An Experiment in Automatic Classification of Pathological Reports

Janneke van der Zwaan, Erik Tjong Kim Sang, and Maarten de Rijke

ISLA, University of Amsterdam, Amsterdam, The Netherlands
{jvdzwaan,erikt,mdr}@science.uva.nl

Abstract. Medical reports are predominantly written in natural language; as such they are not computer-accessible. A common way to make medical narrative accessible to automated systems is by assigning 'computer-understandable' keywords from a controlled vocabulary. Experts usually perform this task by hand. In this paper, we investigate methods to support or automate this type of medical classification. We report on experiments using the PALGA data set, a collection of 14 million pathological reports, each of which has been classified by a domain expert. We describe methods for automatically categorizing the documents in this data set in an accurate way. In order to evaluate the proposed automatic classification approaches, we compare their output with that of two additional human annotators. While the automatic system performs well in comparison with humans, the inconsistencies within the annotated data constrain the maximum attainable performance.

1 Introduction

Increasing amounts of medical data are stored in electronic form. Medical data contains a lot of information that can be used for many different purposes, such as decision support, epidemiological research, quality control, etc. Medical reports are predominantly written in natural language, and as such they are not computer-accessible. Currently, a common way to make medical narrative accessible to automated systems is by assigning 'computer-understandable' keywords. Experts usually perform this task by hand. In this paper, we investigate methods to support or automate medical classification.

The PALGA foundation is the Dutch national network and registry of histo- and cytopathology (in Dutch: Pathologisch Anatomisch Landelijk Geautomatiseerd Archief, http://www.palga.nl). Since 1971, the PALGA foundation has been maintaining a database with abstracts of all histo- and cytopathological examinations that take place in Dutch hospitals. For every examination, a report is written and the conclusion of the report is sent to the PALGA database. The database can be used to find information about the history of a single patient, but it is also available for research and national health care projects.

A key part of the reports in the PALGA database is the set of diagnosis lines which contain a standardized summary of the report. Diagnosis lines are used for indexing and retrieval of the conclusions. They contain a limited number of

fields (four) with terms from a restricted vocabulary. The diagnosis lines have been created over a long period of time by a large group of doctors and there is some uncertainty about the consistency and the quality of their contents. The size of the database does not permit a thorough and complete manual quality check.

In this paper we investigate the potential of machine learning approaches for automatically generating diagnosis lines from summaries of pathological reports. Similar tasks have, of course, been considered before in the literature; see e.g., [1] for a (somewhat dated) overview. Three things set our setting apart from settings reported on in the literature. First, in our case, the task is to generate diagnosis lines from conclusion texts—*summaries* of reports and often very incomplete. Second, we are working with a very large data set (over 14 million records, out of which we use close to .5 million for our experiments; see below for details), while most studies in the literature are based on far smaller data sets. Third, because of its size, its usage across hospitals all over the Netherlands, and its age, the data set is full of errors and inconsistencies, unlike most data sets used for medical coding in the literature. Against this challenging background, we are interested in finding answers to two research questions. First, what level of accuracy can automatic classification obtain for this task? Specifically, what type of feature representation is most effective? Second, and motivated by the bottlenecks that we ran into while trying to increase the recall scores of automatic classifiers, what performance levels do humans attain when given the exact same task as the automatic classification system?

This paper contains five sections. After this introduction, we describe the classification problem in more detail, outline our learning approach and discuss related work. In Section 3, we present our experiments and their results. Section 4 provides an elaborate comparison between the best automatic learner and two domain experts. We conclude in Section 5.

2 Method

In this section, we describe our data, the machine learning method that was applied to the data, the techniques used for evaluation, as well as related work.

