Annotating a corpus of clinical text records for learning to recognize symptoms automatically

Rob Koeling[1], John Carroll[1], A. Rosemary Tate[1], and Amanda Nicholson[2]

[1] School of Informatics, University of Sussex, Brighton, UK
{robk, johnca, rosemary}@sussex.ac.uk
[2] Brighton and Sussex Medical School, Brighton, UK
A.C.Nicholson@bsms.ac.uk

Abstract. We report on a research effort to create a corpus of clinical free text records enriched with annotation for symptoms of a particular disease (ovarian cancer). We describe the original data, the annotation procedure and the resulting corpus. The data (approximately 192K words) was annotated by three clinicians and a procedure was devised to resolve disagreements. We are using the corpus to investigate the amount of symptom-related information in clinical records that is not coded, and to develop techniques for recognizing these symptoms automatically in unseen text.

1 Introduction

UK primary care databases provide a valuable source of information for research into disease epidemiology, drug safety and adverse drug reactions. Analyses of existing large-scale electronic patient records held in the form of large primary care datasets such as the General Practice Research Database have almost exclusively exploited coded data. Such data are readily accessible to the classical methods of epidemiological analysis, once the complexities of defining and selecting a patient cohort have been overcome. However, since clinicians can choose to what extent they code a consultation, an unknown amount of clinical data is not coded, and 'hidden' in free text. Free text records often contain important information on the severity of symptoms or on additional symptoms which have not been coded [6, 3]. The degree to which clinical information is coded and how this varies between by practitioner, practice, or type of clinical problem is currently unknown, as is the impact on public health research results of not using information in free text. The aims of our work are to quantify how much additional information is in the free text and to explore methods for extracting it.

Automatic extraction of complex information from notes written by general practitioners, which may be ungrammatical and often contains ambiguous terms, misspellings and abbreviations, is a very challenging natural language processing (NLP) task. The text is much less uniform than data typically analysed by the NLP research community, and issues of confidentiality make it difficult to gain access to significant amounts of data.

2 The Data

This study builds on previous work [10], in which we used coded records from the General Practice Research database (GPRD [1]) of 344 patients between 40 and 80 years of age (inclusive) diagnosed with ovarian cancer between 1 June 2002 and 31 May 2007. The records use the Read coding system, which was originally developed in the 1980s and is used throughout the United Kingdom for coding clinical events in primary care. Each Read code has an associated textual description e.g. 'Abdominal pain', 'Right iliac fossa pain', 'Constipation', which are available on GP systems as an aid for recording the correct code. In the current study we obtained manually anonymized free text records of all 344 patients for the period 12 months prior to the date of definite diagnosis.

The free text records contain information from a variety of different sources. Mostly they consist of notes typed by the GP during or after a consultation, communication with secondary care (for example referral letters and discharge summaries), and sometimes test results. However, about 90% of the records contain notes typed by the GP, which turn out to be the most challenging category. A typical example is:

> *5 day Hx of umbilical Dx. Smelly Dx. Red inflammed lump within umbilicus. Swab sent. Try fluclox and review. Also ?mass felt left lower abdo. No weight loss, bowels reg. No Jaccol.*

Some differences between this data and standard English are: (1) inconsistent use of capitalization and punctuation; (2) spelling errors and unusual abbreviations, acronyms, and named entities; (3) anomalous tokenization (e.g. missing spaces); and (4) ambiguous use of question marks. These characteristics make it difficult to process the data automatically. They also impact on readability, making human annotation more time-consuming, and therefore costly.

3 The Annotation Process

In order to annotate symptoms associated with ovarian cancer in the free text fields of patient records, we first identified the most commonly experienced ovarian cancer symptoms. Table 1 lists these symptoms, which were taken from a recent paper by Hamilton [3].

To facilitate the work of the annotators, we created an easy to use interactive annotation system. We used the Visual Tagging Tool (VTT), part of the SPECIALIST NLP Tools [2]. VTT allowed us to create an environment in which an annotator can highlight a phrase in the text and choose, from a pull down menu, the most appropriate tag to describe the symptom they had highlighted. A screenshot of the annotation workbench is shown in Figure 1.

