign the medical documents to various subgenres (e.g., clinical, pathology, surgery reports) for subsequent in-depth processing (Hahn et al., 2002). This kind of pre-processing prevents our system from analyzing a priori irrelevant documents, *viz.* the non-medical ones, and, in addition, helps to identify those domain partitions of the background knowledge (Schulz and Hahn, 2001) which are required for different genres. With these precautions, we aim at keeping the processing load for our system within feasible regions.

As a research topic, the whole battery of information retrieval and machine learning methodologies have already been employed to deal with the problem of text classification. In particular, this includes ‘power’ approaches such as support vector machines, linear discriminant techniques, k -nearest-neighbor and Bayesian classifiers, rule induction methods, etc. (cf. the surveys by (Yang, 1999) and (Sebastiani, 2002)). From a system engineering point of view, we considered them computationally too powerful and too expensive for our pre-processing task. We rather focus here on a much simpler, more parsimonious approach based on character n -grams that, nevertheless, should perform at an equal level of classification accuracy. Moving to computationally inexpensive methods also seems inevitable in a framework where documents are extracted from the Web and have to be processed rapidly in large numbers.

The approach we use has originally been tested on language identification and topic discrimination problems using articles from some of the Usenet newsgroups which deal with information technology issues such as graphics, security, artificial intelligence (Cavnar and Trenkle, 1994). Since the results in that domain were promising, we wanted to test the feasibility of this approach in the domain of medicine as well.

2. Methodology

For our experiments, we used the TEXTCAT system.¹ This is off-the-shelf software, which implements the n -gram character based approach to text classification described by (Cavnar and Trenkle, 1994). All the parameters of the tool were left ‘as is’. The system requires various-sized adequate samples for each document category. For Cavnar and Trenkle’s newsgroup text classification task, the sample size for the different categories ranges from 21KB to 132KB. In our case, we decided to decrease the size along the line of increasing category specificity. Hence, for coarse-grained categories (medical vs. non-medical) we chose much larger sample sizes than for fine-grained ones, e.g., surgery vs. pathology reports (cf. Table 3).

The methodology on which this document classifier is based is fairly simple. From a set of pre-existing text categories a set of n -gram frequency profiles is generated ($n=1..5$) to represent each category, the *category profiles*. Each n -gram set is ranked by decreasing frequencies and cut off at rank 400. For each incoming document to be classified, its n -gram frequency profile, the *document profile*, is computed in the same way. Table 1 shows position 350 through 364 of the sorted n -gram frequency ranking² for some of our text category profiles (for a description of the category system, cf. Section 3.).

¹TEXTCAT, a Perl-based program, can be downloaded from <http://odur.let.rug.nl/~vannoord/TextCat/>

²The top ranked n -grams are mostly unigrams and mirror the alphabetic letter distribution of a natural language. They are more relevant for language identification, whereas lower ranked n -grams are more indicative of the text categories to be classified.

Category Profiles					Document Profile
non-med	med	non-clin	clin	histo	histo
ner	ellen	V)	mp	men
ts	i_	L	_(:	re	ophi
nu	_p	_auf	se_	_des	ei_
nte	ber	ko	ellen	lei	mit_
eu	_l	R	atio	nu	topoe
rei	mp	ran	ation	pl	_Er
von	wei	ige	chen_	n,	tze
du	ku	nn_	e.	de_	mie
_is	_r	tu	ek	_Ma	oph
sie	tisc	urc	e_	n_	mit
_ist	tisch	ss	her	och	hie
ges	ebe	_f	_in_	_des_	tisc
_auf	nn	ann	eren	tio	_Ma
von_	zell	urch	tra	ing	wird
dem_	aus	_V	nz	ellen	mozyt

Table 1: Category and Document Profiles from Ranks 350 to 364

The classification decision is based on a fairly simple two-step procedure, which is applied to the category and document profiles.

1. *Measuring the profile distance:* The two profiles are taken and a simple rank-order statistics is computed. It measures how far ‘out of place’ (Cavnar and Trenkle, 1994) a particular n -gram in the document profile is from its place in the category profile.³ For example, for the n -gram *tisc* the distance value is ‘2’ between the histology document profile and the medical category profile. The sum of all out-of-place values for all n -grams is the document’s distance from the category.
2. *Finding the minimum distance:* The profile distances from the document profile to all the category profiles are computed, and the document is assigned to that category to which it has the smallest overall distance.