2.1 The Palga Data

The PALGA data set consists of over 14 million reports. It contains all histological examinations that were performed in the Netherlands from 1990 up to and including 2004. A sample record is shown in Figure 1. The reports contain three parts: main, conclusion and diagnosis lines. The terms in the diagnosis lines are restricted to a set of 14,000 terms, each of which is represented by a code. The coding system used by the PALGA foundation is based on an early (1982) version of the Systemized Nomenclature Of MEDicine (SNOMED). Different types of terms exist; in SNOMED these are called *axes*. The first character of a code represents the axis to which the code belongs. For instance, the term *biopt* is

Record ID	39319785
Patient ID	PATIENT-23 m
Date	07 - 1990
Conclusion	Huidexcisie para-orbitaal links: basaalcelcarcinoom van het solide type. Tumorcellen reiken tot in de excisieranden.
Diagnosis line	huid * gelaat * links * excisie * basaalcelcarcinoom

Fig. 1. Example record from the PALGA data set. Pathologists use everyday terms to code diagnoses. These terms are linked to codes in the PALGA coding system.

linked to code P11400, where P indicates a Procedure. The PALGA coding system consists of the axes Topography, Procedure, Morphology, Etiology, Function and Disease. The coding system is ordered hierarchically.

Every code is linked to one or more terms. A thesaurus is available for enabling pathologists to use everyday language rather than codes in the report writing process. Codes are linked to at most one *preferred term*. Terms linked to identical codes are synonyms. For instance, both *colon* (preferred term) and *dikke darm* are linked to code T67000.

Diagnosis lines, the computer-readable summaries of the contents of pathological reports, consist of PALGA terms only. The lines are used for indexing and retrieving PALGA reports. The quality of the retrieval results is obviously dependent on the quality of the diagnosis lines. Detailed guidelines exist for assuring that the lines are accurate. Among others, these guidelines state that diagnosis lines should be complete and that the report conclusion should contain all relevant information for a pathologist to assign correct diagnostic terms [13]. The diagnosis fields contain three compulsory diagnostic fields: the two axes *Topography* and *Procedure*, and *Diagnosis*, which may contain terms of the other four axes. Conclusions are coded with one to four diagnosis lines.

Diagnosis lines in the PALGA database contain a lot of noise. Creating accurate and precise diagnosis lines is not the main task of a pathologist and recommendations of the type-checking tools which became available in recent years, are often ignored. We are interested in working with a data set which was as clean as possible, and therefore we have restrict ourselves to reports containing a conclusion of at least six characters and a single diagnosis line with valid singular terms. Additionally, we restrict the reports to those that contain the term *colon* or one of its descendants. This results in a data set of 477,734 conclusions with associated diagnosis lines, hereafter called the *Colon data set*. Diagnosis lines in the Colon data set contain 3.4 terms on average. The data set was randomly divided into 75% of training data and 25% of test data.

2.2 Support Vector Machines

We decided to use Support Vector Machines (SVMs) for our medical text classification task, as they belong to the best performing learning algorithms currently available [7]. Fast implementations of SVMs exist; we used SVM^{light} [8] (with default settings) for our experiments.

One of the strengths of SVMs is that the standard linear kernel can be replaced by a non-linear one, e.g., a polynomial or radial basic function (RBF). Most of our experiments are conducted with linear kernels, but we also experiment with polynomial ones.

Another property of SVMs is that a classifier is trained independently of the number of features of the data samples. This is particularly useful for text classification, where the dimensionality of the feature space generally is high with few relevant features [7]. Most other machine learning approaches to text classification apply some sort of feature reduction or transformation. By using SVMs there is no need to reduce the number of features.

2.3 Evaluation

We are interested in classifier performance for individual terms as well as prediction accuracy for complete diagnosis lines. The notion of equivalence of diagnosis lines is problematic, because the coding system allows something being said in different ways. Among other things, the level of detail can differ; if a finding is specified using a (slightly) more general (or specific) term, it is not necessarily wrong. Still, we decided to use a simple notion of 'exact' equivalence to assess the agreement between diagnosis lines. Thus, our evaluation results will underestimate the actual performance. For individual term prediction we treat diagnosis lines as a bags of terms and perform evaluation with precision, recall and $F_{\beta=1}$.