We drafted a detailed set of annotation guidelines. Over several iterations we refined the guidelines to minimize potential disagreement between the annotators — learning from others' experiences in defining a methodology for annotating clinical data [8]. The guidelines ask annotators to:

Table 1. Symptom categories that were annotated

Main symptom category	Subcategory
Abdominal pain	
Pelvic pain	
Back pain	
Abdominal distension NOS	Ascites Bloating
Indigestion NOS	Flatulence Dyspepsia
Nausea & vomiting	
Change in bowel habit NOS	Constipation Diarrhoea
Genito-urinary symptoms NOS	Genital inc bleeding Urinary
Abdominal / pelvic mass	
Weight / appetite change NOS	Weight loss / appetite down Weight gain
Tiredness	
Breathing problems	

identify all the different expressions (words or strings of words) in the notes that represent a symptom from the list of pre-defined relevant symptoms for ovarian cancer. These expressions can be complaints expressed by patient, signs detected on examination or by investigation or findings at operation. All of these should be marked as long as they refer to one of the symptoms in the pre-defined list.

The main supporting instructions are:

- **Do not infer a symptom from the text**: only annotate symptoms if found as such in the text.
- **Presence or absence**: e.g. in the case of 'There was no evidence of abdominal distension', 'abdominal distension' should be marked up.
- **Annotate the bare minimum**: only annotate as much text as you need to identify the symptom. E.g. '...who has had right upper quadrant abdominal pain now for some weeks', only 'abdominal pain' should be marked up.
- **Annotate every occurrence of a symptom in a record**.

Five individuals with a medical background were involved in developing the guidelines, iteratively refining them on the basic of exploratory annotation sessions. Once the final version of the guidelines was produced, three people each annotated the whole dataset independently. One of the annotators was also involved with developing the guidelines. All the annotators have a medical background, either as a General Practitioner, a researcher with GP training, or a

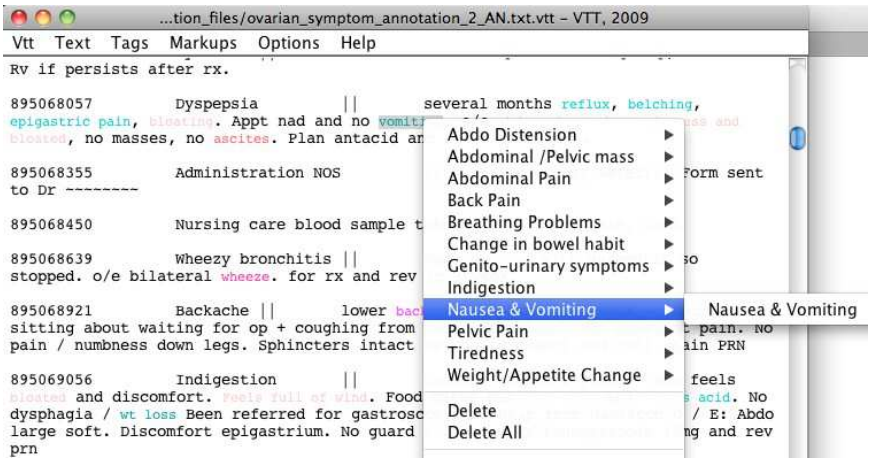

Fig. 1. Screenshot of the annotation work bench

final-year medical student. The annotators were given a copy of the annotation software and worked with it at their convenience. The data was presented to them in batches of about 800 records each (but there was no requirement to finish a batch in a single session).

4 The Corpus

We started with 6141 records, determined by the amount of data available for the 344 patients for the period 12 months prior to the date of definite diagnosis (Section 2). After these records were annotated by the three annotators, we created a gold standard corpus. During the development of the guidelines and exploratory annotation sessions we noticed that there were two main areas of disagreement between annotators. Firstly, annotation is a mentally strenuous task, and as a result annotators occasionally miss symptoms, especially when a record contains a large number of them. Secondly, disagreements arise from the fact that annotators are free to decide the points at which each marked-up symptom starts and ends.

Inter-annotator agreement [4] is an important quality measure. The standard metric for inter-annotator agreement for categorization tasks with two annotators is the kappa statistic, defined as $k{=}P(a)-P(e)/1-P(e)$, where $P(a)$ is the measured probability of agreement between annotators, and $P(e)$ is the probability that agreement is due to chance. However, a complicating factor for our annotation task is that categorization is only one element of the task. The other element is the choice of the *boundaries* of the expression that is associated with a certain class. The set of possible expressions to annotate is extremely large, which makes it impossible to estimate $P(e)$. This issue arises whenever the set

of elements that has to be marked up is not fixed. This has also been noted in annotation efforts that included similar tasks, such as [8] and [9].

