Finding Negative Symptoms of Schizophrenia in Patient Records

Genevieve Gorrell
The University of Sheffield
g.gorrell@sheffield.ac.uk

Angus Roberts
The University of Sheffield
a.roberts@dcs.shef.ac.uk

Richard Jackson
King's College London
Richard.G.Jackson@slam.nhs.uk

Robert Stewart
King's College London
robert.stewart@kcl.ac.uk

Abstract

This paper reports the automatic extraction of eleven negative symptoms of schizophrenia from patient medical records. The task offers a range of difficulties depending on the consistency and complexity with which mental health professionals describe each. In order to reduce the cost of system development, rapid prototypes are built with minimal adaptation and configuration of existing software, and additional training data is obtained by annotating automatically extracted symptoms for which the system has low confidence. The system was further improved by the addition of a manually engineered rule based approach. Rule-based and machine learning approaches are combined in various ways to achieve the optimal result for each symptom. Precisions in the range of 0.8 to 0.99 have been obtained.

1 Introduction

There is a large literature on information extraction (IE) from the unstructured text of medical records (see (Meystre et al., 2008) for the most recent review). Relatively little of this literature, however, is specific to psychiatric records (see (Sohn et al., 2011; Lloyd et al., 2009; Roque et al., 2011) for exceptions to this). The research presented here helps to fill this gap, reporting the extraction of schizophrenia symptomatology from free text in the case register of a large mental health unit, the South London and Maudsley NHS Trust (SLaM).

We report the extraction of negative symptoms of schizophrenia, such as poor motivation, social withdrawal and apathy. These often present in addition to more prominent, positive symptoms such as delusions and hallucinations. Negative symptoms can severely impair the quality of life of affected patients, yet existing antipsychotic medications have poor efficacy in their treatment. As negative symptoms can be measured in quantitative frameworks within a clinical environment (Kay et al., 1987; Andreasen, 1983), they have the potential to reflect the success or failure of new medical interventions, and are of widespread interest in the epidemiology of schizophrenia. The motivation for our work is to provide information on the presence of negative symptoms, for use in such quantitative measures.

SLaM covers a population of 1.1 million, being responsible for close to 100% of the mental health care contacts in four London boroughs. Approximately 225,000 records are stored in the SLaM Electronic Health Record (EHR) system, which supports an average of 35,000 patients at any one time. SLaM hosts the UK National Institute for Health Research (NIHR) Biomedical Research Center (BRC) for Mental Health. The BRC deidentifies all records in the SLaM EHR (Fernandes et al., 2013) to form the largest mental health case register in Europe, the Case Register Interactive Search (CRIS) system (Stewart et al., 2009). CRIS provides BRC epidemiologists with search facilities, via a web front end that allows standard information retrieval queries over an inverted index, and via database query languages. CRIS has been approved as an anonymized data resource for secondary analysis by Oxfordshire Research Ethics Committee C (08/H0606/71). The governance for all CRIS projects and dissemination is managed through a patient-led oversight committee.

CRIS contains both the structured information, and the unstructured free text from the SLaM

Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013, pages 9–17,
Hissar, Bulgaria, 13 September 2013.

EHR. The free text consists of 18 million text field instances – a mix of correspondence and notes describing patient encounters. Much of the information of value to mental health epidemiologists is found in these free text fields. SLaM clinicians record important information in the textual portion of the record, even when facilities are provided for recording the same information in a structured format. For example, a query on the structured fields containing Mini Mental State Examination scores (MMSE, a score of cognitive ability) recently returned 5,700 instances, whereas a keyword search over the free text fields returned an additional 48,750 instances. The CRIS inverted index search system, however, cannot return the specific information of interest (the MMSE score in this case), instead returning each text field that contains a query match, in its entirety. In the case of symptomatology, as examined in this paper, symptoms are rarely recorded in structured fields, but are frequently mentioned in the unstructured text.