3. Experiments

In our research, the need to classify Web documents not only arose with respect to the natural language in which they were written (currently, we focus on selecting and distinguishing English, German and Portuguese documents) but, most importantly, to which domain and domain-specific text genre they belong. For this purpose, we established a classification pipeline with three layers of increasing specificity (cf. Figure 1). Our training set (cf. Table 3) and test set (cf. Table 2) is composed of German-language documents. The categorization decisions can be described as answers to the following questions:

- *Is it a medical document or not?*

Our test set contains a total of 270 medical documents of various categories (see below), as well as a total of 232 online newswire texts from four different categories (100 from politics, 60 from economics, 42 from sports, and 30 from culture/society).

³In case an n -gram does not occur in the category profile, a maximum value is chosen.

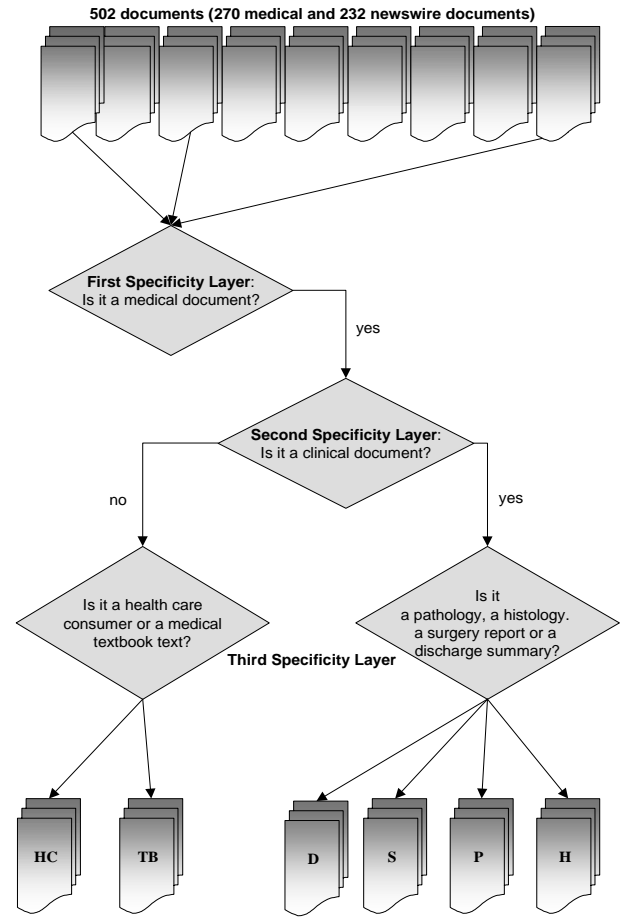


Figure 1: Classification Pipeline

- *For medical documents only, is it a clinical document or not?*
Our test set contains 187 clinical documents (the pathology, histology and surgery reports) and 83 non-clinical ones (the textbook and health care consumer texts).
- *For clinical documents only, is it a pathology, histology or surgery report?*
The test set contains 55 pathology reports, 55 histology reports and 77 surgery reports, all clinical documents.
- *For non-clinical documents only, is it a health care consumer or a medical textbook text?*
The test set contains 48 textbook and 35 health care consumer texts, all non-clinical documents.

Domain	Subdomain	Subsubdomain	# Articles
medical	clinical	pathology	55
		histology	55
		surgery	77
	non-clinical	textbook	48
	consumer text	35	
		$\Sigma = 270$	
non-medical	newswire	different areas	232

Table 2: Text Samples – The Test Set

Best Match	Non-medical	Medical Texts								
Language Model	newspaper & newswire	Clinical Texts			Non-clinical Texts			all non clinical	all medical	
		pathology	histology	surgery	all clinical	textbook	consumer			
medical	37 (15.9%)	54 (98.2%)	55 (100%)	77 (100%)	186 (99.5%)	40 (83.3%)	27 (77.1%)	67 (80.7%)	253 (93.7%)	
non-medical	195 (84.1%)	1 (1.8%)	0 (0%)	0 (0%)	1 (0.5%)	8 (16.7%)	8 (22.9%)	16 (19.3%)	17 (6.3%)	
clinical		45 (81.8%)	55 (100%)	76 (98.7%)	176 (94.1%)	2 (4.2%)	0 (0%)	2 (2.4%)		
non-clinical		10 (18.1%)	0 (0%)	1 (1.3%)	11 (5.9%)	46 (95.8%)	35 (100%)	81 (97.6%)		
pathology		54 (98.2%)	11 (20%)	1 (1.3%)						
histology		0 (0%)	43 (78.2%)	0 (0%)						
surgery		1 (1.8%)	1 (1.8%)	76 (98.7%)						
textbook						36 (75.0%)	10 (28.6%)			
consumer						12 (25.0%)	25 (71.4%)			
Total	232	55	55	77	187	48	35	83	270	

