Automatic rewriting of patient record narratives

Catalina Hallett, David Hardcastle

The Open University
Walton Hall, Milton Keynes, UK
c.hallett@open.ac.uk, d.w.hardcastle@open.ac.uk

Abstract

Patients require access to Electronic Patient Records, however medical language is often too difficult for patients to understand. Explaining records to patients is a time consuming task, which we attempt to simplify by automating the translation procedure. This paper introduces a research project dealing with the automatic rewriting of medical narratives for the benefit of patients. We are looking at various ways in which technical language can be transposed into patient-friendly language by means of a comparison with patient information materials. The text rewriting procedure we describe could potentially have an impact on the quality of information delivered to patients. We report on some preliminary experiments concerning rewriting at lexical and paragaph level. This is an ongoing project which currently addresses a restricted number of issues, including target text modelling and text rewriting at lexical level.

1. Introduction

Allowing patient access to (electronic) records (EPR) in a comprehensive format is a legal requirement in most European countries. Traditionally, providing explanations of EPRshas been the responsibility of the medical units that hold and release the information (GP surgery, hospital, medical trust), which is an expensive and time-consuming task, resulting in a significant delay in delivering this information to patients. With the advent of online access to EPRs, patients will be able to access their records on demand, therefore explanations should also be provided on the fly. In addition to the legal requirement to release information to patients in a form that they can understand, research shows that personalised reports may improve the quality of care(Detmer and Singleton, 2004). Although targeted at producing personalised informational materials, rather than medical reports, several projects have found positive effects of information personalisation in, among others, asthma treatment (Osman et al., 1994) and cancer management (Jones et al., 2006).

Whilst most work on medical report generation systems has concentrated on explaining the structured part of EPR((McKeown et al., 2002),(Torgersson and Falkman, 2002), (Hüske-Kraus, 2003)), there has been very little work on providing automatic explanations of the narratives (letters between health practitioners, hospital discharge notes, notes attached to intervention records) which represent a considerable part of an EPR. Attempting to rewrite narratives in a patient-friendly way is in many ways more difficult than providing suggestions for a natural language generation system that takes as input data records. In text, ambiguity can arise not only from lexical choice but also from aspects over which a natural language generation system has full control, such as syntax, discourse structure, sentence length, formatting and readability.

The research reported in this paper attempts to partially address a gap in the identification of issues likely to impede the patients' understanding of medical records. We are taking a corpus analysis approach in an attempt to answer the following research questions:

- Given the text-based part of a patient record, which segments require explanation before being released to patients?

- Which types of rewriting are appropriate for various text segments that display technical features?

Textual explanations can be constructed using various techniques: text-to-text generation or paraphrasing, inline addition of canned text explanations, linking to trusted medical information sources, information extraction followed by generation. Whilst on long term we envisage being able to provide explanations using a combination of such techniques, in the initial stages we will aproach the problem as a text re-writing exercise, i.e. producing the same type of textual narrative as the input narrative.

Our approach relies on the identification of features that differentiate between medical texts written for doctors and medical texts produced for lay readers. By comparing a corpus of expert text and a corpus of lay texts, we identify a set of features that differentiate medical expert and lay language. We then proceed to use the selected features on a corpus of narratives extracted from a repository of EPRs.

In the first instance, larger sections of text (paragraphs) that contain features characteristic to expert documents are highlighted. A corpus of patient information leaflets is used for tuning the segment selection procedure. Secondly, we identify within those sections the features that contribute to the classification of the section as belonging to the expert register. Thirdly, based on the selected features we provide suggestions for text simplification. The text rewriting process kicks off with the generation of lay equivalents of medical terms, followed by rewriting at sentence and paragraph level.

2. Data profiling

The aim of the data profiling exercise is to inform a natural language generation (NLG) system. The type of information that NLG system developers look for in target texts varies widely according to factors such as the register (a system for generating weather reports will most likely look for the frequency and way of presenting numerical

data, while a decision support system will look for means of presenting logical statements) and intended audience of the system, the capabilities of the system (e.g., if the system is template-based, the developers will not look for complex discourse structures), and ultimately, according to the subjective interest of the system developer in one or another linguistic or textual phenomenon.