2.4 Related Work

Medical coding is the task of assigning one or more keywords to medical text. Reasons to code medical documents include data reduction, standardization, quality control, being able to compare individual cases, and making data available for research. In general, medical coding is considered a difficult and time-consuming task [15, 4, 9, 19]. Ever since the introduction of formal coding systems, attempts have been made to automate the coding process [16, 20, 11]. In [1] a distinction is being made between coding systems that abstract clinical data (such as ICD and MeSH) and those that preserve clinical details. [14] lists a number of information types that medical coding systems can (or should be able to) represent. In [5] manual coding errors in two British hospitals were investigated and compared. It appeared that 'many of the errors seem to be due to laziness in coding, with failure to consult the appropriate manual and reliance on memory for common codes.' A recent study indicates that data quality does indeed improve after the adoption of automatic encoding systems [10].

The PALGA foundation has been involved in an earlier research project regarding medical text classification [2]. In this particular project, complete pathological reports were used to predict appropriate diagnosis lines; 7500 histology reports from two different hospitals were considered and three different document representations were compared. Using the best performing representation—uniform words—a correct diagnosis line could be found within the first five suggestions for 844 of 952 reports. Other experiments showed that a representation

based on words performed better than (character) n-grams with $n > 4$, and that performing training and testing with data from the same site allowed for a better performance than when test data came from another site than the training data. In an additional evaluation, three human experts rating the automatically produced diagnosis lines on a three-point scale, reached an agreement kappa score of only 0.44, which shows that coding pathological reports is not a trivial task.

Recently, Gerard Burger, a pathologist associated with the PALGA foundation, created a term extractor for PALGA conclusions. The program, called AutoDiag, uses domain knowledge and ad hoc rules to propose terms for diagnosis lines associated to the input document. AutoDiag extracts terms from the conclusion part of the documents, keeping terms it considers useful while ignoring other terms. Prior to this paper, AutoDiag had not been properly evaluated. We use it in our work and compare its output with that of our system.

3 Experiments and Results

We treat the task of coding of diagnoses as a text classification task and train binary Support Vector Machines to predict individual diagnostic terms. Below, we describe the experiments that were performed. First, we discuss experiments with different feature representations. After that, we evaluate other variations, changing output class representations, machine learning parameters or data sets. We conclude the section with a discussion.

3.1 Feature Engineering

To create a baseline, we chose SVMs with bag of words (bow) for the feature representation; this is a common representation for text classification. In the bow representation, a document is represented as a feature vector, where each element in the vector indicates the presence or absence of a word in the document. In order to reduce the size of the vector, we discard the least infrequent words (frequency < 2) as well as the most frequent words (so-called stop words; we use a list from the Snowball Porter stemmer for Dutch [18]). The dimensionality of the feature vectors for the *bow* experiment is 21,437. The bow features allowed for a reasonable performance on the Colon test data (section 2.1): precision 83.28%, recall 72.92% and $F_{\beta=1}$ 77.76%.

Since in the baseline results, recall was considerably lower than precision, we focused on improving recall. We evaluated several variations on the bow feature representation to accomplish this:

- replacing binary feature values by *tf-idf* (term frequency-inverse document frequency) weights
- replacing word unigrams by word bigrams (19,641 features)
- adding to the features, terms identified from the conclusion texts (1,286 extra features), and/or their parents according to the thesaurus (1,810)
- reducing the number of features by using stems rather than words (19,098 features) or by splitting compound words (19,006 features)

	Representation	Precis.	Recall	$F_{\beta=1}$		Represent.	Precis.	Recall	$F_{\beta=1}$
a	baseline	83.28%	72.92%	77.76%	j	tf-idf	83.14%	73.71%	78.14%
b	+terms	83.49%	73.16%	77.98%	k	+terms	83.24%	73.94%	78.31%
c	+terms+parents	83.53%	73.22%	78.04%	l	+te+parent	83.35%	74.09%	78.45%
d	terms	79.92%	60.91%	69.13%	m	+te+pa+pr	83.28%	73.69%	78.19%
e	+parents	80.12%	66.41%	72.62%	n	+prep	83.10%	73.26%	77.87%
f	+parents+prep	79.80%	67.92%	73.39%	o	bigrams	84.68%	74.83%	79.45%
g	+prep	80.55%	62.05%	70.10%	p	+terms	**84.85%**	**75.56%**	**79.94%**
h	stems	83.23%	72.71%	77.61%	q	+te+parent	84.83%	75.40%	79.84%
i	split compounds	83.18%	72.37%	77.40%	r	+te+prep	84.84%	75.58%	79.94%
					s	+prep	84.70%	74.88%	79.49%

Table 1. Influence of different feature representations on term identification; highest scores in boldface.