In our setup it is difficult to calculate an exact figure for agreement other than the most basic one. The most straightforward measure is *strict* agreement, defined as the proportion of all cases where all three annotators were in full agreement (those cases where the annotated string starts and ends at exactly the same position in the text and the string is assigned the same label). We found three-way strict agreement in about 62% of cases. This figure is difficult to compare with other work reported in the literature. The closest comparison we are aware of is the inter-annotator agreement for finding 'Signs and Symptoms' in [9]. They report an F-measure (a combination of precision and recall) of 0.61 for double annotated text. However, on the one hand their task was more complex (the text was annotated for several aspects at the same time), but on the other hand the text we annotated is less like standard English, and we had it triple annotated. Considering these factors we were pleasantly surprised by the annotation agreement figure, especially since inspection of the remaining cases suggested that many could be resolved without much effort.

Both [8] and [7] propose second, less strict measure of annotation accuracy that allows for partial matching of annotated strings. This is well-defined for double annotation, but difficult to adapt to triple annotated data. We therefore decided to stick to the strict agreement measure for our data.

In order to maximize the quality of the gold standard, we had to decide on a method for combining the data individually created by the annotators. A typical way of resolving disagreement between annotators is to have the data double annotated and appoint a third annotator to choose between the two conflicting opinions. However, this method does not allow for discussing cases that are inherently difficult. We therefore decided to keep the triple annotation, and employ a variant of the Delphi method [5] to generate consensus. In the Delphi method each annotator works individually, gets feedback on the cases where there is disagreement, and is asked to revisit those cases. We produced an overview of the disagreements and asked the three annotators to come back to discuss and resolve them.

In order to use the annotators' time efficiently, we created five categories from the cases of disagreement, each of which could be approached differently. The categories were:

1. **Trivial difference**: e.g. 'Abdominal pain' vs. 'Abdominal pai'
2. **Added modifier**: e.g. 'slightly bloated abdomen' vs. 'bloated abdomen'
3. **2 agree/1 missed out**: one annotator may have overlooked a symptom
4. **2 agree/1 disagree**: with respect to the span of the string or associated label
5. **Rest**: all other instances

The first two categories, accounting for around 20% of the disagreements, were easy to resolve. The third and fourth categories had to be checked one by one by the annotators, but were mostly easily resolved. The **Rest** category was the most challenging. Many cases in this category had multiple disagreements (over

Table 2. Numbers of occurrences ('tokens') and distinct expressions ('types') by symptom category

Symptom	# tokens	# types	Mean # tokens per type
Abdominal pain	565	233	2.42
Pelvic pain	67	50	1.34
Back pain	69	42	1.64
Abdominal distension	598	159	3.76
Indigestion	134	53	2.53
Nausea & vomiting	339	52	6.52
Change in bowel habit	452	124	3.65
Genito-urinary symptoms	426	222	1.92
Abdominal / pelvic mass	661	344	1.92
Weight / appetite change	322	127	2.54
Tiredness	73	20	3.65
Breathing problems	249	85	2.92

both the span of the string and the associated tag) and often the annotators had to view the full context of the annotated string to reach agreement.

The resulting annotated corpus consists of a total of 6141 records, containing about 192K words. The total number of annotated symptoms is 3955. Table 2 summarizes the symptoms labeled in the corpus. Even though the average number of occurrences of each distinct expression is low, the distribution is very skewed (see Figure 2). For example, 'abdominal swelling' is an expression that the annotators label as 'Abdominal Distension'; there are ten occurrences ('tokens') of this expression ('type') in the corpus. However, there is great variation in the expressions used to describe the same symptom. The more formal ones are used frequently, whereas informal expressions (e.g. 'tummy is strikingly larger' or 'abdo looks sl swollen') often occur only once.

In one experiment we looked at a subset of symptoms, covering about 3281 annotated expressions. Approximately 1200 of these occur only once in the corpus. Figure 2 shows that the most frequent 100 types account for almost 65% of the tokens. This observation is important for designing an automatic system that finds this information in the text. If recall is not an absolute priority, then it is possible to get a long way by concentrating on the high frequency types. Moreover, the most frequently occurring expressions are generally more formal and concise, which would also be of assistance. However, a system would also need to integrate relevant contextual factors, in particular whether indications of symptoms are negated or are attributed to someone other than the patient.