A major factor in choosing the exact features to be investigated is the fact of whether the corpus under investigation contains relevant target texts or only texts resembling the intended output to a certain degree. For example, the intended target text may be a software technical manual, but the developer relies on a corpus of technical instruction manuals for white goods. In this case, it is likely that the developer will be able to learn facts about the structure and style of the document and the nature of the argumentation, as well as some generic facts about the lexical content of the text (for example, frequency of technical terms). However, they cannot rely on observations regarding the frequency of individual words, paraphrasing techniques or content.

In attempting to rewrite reports for patients, we are faced to similar challenges. The overall content of such reports is completely driven by, firstly, the underlying data (the electronic patient record) and secondly by completeness and correctness requirements (no data should be omitted and no extra information should be inferred). However, the manner of conveying this information has to be learned from similar reports. Leaving aside the problem of user-centered generation, this type of report has to take into account features such as complexity and readability, which directly influence verbosity and formatting - since electronic patient records contain complex medical terms, as well as numerical values, how much of this information has to be explained to patients and to which level of detail? In this particular application, obtaining similar reports is especially difficult; strict confidentiality laws governing the disclosure of patient records means that a corpus of reports is not readily available. Constructing one (or even a limited number of reports) is equally prohibitive due to the high cost of involving medical experts. It is however possible to infer, at least partially, some specific features of texts written for patients by analysing materials which, though different in content, are aimed at the same audience and belong to the same domain: in our case, patient information leaflets distributed through hospitals and cancer charities.

In our previous work(Kokkinakis et al., 2007; Hardcastle and Hallett, 2007), we have identified a set of features that differentiate technical and lay medical texts. Our conclusions were drawn from a 3-way corpus comparison where we analysed medical documents written by doctors for doctors, documents written by medical professionals for patients and documents written by patients for patients. Each corpus was constructed from online documents and contained approximately 200,000 words.

In the current project, we are using a selection of these features to construct a model of lay language that we could use as a reference model for text rewriting. Whilst, as previously mentioned, it is generally accepted that medical materials tailored to individual patients are likely to have a greater impact on the level of delivered healthcare, our approach is targeted at generic patients, which we model on the typical reader of patient information leaflets. We construct the model by analysing a 250,000 words corpus of patient information leaflets which we collected online. The choice of corpus was based on our previous experience in analysing medical documents. Patient information leaflets are meant for a wide readership, making only generic assumptions about the reading proficiency of their target audience. They also tend to be homogenous in style, which allows us to use a much smaller corpus that would be otherwise necessary to obtain a correct language model. We will further present succintly the features investigated.

Complexity and readability: The readability of a text is especially important when generating texts for a particular, well defined type of audience - for example, children or people with learning difficulties. We compute measures such as : length of document, sections, paragraphs, sentences and words, percentage of complex words, abbreviations, numerical values, punctuation, readability scores (FOG, FLESCH).

Syntax: We look at a variety of syntactic features, such as verb voice and mood, person, length of noun phrases, frequency of nominalisations and verbalisations.

Lexical content: In this category, we look at the frequency of complex terms in general, the frequency of medical terms and loan words. The analysis of the medical content is based on the MeSH terminology (Canese, 2003) and is further refined into (a) the frequency of MeSH primary concepts and alternative descriptions (synonyms), (b) the frequency of medical terms types and occurences and (c) the frequency of MeSH terms in various top-level categories. The frequency analysis of loan words is based on a selection of affixes that are indicative of medical words of Greek or Latin origin and less likely to appear in general purpose vocabularies. We associated with each affix one or more English words that correspond to the loan suffix, e.g. *mammo-/breast-, angio-/artery-, oesophag-/throat-.*

Our original corpus comparison identified additional discriminatory features (for example, discourse structure features such as the frequency of certain discourse relations and sequences of discourse relations, layout features such as the frequency of tables, headings or figures). However, for the purpose of this project we only selected those features that are not only good discriminators but also likely to be independent of the style and distribution format of the source and target documents. The result of the data profiling exercise is a collection of features and their associated scores, which together constitute the surface model of patient friendly text.