This problem is not unusual. (Meystre et al., 2008) note that free text is "convenient to express concepts and events" (Meystre et al., 2008), but that it is difficult for re-use in other applications, and difficult for statistical analysis. (Rosenbloom et al., 2011) have reviewed the few studies that look at the expressivity of structured clinical documentation systems compared to natural prose notes, and report that prose is more accurate, reliable and understandable. (Powsner et al., 1998) refer to structured data as freezing clinical language, and restricting what may be said. (Greenhalgh et al., 2009), referring to the free text of the paper record, say that it is tolerant of ambiguity, which supports the complexity of clinical practice. Much of medical language is hedged with ambiguity and probability, which is difficult to represent as structured data (Scott et al., 2012).

Given the presence of large quantities of valuable information in the unstructured portion of the BRC case register, and CRIS's inability to extract this information using standard information retrieval techniques, it was decided, in 2009, to implement an IE and text mining capability as a component of CRIS. This comprises tools to develop and evaluate IE applications for specific end-user requirements as they emerge, and the facility to deploy these applications on the BRC compute cluster.

Most IE applications developed by the BRC to date have used a pattern matching approach. In this, simple lexico-syntactic pre-processing and dictionary lookup of technical terms are followed by cascades of pattern matching grammars designed to find the target of extraction. These grammars are hand-written by language engineers. Previous extraction targets have included smoking status, medications, diagnosis, MMSE, level of education, and receipt of social care. Building such pattern matching grammars is often time consuming, in that it takes significant language engineer time to develop and refine grammars. In addition, the process of writing and testing grammars requires examples of the extraction target. These are provided by manual annotation, or labelling, of examples and correction of system output; a task which takes significant domain expert time.

In the case of schizophrenia, the IE applications are required to extract multiple symptoms for use in quantitative measures of the disease. The set of symptoms relevant to such quantitative measures number in the dozens. Given the cost of pattern grammar development, and the cost of manual annotation, it is impractical to develop grammars for each of the required symptoms, and such an approach would not scale up to larger numbers of symptoms and to other diseases. In addition, the cost of domain expert annotation of examples for each individual symptom is also high. The approach taken in our research aims to reduce these two costs.

In order to reduce the cost of system development, and to improve scalability to new symptoms and diseases, we build rapid prototypes, using off-the-shelf NLP and machine learning (ML) toolkits. Such toolkits, and repositories of applications built on them, are becoming increasingly popular. It has been asked (Nadkarni et al., 2011) whether such tools may be used as "commodity software" to create clinical IE applications with little or no specialist skills. In order to help answer this question, we compare the performance of our ML only prototypes to applications that combine ML and pattern matching, and to applications implemented with pattern matching alone.

The second cost considered is that of finding and labelling high quality examples of the extraction target, used to inform and test system development. To deal with this cost, we explore methods of enriching the pool of examples for labelling,

including the use of methods inspired by active learning (Settles, 2012). In active learning, potential examples of the extraction target are selected by the learning algorithm for labelling by the human annotator. The aim is to present instances which will most benefit the ML algorithm, at least human cost. This paper presents results from experiments in training data enrichment, and a simple approach to active learning, applied to symptom extraction.

The paper is organised as follows. Section 2 looks at the task domain in more detail, explaining the symptoms to be extracted, and describing the dataset. Section 3 describes the experimental method used, and the evaluation metrics. This is followed by a presentation of the results in Section 4, and a discussion of these results in Section 5. Finally, we draw some conclusions in Section 6.

2 Analysis of the Task Domain

In this section we will first introduce the concept of negative symptoms and explain what entities we are aiming to extract from the data. We will then discuss the datasets we used, and how each symptom varies in its nature and therefore difficulty.

