Lexical Acquisition for Clinical Text Mining using Distributional Similarity

John Carroll[1], Rob Koeling[1], and Shivani Puri[2]

[1] Department of Informatics, University of Sussex, Brighton BN1 9QH, UK
[2] GPRD, 151 Buckingham Palace Road, London SW1W 9SZ, UK

Abstract. We describe experiments into the use of distributional similarity for acquiring lexical information from clinical free text, in particular notes typed by primary care physicians (general practitioners). We also present a novel approach to lexical acquisition from 'sensitive' text, which does not require the text to be manually anonymised – a very expensive process – and therefore allows much larger datasets to be used than would normally be possible.

1 Introduction

In the UK, almost all primary care physicians (general practitioners, or GPs) use one of a small number of computer systems for managing their patients' medical records. These systems provide facilities for electronic storage, retrieval and modification of records, allowing GPs to enter orders for prescriptions, request tests, view laboratory results, read letters sent by hospital consultants, consult the notes of previous patient consultations and enter new notes. The records contain information in the form of codes, dates, numeric quantities, and free text. From a GP's perspective, the main purpose of the records is to ensure individual patients obtain high quality medical care; however, records for individual patients are also used for medico-legal and health insurance purposes, and on a collective basis to support healthcare audits and to calculate incentivised payments to GPs for specific quality indicators. In addition, samples of records are collected regionally and nationally for other purposes, including health services research, epidemiological studies, and monitoring the safety of medications.

In these electronic medical records, Read codes [1] (the standard system for clinical terminology in the UK) are used to enter symptoms, test results, diagnoses and procedures (and also personal and administrative information). However, during consultations GPs usually do not input all relevant codes due to pressure of time, lack of incentive or relevant training, unwillingness to code symptoms that at the time seem not to be salient or where there is clinical uncertainty, and reluctance to code conditions that would normally be diagnosed by a specialist. In these situations the GP would be likely to type such information into the computer system in the form of unstructured free text. As well as uncoded symptoms and diagnoses, there is often a considerable amount of clinical information in the free text notes, including information on the severity

Published as:
John Carroll, Rob Koeling, and Shivani Puri (2012) Lexical acquisition for clinical text mining using distributional similarity. In Alexander Gelbukh (ed.) *Computational Linguistics and Intelligent Text Processing, Springer Lecture Notes in Computer Science, Volume 7182*, 232-246. DOI: 10.1007/978-3-642-28601-8_20

of symptoms, observations on examinations made by the GP, and information relating to diagnoses that the GP has ruled out. If all information were entered in a structured fashion it would be more amenable to automatic analysis, but the flexibility of the written language is necessary in order to capture the nuanced nature of much of this information and the variability between patients [9, 18].

In recent years there has been a lot of interest in applying natural language processing techniques to clinical text. A further motivation in addition to those mentioned above (particularly in the USA) is to be able to automatically bill medical insurers for medications that have been administered and clinical procedures that have been performed. Some of the research activity has been centered around shared tasks, for example those organised by The Cincinnati Children's Hospital Medical Center involving assignment of clinical codes to radiology reports [16], and by i2b2 ('Informatics for Integrating Biology and the Bedside')[3] including identifying patient smoking status and extraction of medication information from discharge summaries [21, 22]. The information retrieval conference TREC 2011[4] also included a shared task of retrieving clinical cases from narrative records [15].

Although these shared tasks and the deployed application systems that inspired them – as well as many other research efforts in clinical text mining (e.g. [19]) – involve documents of a variety of types (including discharge summaries, diagnostic test reports, and letters written by non-clinicians), the documents were produced from transcribed handwriting or dictation followed by post-transcription checking and editing, and consist of fairly standard language.

In contrast, notes typed by GPs do not go through any transcription or editing process, and are usually not carefully written. As mentioned above, the notes are written to support patient care within the GP practice, and not with the intention that they should be shared with other people. The notes often contain segments of informal language (not using medical terminology) summarising what the patient reported about their condition in their own words, and are usually typed by the GP during the consultation under pressure of time and with the competing requirement to attend simultaneously to the patient. Moreover, since these are notes taken in primary care, they potentially relate to very wide range of medical conditions.