3. Text simplification procedure

Text simplification consist of two phases. In the analysis phase, we identify those features that our data profiling exercise identified as differentiating between medical and lay

Feature	Score
% complex words	14.608
avg sent length	18.953
avg syllables/word	1.628
#MeSH primary concepts	0.100
#MeSH primary types	0.009
#MeSH synonyms	0.047
#MeSH synomyms types	0.004
avg NP length	2.3
%loan words	0.76
%native equivalents	3.99
Fog	12.34
Flesch	54.65

Table 1: Selected features and scores for the reference corpus

documents. Based on these features we compute a technicality score that is used to indicate whether a certain text segment should be rewritten. We then proceed to apply a text simplification process to those segments of text considered too technical, by taking into account the features that contributed to its technicality.

Text rewriting is a two-pass process; the first pass deals with rewriting at the token level, where tokens may be multi-word units, and the second deals with the text at the paragraph level. In the first pass the system looks for tokens that may be difficult for the patient to decipher and attempts to provide inline glosses for complex medical terminology, to unpack acronyms and abbreviations, to replace complex words with simpler synonyms. In the second pass the system analyses the text one paragraph at a time recording scores for a number of features that have an impact on the lay reader's ability to understand the text.

In the first of the two following case studies we present a detailed description of the process through which complex medical terms are analysed, segmented and transposed into a semantic representation from which a short gloss for the lay reader can be generated and inserted inline into the text. In the second case study we consider the research questions posed by the system requirements at the paragraph level: in particular, how to balance a basket of different feature scores which are based on different ranges, scales and distributions, how to tie a paragraph-level failing with respect to a specific feature back to the individual words and phrases that contribute to the problem, and how to parameterise.

We performed our analysis on a corpus of narratives extracted from a large repository of Electronic Patient Records (Rector et al., 2003). We selected approximately 11000 patient records totalling almost 2 million words.

3.1. Rewriting at lexical level

This case study focuses in detail on the strategy that the system uses to handle complex medical terms in the narrative. Our intuition is that it is unhelpful to gloss the med-

ical term with a lengthy explanation, since such explanations often introduce further complex or technical vocabulary and disrupt the flow of the text. Instead the system glosses the medical term explicitly inline (with the addition of a parenthetical glossing expression such as this) with a short paraphrase. For example, the gloss for nephritis (a swelling of the kidney) tells the reader nothing about the possible cause, visible symptoms or treatment associated with the condition. However, since the narrative relates to a patient's own medical experience, the gloss is intended as an aide-memoire rather than a full explanation. Consider for example the following two glosses for *nephritis* from the internet:

```
Nephritis is inflammation of one or both
of the kidneys - the organs that filter
the blood and get rid of excess fluid and
unwanted chemicals.  The inflammation can
be caused by many different conditions.
(From BBC Health http://www.bbc.co.uk/
health/conditions/nephritis1.shtml)

Nephritis is inflammation of the kidney.
The word comes from the Greek nephro-
meaning "of the kidney" and -itis meaning
"inflammation".  Nephritis is often caused
by infections, toxins, and auto-immune
diseases.
(From Wikipedia http://en.wikipedia.org/
wiki/Nephritis)
```

Whereas the two example glosses attempt to explain something about nephritis our aim is to provide a minimal gloss that suffices as a trigger to revive the patient's own lay understanding of the term. For example, the document may contain the term *nephritis* and *neutrophil*, potentially confusable technical terms to the lay reader which can be differentiated with a simple gloss of each one. Our intuition is that there is a substantial gap between the expert and the lay *agenda* in understanding a medical term, especially when the lay understanding is from the perspective of a patient. Many explanations of complex medical terms focus on the biological processes and functions behind body parts, conditions and treatments. In contrast, we believe that patients' interest is more focused on their experience: Is it serious? Will it hurt? How long will it last? Can I still put weight on it? etc.