2.1 Negative Symptoms

In the psychiatric context, negative symptoms are deficit symptoms; those that describe an absence of a behaviour or ability that would normally be present. A positive symptom would be one which is not normally present. In schizophrenia, positive symptoms might include delusions, auditory hallucinations and thought disorder. Here, we are concerned with negative symptoms of schizophrenia, in particular the following eleven, where bold font indicates the feature values we hope to extract from the data (in machine learning terms, the classes, not including the negative class). Examples illustrate something of the ways in which the symptom might be described in text. "ZZZZZ" replaces the patient name for anonymization purposes:

- **Abstract Thinking**: Does the individual show evidence of requiring particularly **concrete** conceptualizations in order to understand? Examples include; "Staff have noted ZZZZZ is very concrete in his thinking", "Thought disordered with concrete thinking", but NOT "However ZZZZZ has no concrete plans to self-harm"

- **Affect**: Is the individual's emotional response **blunted** or **flat**? Is it inappropriate to events (**abnormal**)? Alternatively, does the individual respond appropriately (**reactive**)? Examples include; "Mood: subjectively 'okay' however objectively incongruent", "Denied low mood or suicide ideation", "showed blunting of affect"

- **Apathy**: Does the individual exhibit **apathy**? Examples include; "somewhat apathetic during his engagement in tasks", "Apathy."

- **Emotional Withdrawal**: Does the individual appear **withdrawn** or **detached**? Examples include; "withdrawal from affectational and social contacts", "has been a bit withdrawn recently", NOT "socially withdrawn", which is a separate symptom, described below.

- **Eye Contact**: Does the individual make **good** eye contact, or is it **intermediate** or **poor**? Examples include; "eye contact was poor", "maintaining eye contact longer than required", "made good eye contact"

- **Motivation**: Is motivation **poor**? Examples include; "ZZZZZ struggles to become motivated.", "ZZZZZ lacks motivation.", "This is due to low motivation."

- **Mutism**: A more extreme version of poverty of speech (below), and considered a separate symptom, is the individual **mute** (but not deaf mute)? Examples include; "Was electively mute [...]", "ZZZZZ kept to himself and was mute.", NOT "ZZZZZ is deaf mute."

- **Negative Symptoms**: An umbrella term for the symptoms described here. Do we see any **negative symptom**? Examples include; "main problem seems to be negative symptoms [...]", "[...] having negative symptoms of schizophrenia."

- **Poverty of Speech**: The individual may show a deficit or **poverty** of speech, or their speech may be **abnormal** or **normal**. Examples include; "Speech: normal rate and rhythm", "speech aspontaneous", "speech

was dysarthric", "ongoing marked speech defect", "speech was coherent and not pressured"

- **Rapport**: Individual ability to form conversational rapport may be **poor** or **good**. Examples include; "we could establish a good rapport", "has built a good rapport with her carer"

- **Social Withdrawal**: Do we see indications of **social withdrawal** or not? Examples include; "long term evidence of social withdrawal", "ZZZZZ is quite socially withdrawn"

2.2 Dataset

Different symptoms vary in the challenges they pose. For example, "apathy" is almost exclusively referred to using the word "apathy" or "apathetic", and where this word appears, it is almost certainly a reference to the negative symptom of apathy, whereas concrete thinking is harder to locate because the word "concrete" appears so often in other contexts, and because concrete thinking may be referred to in less obvious ways. In the previous section, we gave some examples of negative symptom mentions that give an idea of the range of possibilities. Exemplars were unevenly distributed among medical records, with some records having several and others having none.

Due to the expertize level required for the annotation part of the task, and strict limitations on who is authorized to view the data, annotation was performed by a single psychiatrist. Data quantity was therefore limited by the amount of time the expert annotator had available for the work. For this reason, formal interannotator agreement assessment was not possible, although a second annotator did perform some consistency checking on the data. Maximizing the utility of a limited dataset therefore constituted an important part of the work.

Because many of the records do not contain any mention of the symptom in question, in order to make a perfect gold standard corpus the expert annotator would have to read a large number of potentially very lengthy documents looking for mentions that are thin on the ground. Because expert annotator time was so scarce, this was likely to lead to a much reduced corpus size, and so a compromise was arrived at whereby simple heuristics were used to select candidate mentions for the annotator to judge rather than having also to find them. For example, in abstract thinking, one heuristic used was to identify all mentions of "concrete". In some cases, the mention is irrelevant to concrete thinking, so the annotator marks it as a negative, whereas in others it is a positive mention. This means that compared with a fully annotated corpus, our data may be lower on recall, since some cases may not have been identified using the simple heuristics, though precision is most likely excellent, since all positive examples have been fully annotated by the expert. In terms of the results reported here, this compromise has little impact, since the task is defined to be replicating the expert annotations, whatever they may be. However, it might be suggested that our task is a little easier than it would have been for a fully annotated corpus, since the simple heuristics used to identify mentions would bias the task toward the easier cases. In terms of the adequacy of the result for future use cases, precision is the priority so this decision was made with end use in mind.