These considerations make automatic processing of free text notes written by GPs a challenging task. However it is potentially a useful task, since epidemiological studies have shown that the free text contains important information relating to symptoms that is not available in the coded part of records [6, 8]. In recent work, we have been applying natural language processing techniques to free text notes in order to extract certain kinds of uncoded information. We have created a corpus of clinical free text records in which we have annotated all mentions of the main symptoms of ovarian cancer [10]. We have been using this corpus to estimate the quantity of symptom-related information in records that

[3] http://www.i2b2.org
[4] http://trec.nist.gov

is not coded, and to develop techniques for automatically recognising mentions of these symptoms in free text [12].

Our current approach to recognising mentions of symptoms is based on manually curated word lists and relatively straightforward string matching technique, in which precision is optimised at the expense of recall. In this paper, we investigate methods based on distributional similarity for improving the recall of this and similar approaches, by automatically acquiring variant ways of expressing relevant words. Section 2 characterises the text data that we are using, summarises our approach to automatic symptom recognition, and motivates our use of distributional similarity methods. Section 3 discusses the datasets of free text records which we use in our investigation. We then go on to describe how we compute distributional similarity (Sect. 4) and the experiments we have conducted (Sect. 5). Finally, we conclude and outline directions for future work.

2 Background and Motivation

2.1 Primary Care Free Text

Our data is drawn from the General Practice Research Database (GPRD),[5] which contains records of about five million currently registered patients from around 630 general practices throughout the UK. GPRD data is used worldwide for research by the pharmaceutical industry, clinical research organisations, regulators, government departments and academic institutions. The free text in these records is of two main types: *notes* written by the GP, and *letters* sent to the practice by other agencies (primarily hospitals and mental health services) that the patient has been referred to. Letters are often entered by clerical staff in the practice, in the form of OCRed versions of the original hardcopy documents.

Notes written by GPs exhibit a very terse, telegraphic style with limited use of full sentence syntax; in particular, sentential subjects are very rare, and even finite verbs are uncommon (see (1a) below). Discourse connectives and conjunctions (e.g. *but*, *however*, *and*, *or*) are oten omitted where they would normally be present, as in (1b).

(1) a *chiropracter seen*
b *sleeps well, low and tearful*

GPs also make widespread use of abbreviations and acronyms, some of which are conventionalised whereas others have a range of variations. For example, in (2a), it is likely that *bl* has been used to abbreviate 'bleeding' and *D* to abbreviate 'diarrhoea'. In (2c), *cpn* is the standard acronym for 'community psychatric nurse'.

(2) a *no bl no D*
b *Had TAH and BSO 4/52 ago*
c *prev h/o depression, ref cpn*

[5] http://www.gprd.com

Abbreviated words can be ambiguous, for example *occ* may mean any of *occurred*, *occasional* or *occupational* depending on the context. Spelling mistakes and anomalous tokenisation (e.g. missing spaces) are ubiquitous.

(3) a *exmiane and futher hx needed and conisder mirena or orther form of contracpetion*
b *Rx amoxiciilin today,now feeling worse,burning up but feels cold,no energy*
c *has 4 children, house needing renovation+working.*

Question marks and other shorthand means of indicating possibility or change are frequently used ambiguously.

(4) a *Postnatal depression ? sl better on lofepramine*
b *small ? 2 outer left breast tender*
c *Shortness of breath +chest tightness*

Punctuation is used inconsistently or idiosyncratically (5a), and is sometimes missing in cases where it would normally be expected (5b).

(5) a *has passed urine only once throughout whole day,/ since Fri/ yes weds/seen dr this am*
b *pt requesting result of preg test result in mail box pt told neg*

In contrast, letters received from external agencies almost always use standard grammar and punctuation, and contain few spelling mistakes, idiosyncratic abbreviations, acronyms or shorthand expressions, as in (6a–c) below (unless the OCR software introduces mistakes).