– preprocessing the input text with Gerard Burger's AutoDiag rule-based term-identification tool (section 2.4)

A summary of the results of the experiments can be found in Table 1. Adding term features and parent features to the baseline set, led to small but significant improvements of both precision and recall (a-c). Replacing the baseline features with term features, had a negative influence on performance (d-g). The stem and the split compound features proved to be worse than the baseline set (h-i). Replacing binary weights by tf-idf weights, resulted in significantly better recall scores (j-n). All experiments with bigram features (o-s) reached significantly better precision and recall scores than the baseline. Bigram features with additional term features (p) reached the highest precision (84.85%) and recall (75.56%) scores. The experiments were inconclusive with respect to preprocessing the input texts. In the terms group, the effect was positive (f-g). With bigrams, scores decreased (r-s) and with tf-idf, performance did not change (m-n).

3.2 Changing Learning Parameters and Output Classes

We performed three alternative experiments to see if an additional performance gain could be obtained. First we, evaluated the influence of an important parameter of the machine learning algorithm: the kernel type. In the previous experiments we used a linear kernel. For the next experiment we tested using a polynomial kernel with three different degrees: 1, 2 and 3. With the baseline feature representation, bag of words, the best results were obtained with degree value 2: precision 84.89% and recall 76.11%, both of which outperform the results of the previous section. However, the performance gain came with a price: the polynomial kernels take much more time to train than the linear ones.

In the next two experiments we attempted to take advantage of the assumption that the terms appearing in diagnosis lines are dependent. First, we trained

SVMs to predict bigrams of terms rather than unigrams. However, for both feature representations that we tested, the baseline set and bigram features plus terms, the recall scores decreased significantly when predicting term bigrams. Next, we evaluated a cascaded learner: first train SVMs to perform the classification task with baseline features and then train a second learner with additional features from the output of the first system. The results were similar to the previous experiment: improved precision scores (85.23%) but lower recall (72.40%).

3.3 Changing Data Sets

Additional experiments were performed to determine whether training and testing with data from different time-periods affects performance. The data was divided into three periods of five years (1990–1994, 1995–1999, 2000–2004). From each period 75,000 records were available for training and 24,995 for testing. Best results were obtained with training and test sets from the same time-periods (bag of words: $F_{\beta=1}$ 77.98%, 78.61% and 77.22% respectively). Performance was significantly worse for experiments with training and test sets from different time-periods (on average, 75.41%). When training and test sets were ten years apart, performance was even lower, 74.09%.

3.4 Discussion

The experiments revealed that compared to precision, the recall scores are rather low. I.e., if a classifier assigns a term to a conclusion, it is probably correct, but many positive instances are missed. Despite several attempts to increase recall, it was mainly precision that went up and recall remained relatively low.

Several reasons can be given for explaining why recall scores are lower than precision scores. First, many terms in the diagnosis terms are infrequent and it is hard to train classifiers for classes with a small amount of positive samples. Second, the information needed in a diagnosis line case might not always be available in the corresponding conclusion or that information might be lost in the conversion to features. And third, low recall scores might be caused by the incorrect or inconsistently tagged data.

So there are different possible causes for low recall. But can we expect to attain higher recall scores, or is the problem simply very hard? How well do experts perform if they only have access to the conclusions (instead of the complete report) for coding purposes? Do experts consistently assign the same codes to conclusions? These matters will be investigated in the next section.