2.2.1 Selecting examples for training

As a further attempt to obtain more expert-annotated data, the principles of active learning were applied in order to strategically leverage annotator time on the most difficult cases and for the most difficult symptoms. Candidate mentions were extracted with full sentence context on the basis of their confidence scores, as supplied by the classifier algorithm, and presented to the annotator for judgement. Mentions were presented in reverse confidence score order, so that annotator time was prioritized on those examples where the classifier was most confused.

3 Method

Because the boundaries of a mention of a negative symptom are somewhat open to debate, due to the wide variety of ways in which psychiatric professionals may describe a negative symptom, we defined the boundaries to be sentence boundaries, thus transforming it into a sentence classification task. However, for evaluation purposes, precision, recall and F1 are used here, since observed agreement is not appropriate for an entity extraction task, giving an inflated result due to the inevitably large number of correctly classified negative examples.

Due to the requirements of the use case, our work was biased toward achieving a good preci-

sion. Future work making use of the data depends upon the results being of good quality, whereas a lower recall will only mean that a smaller proportion of the very large amount of data is available. For this reason, we aimed, where possible, to achieve precisions in the region of 0.9 or higher, even at the expense of recalls below 0.6.

Our approach was to produce a rapid prototype with a machine learning approach, and then to combine this with rule-based approaches in an attempt to improve performance. Various methods of combining the two approaches were tried. Machine learning alone was performed using support vector machines (SVMs). Two rule phases were then added, each with a separate emphasis on improving either precision or recall. The rule-based approach was then tried in the absence of a machine learning component, and in addition both overriding the ML where it disagreed and being overridden by it. Rules were created using the JAPE language (Cunningham et al., 2000). Experiments were performed using GATE (Cunningham et al., 2011; Cunningham et al., 2013), and the SVM implementation provided with GATE (Li et al., 2009).

Evaluation was performed using fivefold cross-validation, to give values for precision, recall and F1 using standard definitions. For some symptoms, active learning data were available (see Section 2.2.1) comprising a list of examples chosen for having a low confidence score on earlier versions of the system. For these symptoms, we first give a result for systems trained on the original dataset. Then, in order to evaluate the impact of this intervention, we give results for systems trained on data including the specially selected data. However, at test time, these data constitute a glut of misrepresentatively difficult examples that would have given a deflated result. We want to include these only at training time and not at test time. Therefore, the fold that contained these data in the test set was excluded from the calculation. For these symptoms, evaluation was based on the four out of five folds where the active learning data fell in the training set. The symptoms to which this applies are abstract thinking, affect, emotional withdrawal, poverty of speech and rapport.

In the next section, results are presented for these experiments. The discussion section focuses on how results varied for different symptoms, both in the approach found optimal and the result achieved, and why this might have been the case.

4 Results

Table 1 shows results for each symptom obtained using an initial "rapid prototype" support vector machine learner. Confidence threshold in all cases is 0.4 except for negative symptoms, where the confidence threshold is 0.6 to improve precision. Features used were word unigrams in the sentence in conjunction with part of speech (to distinguish for example "affect" as a noun from "affect" as a verb) as well as some key terms flagged as relevant to the domain. Longer n-grams were rejected as a feature due to the small corpus sizes and consequent risk of overfitting. A linear kernel was used. The soft margins parameter was set to 0.7, allowing some strategic misclassification in boundary selection. An uneven margins parameter was used (Li and Shawe-Taylor, 2003; Li et al., 2005) and set to 0.4, indicating that the boundary should be positioned closer to the negative data to compensate for uneven class sizes and guard against small classes being penalized for their rarity. Since the amount of data available was small, we were not able to reserve a validation set, so care was taken to select parameter values on the basis of theory rather than experimentation on the test set, although confidence thresholds were set pragmatically. Table 1 also gives the number of classes, including the negative class (recall that different symptoms have different numbers of classes), and number of training examples, which give some information about task difficulty.