(6) a *It might be worthwhile continuing with some regular Nasal Steroids but if there is a continued deterioration then she could try an alternative preparation such as Accolate 20mgs bd.*
b *She remained positive and was willing to discuss her problem and look at change. She is presently medication free and appears to be coping well.*
c *She does describe some chest tightness and the symptoms are certainly worse first thing in the morning but some of her symptoms would seem to be related to a musculoskeletal element of the anterior chest.*

The large majority of tools for grammatical analysis of text which have been built by the natural language processing research community are based on statistical models and lexicons trained on edited text genres such as newspaper articles or scientific papers. The vocabulary and the way language is used in these genres is very different to GP-written free text notes, and we have observed that standard NLP tools make many errors when applied to these notes. Retraining the tools on GP notes would be very expensive, since we would re-

quire access to significant amounts of text, which would have to be anonymised manually (to conceal all 'sensitive' information that could identify a patient) before it could be released by the GPRD and made available for part-of-speech or syntactic annotation. We therefore do not attempt any form of grammatical analysis.

2.2 Automatic Symptom Recognition

Koeling et al. [12] describe an investigation into automatic estimation of the incidence of symptoms using coded and free text information in primary care medical records. The experiments were run on records from the General Practice Research Database for 344 patients who were ultimately diagnosed as having ovarian cancer, and used an algorithm based on matching the textual descriptions of Read codes (for example 'abdominal pain') for the five most commonly presenting symptoms of the disease. The algorithm consists of three steps, performed in sequence:

1. Locate an occurrence of textual description of Read code in the text
2. Check whether there is evidence of negation
3. Determine whether the located textual description is within the scope of the negation

In the first step, sometimes an exact match of the textual description of the Read code is found in the text, but often the GP used a variant of the textual description. To deal with this, Koeling et al. manually compiled lists of common abbreviations of each word used in the Read code descriptions. Each list was augmented with a small number of semantically similar variants, selected from words which had a very similar distribution in medical record text. The algorithm allowed for spelling mistakes by matching words that are a small edit distance from the original word. For some symptoms the algorithm was able to double the amount of information extracted from the records compared to just considering coded information.

In Koeling et al.'s investigation, precision was optimised at the expense of recall; better results might therefore be obtained by improving the recall of the algorithm by producing more comprehensive lists of ways of expressing symptoms. Unfortunately, standard biomedical ontologies have poor coverage of many important phenomena in GP notes, especially abbreviations, acronyms, common shortenings of words and alternative spellings; they also do not cover informal but still medically-relevant language, as used by patients and reported by GPs (e.g. *tummy*). We are currently working on ways to automatically derive variations of surface realisations of words in order to obviate the need to compile lists of common variants manually, and to improve coverage of non-obvious variants. One approach we are exploring involves improving the process which identifies semantically similar variants of words so that it can find larger numbers of suitable candidates.

2.3 Distributional Similarity

It is often the case that semantically similar words are distributed similarly in text – that is, they occur in similar contexts. This idea goes back to observations by J. R. Firth who summarised it as "You shall know a word by the company it keeps" [4]. The *distributional similarity* of a pair of words is computed based on the shared contexts of the two words. Several measures of distributional similarity have been described in the literature. In the experiments described in this paper, we compute distributional similarity scores between words using Lin's measure [13]. We use the scores to create a *distributional thesaurus*, in which each word is associated with a list of other words with the highest distributional similarity scores.

More formally, to encode context information a word w is associated with a set of features, f, each feature having an associated frequency. Each feature is a pair $< r; x >$ consisting of a relation name[6] and a word x that is related to w via r. To create a distributional thesaurus we compute similarity of contexts for every pair of words in the free text, but limited to those words that have a total feature frequency of at least N.[7] A thesaurus entry of size k for some word w consists of the k most similar words to w.

Weeds and Weir [24] provide empirical insights into what makes a 'good' distributional similarity measure for semantic similarity prediction. They observe that weighting features by pointwise mutual information appears to be beneficial. The intuition behind this is that the occurrence of a less common feature is more important in characterising a word than a more common feature. For example, the verb *to eat* is more selective and tells us more about the meaning of its grammatical arguments than the verb *to be*.