Table 4: Classification Results

(Sub)Domains	Size	Contains/Covers
medical	22MB	all medical (sub)domains
non-medical	19MB	newspaper & newswire texts
clinical	235KB	all clinical subdomains
pathology	74KB	pathology reports
histology	78KB	histology reports
surgery	160KB	surgery reports
non-clinical	266KB	all non-clinical subdomains
textbook	104KB	medical expert texts
consumer texts	161KB	web health portal texts

Table 3: The Training Set

In our experiments, we did neither sort out those documents which were incorrectly classified, nor propagate misclassified documents up or down the classification hierarchy. Rather, the goal of our classification experiment was to evaluate each of the three specificity layers of our pipeline. For this purpose, we judged all the available candidate documents at a given decision point as to whether they were correctly classified or not.

The training set (cf. Table 3) shows that we have an almost even distribution of medical and non-medical documents (around 20MB). In the clinical document set, surgery reports (160KB) have almost double the size of the set of pathology (74KB) and histology (78KB) reports. The samples for textbook articles and health care consumer texts come with 104KB and 161KB, respectively. Future work will have to further balance the size of all samples for all subdomains involved.

4. Results and Discussion

Table 4 summarizes the results of our classification experiment. Out of all medical documents, almost 94% (253 out of 270) were correctly classified as medical; out of all newswire documents, 84% were classified as non-medical. Thus, the overall results for the most coarse-grained specificity layer of our classification pipeline seem solid.

All but one out of 187 clinical documents were classified correctly (99.5%). Fewer non-clinical documents were correctly classified as medical documents (80.7%, viz. 67 out of 83). This data indicates that the categorization of non-clinical textbook and consumer text material marks the borderline case from science or jargon to (more) popular, in the sense of non-medical, writing. This is further evidenced

by looking at the individual non-clinical document types: 83% (40 out of 48) of the textbook texts were correctly sent down the medical drain of our classification pipeline, whereas only 77% of the health care consumer documents were done so.

With respect to the individual clinical document types, 98% (54 out of 55) of the pathology reports were correctly classified as medical, as well as 100% (all 55) of the histology reports, and 100% (all 77) of the surgery reports. We attribute this to the articulate form of ‘medical writing’ one finds in these clinical documents.

The second, more fine-grained decision point in our classification pipeline – whether a medical document is clinical or not – also shows good results, with 94.1% (176 out of 187) of the clinical and 97.6% (81 out of 83) of the non-clinical documents being sent down the correct drain.

On the third and most specific decision layer, the documents are pumped into their corresponding document type category. The data shows that the results for clinical texts are better than those for non-clinical ones. 98.7% of all surgery reports and 98.2% of all pathology reports were assigned the correct classification tag. Thus, compared to all other subdomains, the surgery reports show the best overall performance in terms of classification accuracy. The results for the histology reports (78%) are considerably worse, with more than 20% being falsely assigned to the pathology category. Given the similarity of the contents between the two text categories, this, at first sight, may not be surprising. Conversely, however, none of the pathology reports were falsely put in the histology category. At this point, we have no explanation for this rather puzzling state of affairs.

As for non-clinical documents, more textbook (over 75%) than consumer texts (71.4%) were correctly assigned to their corresponding category. Still, the numbers do not at all match those of the clinical texts. This suggests that, in general, these two text types are far more distant from the clinical text genres than one might expect. Overall, we achieve an average score of 90.6% for our 9-category classification system.

5. Related Work

Up until now, almost no effort has been made on medical text genre classification. In terms of message routing, the state-of-the-art performance figures for medical

text categorization are set in a recent study of (Kornai and Richards, 2002). They achieve 86% correct classifications of clinical verbatims on a 12-category system, using linear discriminant analysis, one of the already mentioned ‘power’ methods. However, their classification system consists of standardized and sophisticated medical phenomena categories, and is thus not comparable to a higher-level text genre system like ours.