As described in Section 2.2.1, active learning-style training examples were also included for symptoms where it was deemed likely to be of benefit. Table 2 provides performance statistics for these symptoms alongside the original machine learning result for comparison. In all cases, some improvement was observed, though the extent of the improvement was highly variable.

Central to our work is investigating the interplay between rule-based and machine learning approaches. Rules were prepared for most symptoms, with the intention that they should be complementary to the machine learning system, rather than a competitor. The emphasis with the rules is on coding for the common patterns in both positive and negative examples, though coding the ways in which a symptom might not be referred

Table 1: Machine Learning Only, SVM

Symptom	Classes	Training Ex.	Precision	Recall	F1
Abstract Thinking	2	118	0.615	0.899	0.731
Affect	5	103	0.949	0.691	0.8
Apathy	2	145	0.880	0.965	0.921
Emotional Withdrawal	3	118	0.688	0.815	0.746
Eye Contact	4	35	0.827	0.677	0.745
Motivation	2	259	0.878	0.531	0.662
Mutism	2	234	0.978	0.936	0.956
Negative Symptoms	2	185	0.818	0.897	0.856
Poverty of Speech	4	263	0.772	0.597	0.674
Rapport	3	139	0.775	0.693	0.731
Social Withdrawal	2	166	0.940	0.958	0.949

Table 2: Active Learning

Symptom	Ex.	Without AL-Style Examples			With AL-Style Examples			Difference
		Prec	Rec	F1	Prec	Rec	F1	
Abstract Thinking	99	0.595	0.940	0.728	0.615	0.899	0.731	0.003
Affect	200	0.947	0.529	0.679	0.949	0.691	0.8	0.121
Emotional Withdrawal	100	0.726	0.517	0.604	0.688	0.815	0.746	0.142
Poverty of Speech	62	0.721	0.515	0.601	0.772	0.597	0.674	0.073
Rapport	37	0.725	0.621	0.669	0.775	0.693	0.731	0.062

to is considerably harder. F1 results for the standalone rule-based systems where sufficiently complete are given in Table 4; however, for now, we focus on the results of our experiments in combining the two approaches, which are given in Table 3. Here, we give results for layering rules with machine learning. On the left, we see results obtained where ML first classifies the examples, then the rule-based approach overrides any ML classification it disagrees with. In this way, the rules take priority. On the right, we see results obtained where machine learning overrides any rule-based classification it disagrees with. The higher of the F1 scores is given in bold. Results suggest that the more successful system is obtained by overriding machine learning with rules rather than vice versa.

Table 4 gives a summary of the best results obtained by symptom, using all training data, including active learning instances. We focus on F1 scores only here for conciseness. The baseline machine learning result is first recapped, along with the rule-based F1 where this was sufficiently complete to stand alone. Since in all cases, overriding machine learning with rules led to the best result of the two combination experiments, we give the F1 for this, which in all cases, where available, proves the best result of all. We provide the percentage improvement generated relative to the ML baseline by the combined approach. The final column recaps the best F1 obtained for that symptom. We can clearly see from Table 4 that in all cases, the result obtained from combining approaches outperforms either of the approaches taken alone.

5 Discussion

In summary, the best results were obtained by building upon a basic SVM system with layers of rules that completed and corrected areas of weakness in the machine learning. Note that the symptoms where this approach yielded the most striking improvements tended to be those with the fewer training examples and the larger numbers of classes. In these cases, the machine learning approach is both easier to supplement using rules and easier to beat. A high performing rule-based system certainly correlates with a substantial improvement over the ML baseline; however, we