4 A Comparison with Domain Experts

In this section, we compare two of the annotation approaches discussed in the previous section with human expert annotators. Based on earlier work, we created a balanced corpus with 1000 texts of which 35% were records for which the baseline obtained a high score, 19% were records with a low score while the

				Kappa scores		
	Precision	**Recall**	$\mathbf{F}_{\beta=1}$	Corpus	P A	P B
bow	83.62%	72.87%	77.87%	0.65	0.58	0.61
bigrams+terms	84.88%	75.47%	79.90%	0.80	0.52	0.56
Pathologist A	71.75%	72.54%	72.14%	0.44		0.65
Pathologist B	66.75%	67.33%	67.04%	0.42	0.55	
AutoDiag	53.10%	62.19%	57.28%	0.22	0.31	0.46

Table 2. Precision, recall, and $F_{\beta=1}$ of new expert ratings compared to the diagnosis lines in the corpus and Kappa agreement scores between the automatic systems, the humans and the corpus, where "P A" ("P B") stands for "Pathologist A" ("Pathologist B"). Scores have been averaged over terms suggested by raters in the first column.

remaining 46% had a medium classification score of the baseline system (details can be found in [21]). Next, two experts were invited to re-annotate the texts based on only the conclusion part. Even though each text had already been coded by experts, it is not obvious that their ratings are correct (or optimal) or that conclusions contain sufficient information for coding. Comparing the new expert ratings to the corpus will enable us to identify differences in term assignments.

We created a web interface which presented a conclusion text to the annotator together with terms predicted by the baseline system and terms that were extracted from the conclusion with a basic term extractor. Terms were grouped into the three main parts of the diagnosis lines: Topography, Procedure and Diagnosis. The two annotators from different hospitals had the opportunity to suggest alternative terms when they regarded the suggested terms as incomplete. Each of the two pathologists took over four and a half hours to complete this task (about sixteen seconds per conclusion text).

The diagnosis lines created by the two experts were compared with the lines in the corpus. Table 2 lists the results as well as the scores of the baseline system, the best bigram system and the rule-based term extractor AutoDiag. The classifiers proved to be better at reproducing the corpus' term assignments than the experts. Another aspect worth noting is that our human annotators score better on recall than on precision, suggesting that the classification task is inherently hard (and not "just" a recall problem).

These are our explanations for the differences between humans and systems:

- some (complex) terms consist of multiple simple terms, and replacing one by the other results in an error
- often when there is a mismatch between two terms, one is just higher or lower in the same hierarchy
- human annotations proved to be more elaborate than system annotations
- humans also had a larger number of conclusion texts with multiple diagnosis lines (10% versus 0)
- while systems always assign terms to conclusion texts, humans frequently assigned the term *unknown* (18% compared with 1% in the corpus)

As an aside, in our evaluation we also included AutoDiag, the rule-based term extractor mentioned before. With an F-score of 57.28% it performed worse than all other methods we considered.

In general, large differences exist between the diagnosis lines in the corpus and the new expert ratings. Amongst themselves the experts also disagree about the terms that should be assigned to conclusions. So, again, the task of assigning diagnostic terms to PALGA conclusions is hard, and it is not just recall that is a problem. At higher levels in the hierarchy of terms, agreement seems to be much better. These results (on the PALGA data set) confirm findings of earlier studies investigating the reliability of coded diagnoses [12, 3], and more general work on the selection of search terms [17, 6].

5 Concluding Remarks

We have examined the potential of machine learning approaches for automatically generating diagnosis lines from summaries of pathological reports. We found that automatic systems perform well in predicting individual diagnosis line terms from text conclusions (precision 85% and recall 75%). However, it proved to be difficult to attain performance levels that were distinctively higher than the baseline scores (83% and 73%).

In a follow-up study, we found that, when restricting access to only the conclusion part of the texts, human experts perform worse than the automatic systems when tested on their ability to reproduce the exact diagnosis lines of the evaluation corpus. This is partly caused by conclusions being incomplete. However, there was also a lack of agreement between the two expert annotators, for example on term specificity. Assigning diagnosis lines to text conclusions proves to be a difficult task.