3 Datasets

We have access to two datasets of free text records from which we can construct distributional thesauruses. The first of these datasets, described in Sect. 3.1 below, consists of a relatively small amount of manually anonymised data which we have stored within our institution. The other dataset (Sect. 3.2) is rather different, in that it is much larger and is a 'virtual dataset' of un-anonymised text which we cannot view in raw form and can only access for running experiments via an intermediary, due to reasons of confidentiality.

3.1 Anonymised Dataset

As part of an interdisciplinary research project, PREP[8], which is exploring the utility of free text in primary care medical records, we have been focussing on

[6] Previous studies have used grammatical relations (such as *subject* or *direct object*), or proximity relationships (such as *next word to the right*).

[7] For the experiments reported in this paper we set the frequency threshold N to 10 for smaller datasets, and 25 for larger ones.

[8] http://www.informatics.sussex.ac.uk/research/projects/PREP/

the records of women diagnosed with ovarian cancer. In this project, standard epidemiological methods were used to select a cohort of 344 patients and obtain from the General Practice Research Database all the records for these patients dating from 12 months before the diagnosis until 2 weeks after [20]. The resulting corpus consists of just over 6100 records, containing about 192K words. This corpus was manually anonymised by staff at the GPRD. Even though this dataset is large enough to answer many epidemiological questions, for most NLP purposes it is very small (especially considering the variety and amount of noise in the data). Fortunately, the GPRD have previously dealt with requests for anonymised free text and were also able to share with us text that had been anonymised for previous research projects. Even though this data is not relevant for studying ovarian cancer, it gives us more data that is representative of language in the database. The complete anonymised dataset contains around 3.5 million words, of which 3 million words are GP-written notes, and 500,000 words are letters.

3.2 Un-anonymised Dataset

Previous research into the distributional similarity technique has demonstrated that the quality of a distributional thesaurus improves as the amount of data it is derived from increases [2]. Given that manual anonymisation is expensive, it is unlikely that we will be able to obtain significantly more anonymised text in the near future. However, the full General Practice Rsearch Database is orders of magnitude larger than our anonymised dataset (and is growing all the time, as more records are collected from participating practices). We would therefore like to be able to draw on this much larger source of data in order to create thesauruses.

Another reason for wanting to use more data is that our 3.5 million words of anonymised text is a very small amount in comparison to previous work in distributional thesaurus building. Moreover, it contains less useful information for thesaurus building even than that figure might suggest. As Sect. 2.1 argues, the text is difficult to parse, and so the distributional similarity computation between two words should probably be with respect to the words that are proximate to both rather than the words that are grammatically related to both. McCarthy et al. [14] found that to obtain similar results, ten times more input data was needed when using proximity relationships compared with using grammatical relations.

A further problem with the anonymised dataset is the fact that it is not a random sample of the database. The records from which the dataset is derived were selected for a small set of studies – most of which were concerned with cardiovascular disease – so the dataset is biased.

These problems of size and bias inspired us to devise a better approach. Since thesaurus creation only needs to establish relationships between words in free text records and takes no account of the surface form of sensitive words (i.e. whether these words are anonymised or not), we can expect to obtain similar thesauruses from un-anonymised text as from anonymised text. So if we can arrange for the thesaurus building software to run at the GPRD, on a machine

behind their firewall, the free text records themselves do not need to leave GPRD premises. Instead of anonymising the *input* data, any identifiable information would be removed from the *output* thesaurus before it left the GPRD. The set of words comprising the thesaurus would be much smaller than the input data, so this would be a much cheaper exercise.

This novel approach allows us to use a much larger, balanced sample of text from the General Practice Research Database. We are currently running experiments with a random sample of records comprising 55 million words. We refer to this sample as the '*un-anonymised* dataset'.

4 Creating Distributional Thesauruses

We compute the distributional similarity of a pair of words based on the extent to which the words occur in the same contexts in some body of text. In order to be able to record the surrounding contexts for each word, we need to define the set of relations used to compute the features.