Table 3: Machine Learning Layered with Rules

Symptom	**Rules Override ML**			**ML Overrides Rules**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Abstract Thinking	0.914	0.719	**0.805**	0.935	0.652	0.768
Affect	0.931	0.827	0.876	0.931	0.827	0.876
Emotional Withdrawal	0.840	0.778	**0.808**	0.691	0.827	0.753
Eye Contact	0.88	0.852	**0.866**	0.779	0.611	0.684
Mutism	0.986	0.936	**0.960**	0.978	0.936	0.956
Negative Symptoms	0.851	0.897	**0.874**	0.818	0.897	0.856
Poverty of Speech	0.8	0.730	**0.763**	0.793	0.723	0.757
Rapport	0.839	0.868	**0.853**	0.907	0.772	0.834

Table 4: Best Result Per Symptom

Symptom	**Classes**	**Ex.**	**ML F1**	**Rules F1**	**Rules>ML F1**	**% Imp**	**Best F1**
Abstract Thinking	2	217	0.731	0.765	0.805	10%	0.805
Affect	5	303	0.800	0.820	0.876	9%	0.876
Apathy	2	145	0.921	n/a	n/a	n/a	0.921
Emotional withdrawal	3	218	0.746	0.452	0.808	8%	0.808
Eye contact	4	35	0.745	0.859	0.866	16%	0.866
Motivation	2	259	0.662	n/a	n/a	n/a	0.662
Mutism	2	234	0.956	n/a	0.960	0%	0.960
Negative Symptoms	2	185	0.856	n/a	0.874	2%	0.874
Poverty of speech	4	325	0.674	0.689	0.763	13%	0.763
Rapport	3	176	0.731	0.826	0.853	17%	0.853
Social withdrawal	2	166	0.949	n/a	n/a	n/a	0.949

do also consistently see the combined approach outperforming both the ML and rule-based approaches as taken separately. We infer that this approach is of the most value in cases where training data is scarce.

Where machine learning was removed completely, we tended to see small performance decreases, but in particular, recall was badly affected. Precision, in some cases, improved, but not by as much as recall decreased. This seems to suggest that where datasets are limited, machine learning is of value in picking up a wider variety of ways of expressing symptoms. Of course, this depends on a) the coverage of the rules against which the SVM is being contrasted, and b) the confidence threshold of the SVM and other relevant parameters. However, this effect persisted even after varying the confidence threshold of the SVM quite substantially.

Optimizing precision presented more difficulties than improving recall. Varying the confidence threshold of the SVM to improve recall tended to cost more in recall than was gained in precision, so rule-based approaches were employed. However, it is much easier to specify what patterns do indicate a particular symptom than list all the ways in which the symptom might *not* be referred to. Symptoms varied a lot with respect to the extent of the precision problem. In particular, abstract thinking, which relies a lot on the word "concrete", which may appear in many contexts, posed problems, as did emotional withdrawal, which is often indicated by quite varied use of the word "withdrawn", which may occur in many contexts. Other symptoms, whilst easier than abstract thinking and social withdrawal, are also variable in the way they are expressed. Mood, for example, is often described in expressive and indirect ways, as is poverty of speech. On the other hand, mutism is usually very simply described, as is eye contact. It is an aid in this task that medical professionals often use quite formalized and predictable ways of referring to symptoms.

Aside from that, task difficulty depended to a large extent on the number of categories into which symptoms may be split. For example, the simple "mute" category is easier than eye contact, which may be good, intermediate or poor, with intermediate often being difficult to separate from good and poor. Likewise, speech may show poverty or be normal or abnormal, with many different types of problem indicating abnormality.

We chose to use an existing open-source language engineering toolkit for the creation of our applications; namely GATE (Cunningham et al., 2011). This approach enabled rapid prototyping, allowing us to make substantial progress on a large number of symptoms in a short space of time. The first version of a new symptom was added using default tool settings and with no additional programming. It was often added to the repertoire in under an hour, and although not giving the best results, this did achieve a fair degree of success, as seen in Table 1 which presents the machine learning-only results. In the case of the simpler symptoms (apathy and social withdrawal), this initial system gave sufficient performance to require no further development.

Additional training data was obtained for five symptoms, by presenting labelled sentences with low classifier confidence to the annotator (Table 2). Although this did improve performance, it is unclear whether this was due to an increase in training data alone, or whether concentrating on the low confidence examples made a difference. The annotator did, however, report that they found this approach easier, and that it took less time than annotating full documents for each symptom.