5 Discussion

The corpus we have produced is a rich resource, which we are using for two purposes. Firstly, we are investigating the amount of symptom-related information

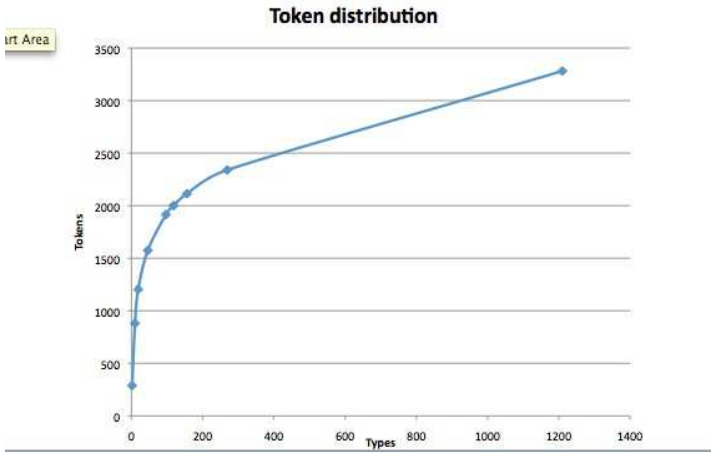

Fig. 2. Cumulative number of annotated tokens, ordered by type frequency

available in the free text fields of primary care patient records. By comparing symptom annotations with coded information in the same records we are exploring the hypothesis that a significant amount of information is missed when the contents of free text fields are not taken account of in epidemiological research. Secondly, we are starting to use the data for creating and evaluating techniques for automatic recognition of symptoms in free text. Although we are ultimately interested in developing machine learning-based models that precisely capture as wide a variety of symptoms as possible, the annotated corpus also allows us to estimate the utility of unsophisticated techniques such as approximate string matching and thesaurus-based expansion. If significant amounts of information can be uncovered with such methods, then the epidemiological research community would not require specialised natural language processing expertise to be able to exploit free text resources. Automatic processing of information in free text fields also opens up opportunities to work with un-anonymized data. Restricted access to textual data is a major hurdle in research using electronic patient records. The possibility of retrieving information without the need for manual anonymization would open up many new opportunities.

The process of establishing consensus between annotators has given us new insights into the nature of the data and has highlighted issues that need to be addressed when processing this data and when defining further annotation tasks. One of the main issues relates to symptoms recorded as a result of a complaint by the patient versus the outcome of an examination or a test result. From a clinical point of view, a complaint has a different status to an examination result. Even though many potential disagreements of this type were resolved in the course of developing the annotation guidelines, new issues will crop up when annotating large amounts of text. Decisions about how to deal with these cases might need to be informed by the research question being addressed.

There is evidence that important information is missed when epidemiological research using patient records relies only on coded information. The research reported here is a step towards quantifying how much additional information is potentially available, and is a prerequisite for research into automatic retrieval of this information and making it available to the research community.

Acknowledgments This work was supported by the Wellcome Trust [086105] (*The ergonomics of electronic patient records: an interdisciplinary development of methodologies for understanding and exploiting free text to enhance the utility of primary care electronic patient records ("PREP")*). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. GPRD: GPRD. Excellence in public health research. http://www.gprd.com) (2009)
2. Group, L.S.: http://lexsrv3.nlm.nih.gov/LexSysGroup/Home/ (2010)
3. Hamilton, W., Peters, T.J., Bankhead, C., Sharp, D.: Risk of ovarian cancer in women with symptoms in primary care: population based case-control study. British Medical J. 339 (AUG 25 2009)
4. Hripcsak, G., Rothschild, A.S.: Agreement, the f-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association 12, 296–298 (2005)
5. Hripcsak, G., Wilcox, A.: Reference standards, judges, and comparison subjects. Journal of the American Medical Informatics Association 9 (2002)
6. Johansen, M.A., Scholl, J., Hasvold, P., Ellingsen, G., Bellika, J.G.: "Garbage In, Garbage Out" - Extracting Disease Surveillance Data from EPR Systems in Primary Care. pp. 525–534. ACM; ACM SIGCHI, ACM (2008), ACM Conference on Computer Supported Cooperative Work, San Diego, CA, NOV 08-12, 2008
7. Pyysalo, S., Ginter, F., Heimonen, J., Bjrne, J., Boberg, J., Jrvinen, J., Salakoski, T.: Bioinfer: a corpus for information extraction in the biomedical domain. BMC BioInformatics 8 (2007)
8. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., Setzer, A.: Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics 42 (2009)
9. South, B.R., Shen, S., Jones, M., Garvin, J., Samore, M.H., Chapman, W.W., Gundlapalli, A.V.: Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics 10 (2009)
10. Tate, A.R., Martin, A.G.R., Murray-Thomas, T., Anderson, S.R., Cassell, J.A.: Determining the date of diagnosis - is it a simple matter? The impact of different approaches to dating diagnosis on estimates of delayed care for ovarian cancer in UK primary care. BMC Medical Research Methodology 9 (June 2009)