Medical experts are familiar with the loan stems that underpin the etymology of many complex medical terms, such as the segmentation provided by the Wikipedia entry for *nephritis*. An expert can 'parse' the term and recover the minimal description of the condition implied by the etymology (an inflammation of the kidney) and can then rely on their expert knowledge to infer further details. We aim to provide the patient with the same minimal description and allow then to rely on their lay understanding of their own medical history to infer the information that matters to them. As we set out below, we approach the rewriting process as a natural language generation task, where a semantic representation of the input term is automatically created and then regenerated into lay language. We allow the

segment structure of the original medical term to determine the content of the gloss, rely on a well-known and well-founded ontology to derive a semantic representation and use common terms to lexicalize the derived concepts. Since we want the gloss to distract the reader as little as possible we reanchor the representation to ensure that the syntactic category of the gloss matches that of the original term to ensure fluency. In the following sections we describe in detail the process through which the term is semantically interpreted through segmentation, and subsequently how it is anchored appropriately and rewritten.

3.1.1. Term interpretation
Modelling medical concepts
We construct a descriptive ontology of medical concepts which we will refer to as T-box, using description logic terminology. The T-box is based on standard compositional ontological models such as Galen and Snomed. Our medical concept model is based on the assumption that the semantics of a medical term is recoverable from the semantics of its components. As we later discuss in section 3.1.4., this is a somewhat restrictive assumption that may lead to failures in the semantic parsing of the terms, however it offers a good starting point for describing medical term semantics. Medical concepts are classified into 6 high level categories: *disorders*, *body structures*, *procedures*, *devices*, *agents*, *physical agents* which are further divided into a small number of subcategories (e.g., *body structures* are divided into *body parts* and *body functions*. Each high level medical concept is further described in terms of its attributes and relationships with other concepts. Figure 1 shows a section of our conceptual model and some examples of medical concepts that it models, together with their intended translation.

Whilst the conceptual models used in large scale medical terminologies are built with the purpose of accurately and unambiguously describing and classifying medical terms, our much more simplistic concept model only roughly models the exact meaning of medical terms. Its main purpose is to provide a bridge from medical terminologies to natural language representations, therefore it only models those term components that could be potentially identified by automatically parsing medical terms. For this reason we do not differentiate, for example, between qualifying and defining attributes but instead we collapse them into a general *Modifier* category. Although the resulting conceptual model has quite a flat structure, our initial experiments showed that it provides sufficient coverage for mapping the MeSH terms encountered in a 10 million word corpus of medical narratives.

Term segmentation
The starting point for the semantic term segmentation process is a dictionary of medical specific stems which we compiled from various online sources. The dictionary consists of over 300 stems and their lay definition and contains suffixes (*itis, ic*), prefixes (*an*) and infixes.

The term segmenter uses pattern matching to split MeSH medical terms into components and return a list of stems. For example:

```
nephritis: [nephr = kidney, itis =
inflammation]
adenocarcinoma: [aden = gland, carcinoma =
cancer]
osteoarthritis: [osteo = bone, arthr =
joint, itis = inflammation]
```

3.1.2. Semantic interpretation
The stemming does not provide any information about the way in which term components relate to each other. In the example above, we are not able to tell that *bone* modifies *joint*. In order to achieve this meaning, we try to construct an instantiation of the descriptive ontological model (A-box in description logic terminology). We start by anchoring the medical term into the correct high level concept in the T-box. This anchoring process is facilitated by the fact that the medical term we are trying to interpret was initially identified by MeSH look-up, so we are able to retrieve the MeSH category (or, in some instances, multiple categories) that the term appears in. We do this by mapping top-level MeSH categories into high level categories in our conceptual model. A one-to-one mapping is not possible, one MeSH category may correspond to multiple T-box categories, however we are still able to significantly restrict our search to a small number of initial anchors. Some suffix information relates syntactic category information, for example the suffix *ic* which indicates that the term is expressed as an adjectival modifier, and so these segments are dropped from the semantic representation.

For example, when the system encounters the term *osteoarthritis* it begins by segmenting the term and mapping the segments into the T-box as follows:

- **osteo** maps to **bone** which is a subtype of **body_structure**
- **arthr** maps to **joint** which is a subtype of **body_structure**
- **itis** maps to **swelling** which is a subtype of **disorder**

The A-Box is constructed using a chart-based, breadth-first, recursive algorithm. All three segments, including the mapped T-Box concepts, are added to the chart. The A-box is then anchored on an appropriate root, as explained above, in this case on the concept *swelling* and this root is removed from the chart. Next we explore all of the attributes of the root, marked as labelled arcs in Figure 2, and check to see if there are any entries on the chart which are subtypes of the filler concept that constrains each one. In this case there are two entries on the chart that are subtypes of *body_structure* and so we apply some heuristics to determine which to use - we use *joint* because the segment *arthr* borders on the segment annotating the attributes parent concept (*itis*) whereas the other does not.