6 Conclusion

In conclusion, a good degree of success has been achieved in finding and classifying negative symptoms of schizophrenia in medical records, with precisions in the range of 0.8 to 0.99 being achieved whilst retaining recalls in excess of 0.5 and in some cases as high as 0.96. The work has unlocked key variables that were previously inaccessible within the unstructured free text of clinical records. The resulting output will now feed into epidemiological studies by the NIHR Biomedical Research Centre for Mental Health.

We asked whether off-the-shelf language engineering software could be used to build symptom extraction applications, with little or no additional configuration. We found that it is possible to create prototypes using such a tool, and that in the case of straightforward symptoms, these perform well. In the case of other symptoms, however, language engineering skills are required to enhance performance. The best results were obtained by adding hand-crafted rules that dealt with weakness in the machine learning.

References

N. C. Andreasen. 1983. *Scale for the Assessment of Negative Symptoms*. University of Iowa Press, Iowa City. Cited by 0000.

H. Cunningham, D. Maynard, and V. Tablan. 2000. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, Sheffield, UK, November.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M.A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. 2011. *Text Processing with GATE (Version 6)*.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2):e1002854, 02.

Andrea C Fernandes, Danielle Cloete, Matthew TM Broadbent, Richard D Hayes, Chin-Kuo Chang, Angus Roberts, Jason Tsang, Murat Soncul, Jennifer Liebscher, Richard G Jackson, Robert Stewart, and Felicity Callard. 2013. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Medical Informatics and Decision Making*. Accepted for publication.

T. Greenhalgh, H. W. Potts, G. Wong, P. Bark, and D. Swinglehurst. 2009. Tensions and paradoxes in electronic patient record research: a systematic literature review using the meta-narrative method. *Milbank Quarterly*, 87(4):729–788, Dec.

S R Kay, A Fiszbein, and L A Opler. 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin*, 13(2):261–276. Cited by 8221.

Y. Li and J. Shawe-Taylor. 2003. The SVM with Uneven Margins and Chinese Document Categorization. In *Proceedings of The 17th Pacific Asia Conference on Language, Information and Computation (PACLIC17)*, Singapore, Oct.

Y. Li, K. Bontcheva, and H. Cunningham. 2005. Using Uneven Margins SVM and Perceptron for Information Extraction. In *Proceedings of Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*.

Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. 2009. Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. *Natural Language Engineering*, 15(2):241–271.

Keith Lloyd, Matteo Cella, Michael Tanenblatt, and Anni Coden. 2009. Analysis of clinical uncertainties by health professionals and patients: an example from mental health. *BMC Medical Informatics and Decision Making*, 9(1):34.

S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and J.F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *IMIA Yearbook of Medical Informatics*, pages 128–144.

P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. 2011. Natural language processing: an introduction. *J Am Med Inform Assoc*, 18(5):544–551.

S. M. Powsner, J. C. Wyatt, and P. Wright. 1998. Opportunities for and challenges of computerisation. *Lancet*, 352(9140):1617–1622, Nov.

Francisco S. Roque, Peter B. Jensen, Henriette Schmock, Marlene Dalgaard, Massimo Andreatta, Thomas Hansen, Karen Seby, Sren Bredkjr, Anders Juul, Thomas Werge, Lars J. Jensen, and Sren Brunak. 2011. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol*, 7(8):e1002141, 08.

S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc*, 18(2):181–186.

Donia Scott, Rossano Barone, and Rob Koeling. 2012. Corpus annotation as a scientific task. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Burr Settles. 2012. *Active Learning*. Morgan and Claypool.

Sunghwan Sohn, Jean-Pierre A Kocher, Christopher G Chute, and Guergana K Savova. 2011. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *Journal of the American Medical Informatics Association*, 18(Suppl 1):i144–i149.

Robert Stewart, Mishael Soremekun, Gayan Perera, Matthew Broadbent, Felicity Callard, Mike Denis, Matthew Hotopf, Graham Thornicroft, and Simon Lovestone. 2009. The South London and Maudsley NHS foundation trust biomedical research centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry*, 9:51–62.