Many approaches to distributional similarity use sets of grammatical relations. The motivation is that, for example, if the word *codeine* appears in the direct object relation to the verb *prescribe*, it is likely that other words that appear in the direct object relation with *prescribe* are semanticaly close to the word *codeine*. The fact that two words share one particular grammatical relation might not be of much consequence, but the more grammatical relations two words share, the more likely it is that those two words have a related meaning. However, as discussed in Sect. 2.1, standard parsers perform poorly on GP-written free text notes. We therefore use an alternative, more robust way of establishing a relation between two words in which we define a window around each word and relate it to other words in terms of proximity. In view of the telegraphic style of GP-written notes, we use a small window. We disregard apparent sentence boundaries and consider all the words that appear within the window. The relationships we use are summarized in Table 1.

Table 1. Proximity relationships capturing context

prev	previous word
prev_window	word within a distance of 2–5 words to the left
next	next word
next_window	word within a distance of 2–5 words to the right

5 Experiments

5.1 Variant Realisations

Section 2.2 summarised our approach to symptom recognition, which requires the ability to recognise variants of words. A key resource for this is an automatically

built distributional thesaurus, which allows us to harvest variant realisations automatically from relevant free text, thus avoiding the time-consuming and error-prone task of manually compiling lists of variants. In future studies, as well as recognising symptoms, we also will want to recognise mentions of specific tests, diagnoses and treatments in a similar way.

In GPRD text records, many commonly occurring words have a large number of distinct realisations. Reasons include the use of abbreviations, spelling errors and idiosyncratic capitalisation and punctuation. For example, we have found 15 variants of the word *patient*, including *pat.*, *pat*, *Pt*, *pt.*, *pateint*, *aptient* and *ptient*. Even though some of these are clearly unintentional misspellings, they are used in sufficiently consistent ways to be identified on the basis of the contexts in which they appear.

An example of how distributional methods can be used to identify candidate variants is the thesaurus entry for *abx* (a common abbreviation for *antibiotics*), shown in Fig. 1. The column to the right of the entry (*abx*) contains a list of the twenty words scored as most similar to *abx*. The numbers in the next column are scores that indicate the degree of similarity. The first thing we note about the related words in Fig. 1(a), is that the most similar word is the word that is being abbreviated (*antibiotics*). Even though the rest of the list does not contain many highly relevant words for this purpose, most of the words are related in one way or another. The list in Fig. 1(b) gives a strong indication that more data results in a better quality thesaurus. Out of the twenty words, more than half are variant realisations of *abx*, and half of the remainder are names of specific antibiotics.

It turns out that *abx* is a very frequently occurring term. Less frequently used terms do not always end up with such accurate results. However, the thesaurus usually gives an acceptable pre-selection of candidate variants. We are working on other methods to distinguish between the full version of an abbreviated word, other surface variations, related words and unrelated words.

5.2 Related words

In addition to identifying variant realisations of a single word, some information extraction tasks might require the ability to recognise sets of *related* words expressing qualities or attributes of a clinically-relevant entity (such as a symptom, part of the body, or mental state). In contrast to symptoms, tests, diagnoses and treatments which are typically nouns, such words would usually be adjectives. In general, adjectives are more polysemous than nouns;[9] we might therefore expect slightly lower quality distributional thesaurus entries for adjectives, since polysemy is acknowledged problem for distributional approaches [5].

For example, we might need words related to the attribute *swollen*, such as *bruised*, *enlarged*, *inflamed*, *painful*, *red* and *sore*. Figure 2 shows thesaurus entries for this word; comparing (a) and (b), it is evident that the larger unanonymised dataset again gives good quality results.

[9] In WordNet 3.0, the average polysemy of nouns is 1.24, whereas that of adjectives is 1.40 (http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html).

abx	abx	1.0	abx	abx	1.0
	antibiotics	0.1231		abs	0.1574
	antibiotics.	0.1102		antibiotics	0.1509
	calpol	0.1088		ab	0.1416
	msu	0.1064		abx,	0.1330
	fluids	0.1053		a/b	0.1313
	steroids	0.1042		abx.	0.1303
	diuretic	0.1039		antibiotic	0.1276
	pred	0.1018		amoxil	0.1252
	diuretics	0.1016		amox	0.1248
	treat	0.0992		ab's	0.1241
	rx.	0.0988		amoxicillin	0.1238
	uti	0.0985		erythromycin	0.1196
	meds	0.0981		calpol	0.1181
	infection	0.0978		steroids	0.1178
	analgesia	0.0963		fluclox	0.1159
	observe	0.0958		treat	0.1158
	1/52	0.0949		analgesia	0.1150
	ibuprofen	0.0940		Abx	0.1136
	tomorrow	0.0936		antibiotics.	0.1099
	sos	0.0930		ABs	0.1089
	(a)			(b)	