(Karlgrén and Cutting, 1994), who use the same computational approach, report on experiments run on the mixed-domain Brown corpus. For a 2-category classification scheme (informative vs. imaginative texts), they come up with 96% and 95% accuracy, respectively. For a 4-category classification (which they also blow up to 10/15-categories) involving press, fiction, non-fiction, and miscellaneous texts, they achieve 83%, 95%, 75% and 53% accuracy, respectively. Their data is hard to compare to ours, because not even a category like ‘science’ is mentioned. The parameters they feed the discriminant analysis with are really diverse and (using terminology introduced by (Kessler et al., 1997) range from *lexical* cues involving specific word counts (e.g., ‘therefore’, ‘that’, ‘which’), *structural* cues involving part of speech counts (e.g., adverbs, nouns, pronouns) to *formal pattern* cues such as counts of long words (exceeding six characters), as well as *derivative* cues such as type/token ratios, word length and sentence length averages. Their observation, however, that the ‘error rates climb steeply with the number of categories’ (Karlgrén and Cutting, 1994, p.1073) does certainly not hold for our test set.

(Kessler et al., 1997) employ a 13-category classification system, one category of which is scientific or technical writings (scitech). Their basic numerical methods are logistic regression and 2- as well as 3-layer perceptrons, which achieve a classification accuracy ranging between 93-100% for the scitech category (depending on the parametrization) given, however, a stunning baseline of 94%.

A recent study by (Lee and Myaeng, 2002) on text genre classification is more comparable to ours. On a 7-category system and a 6-category system, they achieve 87% and 90% correct assignments, respectively. Although their results match ours, their statistical classifier (term and document frequency as well as naive Bayesian statistics and a cosine-based similarity measure) is still computationally more expensive than ours, and hence would increase the computational load of our pre-processing task.

(Stamatatos et al., 2000) compute genre categorization on a 4-category system by comparing absolute word frequencies in general language (derived from the BNC) with those from four genre-specific corpora built from the *Wall Street Journal*. They also use linear discriminant analysis for genre identification and achieve around 97% accuracy.

6. Conclusion

We proposed a simple, yet effective, pipeline architecture for document classification. Our system classifies large, continuous and content-wise heterogeneous document streams on a 9-category system, distinguishing medical from non-medical texts and sorting medical texts to various subgenres. While the document classification problem has already been dealt with using computationally expen-

sive classifiers (SVMs, Bayesian models, etc.), we gathered empirical evidence that a conceptually and computationally much simpler approach based on n -gram statistics (Cavnar and Trenkle, 1994) achieves a comparable level of classification performance. Rather than considering different dimensions of parameters (e.g., structural, lexical, and derived ones) we just rely on formal parameters, viz. the n -gram structure ($n=1..5$) of documents to be classified. We intentionally chose such a simple approach, because it is sufficient for our pre-processing purposes. Unlike the ‘power’ methods which devote all computational efforts to this step, we save power for the subsequent stages of ‘real’ text analysis, viz. parsing, semantic and conceptual interpretation (Hahn et al., 2002).

Acknowledgements. This work was supported by Deutsche Forschungsgemeinschaft (DFG), grant KL 640/5-1, and by the Faculty of Medicine at Freiburg University, grant KLA231/03.

7. References

- Cavnar, W. and J. Trenkle, 1994. N-gram-based text categorization. In *SDAIR’94 – Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*. pages 161–175.
- Hahn, U., M. Romacker, and S. Schulz, 2002. MEDSYNDIKATE: A natural language system for the extraction of medical information from finding reports. *International Journal of Medical Informatics*, 67(1/3):63–74.
- ICD-10, 1992. *International Statistical Classification of Diseases and Health Related Problems. 10th Revision*. Geneva: World Health Organization.
- Karlgrén, J. and D. Cutting, 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING’94 – Proceedings 15th International Conference on Computational Linguistics*. pages 1071–1075.
- Kessler, B., G. Nunberg, and H. Schütze, 1997. Automatic detection of text genre. In *ACL’97/EACL’97 – Proceedings 35th Annual Meeting of the ACL & 8th Conference of the European Chapter of the ACL*. pages 32–38.
- Kornai, A. and J. Richards, 2002. Linear discriminant text classification in high dimension. In *Hybrid Information Systems. Proceedings of the 1st International Workshop on Hybrid Intelligent Systems*. pages 527–537.
- Lee, Y.-B. and S. Myaeng, 2002. Text genre classification with genre-revealing and subject-revealing features. In *SIGIR 2002 – Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pages 145–150.
- Schulz, S. and U. Hahn, 2001. Medical knowledge reengineering – converting major portions of the UMLS into a terminological knowledge base. *International Journal of Medical Informatics*, 64(2/3):207–221.
- Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.
- Stamatatos, E., N. Fakotakis, and G. Kokkinakis, 2000. Text genre detection using common word frequencies. In *COLING 2000 – Proceedings 18th International Conference on Computational Linguistics*. pages 808–814.
- Yang, Y., 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90.