We conclude that machine learning approaches can achieve good performances in predicting diagnosis lines. By selecting pairs of text conclusions and diagnosis lines for which they perform less well, they can be applied for spotting mismatches between such pairs. Using the predicted diagnosis lines of the systems without an additional manual check would be less appropriate given the machine learner's inability to identify incomplete conclusions. As to supporting the coding task of pathologists, we expect the best results from systems trained on documents of individual doctors, as personal coding assistants.

Acknowledgements

We are very grateful to Ton Tiebosch, Gerard Burger, Arjen van de Pol, and Loes van Velthuysen for data, feedback, and the research problem.

This research was supported by various grants from the Netherlands Organisation for Scientific Research (NWO). Erik Tjong Kim Sang was supported under project number 264.70.050. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220.80.001, 264.70.050, 354.20.005, 600.065.120, 612.13.001, 612.000.106, 612.066.302, 612.069.006, 640.001.501, and 640.002.501.

References

1. J.J. Cimino. Review paper: coding systems in health care. *Methods Inf Med*, 35(4-5):273–84, 1996.
2. B. de Bruijn. *Automatic Classification of Pathology Reports*. PhD thesis, Maastricht University, 1997.
3. J. Dixon, C. Sanderson, P. Elliott, P. Walls, J. Jones, and M. Petticrew. Assessment of the reproducibility of clinical coding in routinely collected hospital activity data: a study in two hospitals. *J Public Health Med*, 20(1):63–9, Mar 1998.
4. P. Franz, A. Zaiss, S. Schulz, U. Hahn, and R. Klar. Automated coding of diagnoses–three methods compared. *Proc AMIA Symp*, pages 250–4, 2000.
5. P. A. Hall and N. R. Lemoine. Comparison of manual data coding errors in two hospitals. *J Clin Pathol*, 39(6):622–6, Jun 1986.
6. M. Iivonen. Consistency in the selection of search concepts and search terms. *Information Processing and Management*, 31:173–190, 1995.
7. T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning*. Springer, 1998.
8. T. Joachims. Making large-scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.
9. L. Letrilliart, C. Viboud, P.Y. Boëlle, and A. Flahault. Automatic coding of reasons for hospital referral from general medicine free-text reports. *Proc AMIA Symp*, pages 487–91, 2000.
10. D. P. Lorence and R. Jameson. Managers reports of automated coding system adoption and effects on data quality. *Methods Inf Med*, 42(3):236–42, 2003.
11. Robert Moskovitch, Shiva Cohen-Kashi, Uzi Dror, Iftah Levy, Amit Maimon, and Yuval Shahar. Multiple hierarchical classification of free-text clinical guidelines. *Artificial Intelligence in Medicine*, 37(3):177–190, 2006.
12. G. Nilsson, H. Petersson, H. Ahlfeldt, and L. E. Strender. Evaluation of three Swedish ICD-10 primary care versions: reliability and ease of use in diagnostic coding. *Methods of information in medicine*, 39(4-5):325–31, Dec 2000.
13. PALGA. Thesaurus coderen (PALGA), 2005.
14. A. L. Rector. Clinical terminology: why is it so hard? *Methods Inf Med*, 38(4-5):239–52, Dec 1999.
15. B. Ribeiro-Neto, A.H.F. Laender, and L.R.S. de Lima. An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science and Technology*, 52(5):391–401, 2001.
16. N. Sager, C. Friedman, and S. Margaret. *Medical language processing: computer management of narrative data*. Addison-Wesley, 1987.
17. T. Saracevic and P.B. Kantor. A study of information seeking and retrieving. III. searchers, searches, overlap. *Journal of the American Society for Information Science and Technology*, 39:197–216, 1988.
18. Snowball. Porter stemmer for Dutch. http://www.snowball.tartarus.org/.
19. G. Surján. Questions on validity of International Classification of Diseases-coded diagnoses. *Int J Med Inform*, 54(2):77–95, May 1999.
20. F. Wingert. Automated indexing based on SNOMED. *Methods Inf Med*, 24(1):27–34, 1985.
21. J. van der Zwaan. Development and evaluation of a method for the automatic coding of pathology report conclusions. Master's thesis, Faculty of Science, University of Amsterdam, 2006.