Fig. 1. Distributional thesaurus entries for *abx*, derived from (a) the 3.5 million word anonymised dataset, and (b) the 55 million word un-anonymised dataset

For other information extraction-type tasks, we might want to relate words expressing similar qualities – for example a system that recognised whether a patient has reported pain (through words such as *discomfort*, *ache* or the word *pain* itself) would also probably need to determine gradation in the level of pain (from slight to severe), since *severe discomfort* might be of interest whereas *minimal pain* might not be. Figure 3 shows the thesaurus entry for *slight* derived from the 3.5 million word anonymised dataset. Since this is a small dataset, it should be expected that some of the words with high similarity scores are not relevant to this type of gradation; however, it is surprising that although in general most words are distributionally very similar to their antonyms, in this entry all of the relevant words are close in meaning to *slight* and there are no antonyms or close antonyms.

6 Discussion and Future Work

GP-written free text notes differ in many ways from the types of edited text that are commonly used in the natural language processing research community. Standard tools for grammatical analysis of text in general give poor results when applied to GP notes. However, our experiments into creating distributional similarity thesauruses from this text give promising results, and are able to

swollen	swollen	1.0
	swelling	0.1467
	painful	0.1458
	red	0.1440
	thigh	0.1406
	ankles	0.1291
	leg	0.1233
	hot	0.1200
	legs	0.1190
	tender	0.1183
	sob	0.1144
	ankle	0.1130
	swollen,	0.1090
	unwell	0.1087
	sore	0.1085
	oedema	0.1080
	dry	0.1055
	calf	0.1051
	crying	0.1046
	worse	0.1040

(a)

swollen	swollen	1.0
	painful	0.1542
	inflamed	0.1416
	red	0.1383
	swelling	0.1365
	inflammed	0.1357
	red,	0.1306
	swollen,	0.1251
	sore	0.1222
	redness	0.1179
	infected	0.1175
	slightly	0.1171
	itchy	0.1158
	sl	0.1141
	tender	0.1140
	hot	0.1140
	swelling,	0.1119
	enlarged	0.1105
	tonsils	0.1091
	swollen.	0.1082

(b)

Fig. 2. Distributional thesaurus entries for *swollen*, derived from (a) the 3.5 million word anonymised dataset, and (b) the 55 million word un-anonymised dataset; in GP notes *sob* is an acronym for 'shortness of breath', and *sl* is commonly used to abbreviate 'slight' or 'slightly'

slight	slight	1.0
	some	0.1291
	sl	0.1246
	mild	0.1199
	cough	0.1191
	c/o	0.1096
	slightly	0.1084
	minimal	0.1074
	swelling	0.1072
	ankle	0.1065
	little	0.1047
	tender	0.1027
	dry	0.0993

Fig. 3. Distributional thesaurus entry for *slight*, derived from the anonymised dataset; *c/o* is a standard abbreviation in GP notes for 'complains of'

extract lexical information that that is suitable for clinical text mining tasks, and which cannot be obtained from standard resources such as (bio)medical lexicons or ontologies.

Our successful use of distributional similarity for unsupervised lexical acquisition in the medical domain accords with other recent research efforts. A number of these have focussed on organising biomedical terminology with respect to (bio)medical ontologies and encyclopedias. In particular, Weeds et al. [23] apply distributional techniques to determine semantic proximity in order to classify terminology drawn from the GENIA corpus of biomedical research abstracts with respect to an associated ontology of terminological types. Fan and Friedman [3] reclassify UMLS concepts into broader semantic classes, using text from MEDLINE/PubMed titles and abstracts to compute distributional information. Van der Plas and Tiedemann [17] describe a system for identifying variants of Dutch terms in a medical encyclopedia using statistical information extracted from raw text, using word-aligned parallel corpora to establish co-occurrence relationships between terms in Dutch and their translations.

The study that is probably the most similar to ours – in that it is concerned with clinical text rather than edited biomedical text – is that of Henriksson et al. [7]. They describe an approach to assigning ICD-10 codes for diagnoses to uncoded medical records in Swedish which comprise text that is often semi-structured, but still contains many typing errors and non-standard abbreviations. Their approach computes word and code co-occurrences at the document level in order to capture information about the semantic similarity of individual words and codes, which is then used to classify uncoded documents.

The main novel aspect of our work is a new approach for acquiring lexical information from sensitive text which does not require the text to be anonymised before processing. This makes it possible to create thesauruses from un-anonymised text, and thus process much larger quantities of data than would normally be the case. However, even the un-anonymised dataset we are using is much smaller than the corpora of general text used in previous investigations into distributional similarity, which have shown that the larger the corpus the higher the quality of the resulting thesaurus [2, 14, 24]. We therefore intend to scale up this aspect of our processing – although at a purely practical level it does depend on having access to sufficiently powerful remote computing infrastructure, which in turn relies on purchasing decisions outside our control. In addition, although this approach means that the input text does not need to be anonymised, any identifiable information must removed from the output thesaurus before it leaves the GPRD. While this should be relatively cheap exercise, we are currently investigating a set of safeguards which might mean that even this step might be unnecessary.

Previous work has found that distributional thesauruses can reflect latent aspects of the text they were built from. In particular, Koeling et al. [11] demonstrate that differences in the most frequent meaning of an ambiguous word between domains (for example the meanings of *bypass* in medical text and in current affairs news articles) can be predicted by creating separate distributional

thesauruses from documents in each domain. One of the reasons why this works is that for many words whose predominant meaning changes between domains, the most similar words in a thesaurus are specific to the domain (for example *catheter* might be similar to *bypass* but would only appear in medical text). This observation is relevant to our processing of clinical free text. As mentioned in Sect. 3.1, the dataset of anonymised text is a by-product of a relatively small number of research projects, and contains a large proportion of text relating to cardiovascular disease and prostate cancer. If we examine the thesaurus that was built using this dataset we can detect a bias towards these diseases. Although this is a weakness in our current experiments, we may be able to take advantage of it start processing larger amounts of GPRD data. We intend to explore how we can specialise our thesauruses for certain disease areas and see if it improves the utility of the resulting thesauruses.

In this paper we have not carried out an objective assessment of the thesauruses we have created, nor have we performed a quantitative evaluation of how they can contribute to a relevant application task. One of our next steps will be to conduct an extrinsic evaluation of thesaurus data applied to a symptom recognition problem. We already have a suitable prototype system (outlined in Sect. 2.2) which we can adapt, and also an annotated corpus [10]. This should constitute a good test of our techniques.

Acknowledgments. We are grateful to Jackie Cassell, Clare Laxton and Rosemary Tate for advice and suggestions on the direction of this research. The work was supported by the Wellcome Trust [086105/Z/08/Z]. Access to the GPRD database was funded through the Medical Research Council's licence agreement with MHRA. The authors were independent from the funder and sponsor, who had no role in conduct, analysis or the decision to publish. This study is based in part on data from the Full Feature General Practice Research Database obtained under licence from the UK Medicines and Healthcare Products Regulatory Agency. However, the interpretation and conclusions contained in this study are those of the authors alone.

References

1. Bentley, T., Price, C., Brown, P.: Structural and lexical features of successive versions of the Read Codes. In: Teasdale, S. (ed.) Proceedings of the Annual Conference of The Primary Health Care Specialist Group of the British Computer Society. pp. 91–103. Worcester, UK (1996), http://www.phcsg.org/main/pastconf/camb96/readcode.htm
2. Curran, J., Moens, M.: Scaling context space. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 231–238. Philadelphia, PA (2002)
3. Fan, J.W., Friedman, C.: Semantic classification of biomedical concepts using distributional similarity. JAMIA 14(4), 467–477 (2007)
4. Firth, J.R.: A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis pp. 1–32 (1957)

5. Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., Wang, Z.: New experiments in distributional representations of synonymy. In: Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL). pp. 25–32. Ann Arbor, MI (2005)
6. Hamilton, W., Peters, T., Bankhead, C., Sharp, D.: Risk of ovarian cancer in women with symptoms in primary care: population based case-control study. British Medical Journal 339, b2998 (2009)
7. Henriksson, A., Hassel, M., Kvist, M.: Diagnosis code assignment support using random indexing of patient records a qualitative feasibility study. In: Proceedings of AIME, the 13th Conference on Artificial Intelligence in Medicine. pp. 348–352 (2011)
8. Johansen, M., Scholl, J., Hasvold, P., Ellingsen, G., Bellika, J.: "Garbage in, garbage out" – extracting disease surveillance data from EPR systems in primary care. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work. pp. 525–534. San Diego, CA (2008)
9. Kalra, D., Ingram, D.: Electronic health records. In: Zielinski, K., Duplaga, M., Ingram, D. (eds.) Information Technology Solutions for Healthcare. Springer-Verlag, http://eprints.ucl.ac.uk/1598/ (2006)
10. Koeling, R., Carroll, J., Tate, A.R., Nicholson, A.: Annotating a corpus of clinical text records for learning to recognize symptoms automatically. In: Proceedings of the 3rd Louhi Workshop on Text and Data Mining of Health Documents. pp. 43–50. Bled, Slovenia (2011)
11. Koeling, R., McCarthy, D., Carroll, J.: Domain-specific sense distributions and predominant sense acquisition. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 419–426. Vancouver, Canada (2005)
12. Koeling, R., Tate, A.R., Carroll, J.: Automatically estimating the incidence of symptoms recorded in GP free text notes. In: Proceedings of the First International Workshop on Managing Interoperability and Complexity in Health Systems. pp. 43–50. Glasgow, UK (2011)
13. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the ACL. pp. 768–774. Montreal, Canada (1998)
14. McCarthy, D., Koeling, R., Weeds, J., Carroll, J.: Unsupervised acquisition of predominant word senses. Computational Linguistics 33(4), 553–590 (2007)
15. NIST: Proceedings of the 2011 Text REtrieval Conference (TREC 2011). National Institute for Standards in Technology, Gaithersburg, MD (2011)
16. Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K.B., Duch, W.: A shared task involving multi-label classification of clinical free text. In: Proceedings of BioNLP 2007: Biological, Translational, and Clinical Language Processing. pp. 97–104. Prague, Czech Republic (2007)
17. van der Plas, L., Tiedemann, J.: Finding medical term variations using parallel corpora and distributional similarity. In: Proceedings of the 6th Workshop on Ontologies and Lexical Resources. pp. 28–37. Beijing, China (2010)
18. Resnik, P., Niv, M., Nossal, M., Kapit, A., Toren, R.: Communication of clinically relevant information in electronic health records: a comparison between structured data and unrestricted physician language. Perspectives in Health Information Management (2008)
19. Roberts, A., Gaizauskas, R., Hepple, M., Guo, Y.: Mining clinical relationships from patient narratives. BMC Bioinformatics 9(Suppl 11), S3 (2008)

20. Tate, A.R., Martin, A., Ali, A., Cassell, J.: Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational study using primary care records of patients with ovarian cancer. BMJ Open doi:10.1136/bmjopen-2010-000025 (2011)
21. Uzuner, Ö., Goldstein, I., Luo, Y., Kohane, I.: Identifying patient smoking status from medical discharge records. JAMIA 15(1), 14–24 (2008)
22. Uzuner, Ö., Solti, I., Cadag, E.: Extracting medication information from clinical text. JAMIA 17(5), 514–518 (2010)
23. Weeds, J., Dowdall, J., Schneider, G., Keller, B., Weir, D.: Using distributional similarity to organise biomedical terminology. Terminology 11(1), 107–141 (2005)
24. Weeds, J., Weir, D.: Co-occurrence Retrieval: a flexible framework for lexical distributional similarity. Computational Linguistics 31(4), 439–476 (2005)