We chose to use breadth-first search following an intuition that the resulting structure should be as flat as possible, with most terms relating directly to the root. The heuristics are based on an assumption that the system should prefer to link segments that are closest together, or if both candidates are contiguous with the parent segment then the system prefers to chain compounds of the same T-Box type, and failing that it uses segments from the right hand side of the term
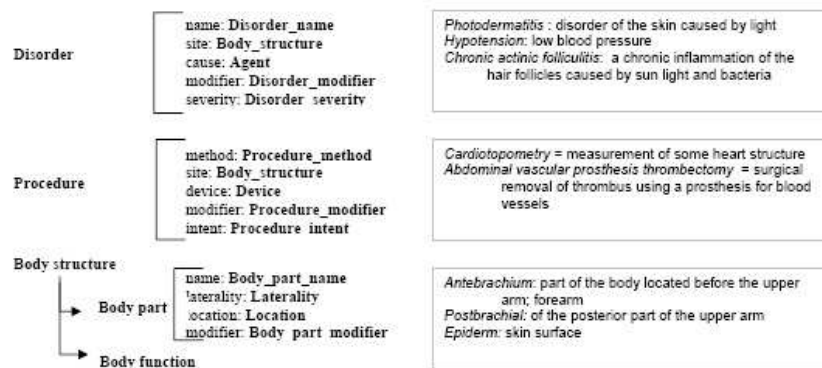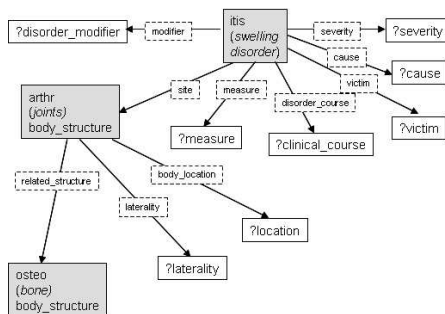
Figure 1: Conceptual model

first.

Figure 2: Abox constructed for the term *osteoarthritis*

3.1.3. Term rewriting procedure

Generating text from semantic representations is a classical Natural Language Generation task, which can be approached either as deep generation or surface generation. Whilst deep generation is generally preferred in cases where the input is a complex semantic representation and flexible output consisting of several coordinated clauses and sentence is required, template-based surface generation is the preferred approach for generating short textual description with little or no syntactical variation required.

We define templates that correspond to primitive types, which have slots representing attributes of that type, and by defining clear realisation rules for each template. For example, a *Disorder* template will have slots for site, cause, agent, etc. and will be realised as "[Disorder] of [Site] caused by [Cause]":

```
Disorder:
Name:   Disorder_name
Site:   Body structure
Cause:  Agent
```

Some slots in the template can also be templates. For example, the disorder site will be a template of type Body structure, representing a concept which can be further qualified, by specifying its laterality, related body structure and other qualifying attributes. Thus, in constructing the textual representation *Inflammation of the joint bone*, one has

to combine the templates *Disorder* and *Body part*. Since the aim of the concept rewriting is to provide inline explanations, we anchor the generated description back in text by associating the same syntactic features as the original concept.

3.1.4. Discussion

Although many of the affixes in the list are substrings of other affixes, as a result of which the segmentation algorithm rejects any segmentation path that cannot fully explain the term to avoid false segmentations, very few of the affixes appear to be problematically polysemous. In many cases the affix can relate to other segments of the term in different ways, but the system is able to capture this flexibility appropriately. In some cases, however, the T-Box is insufficiently fine-grained to capture the semantics correctly. For example the affix *aden(o)* is mapped to the concept *gland*, and this can denote a location in the body or an association with it. However, *adenocarcinoma* can mean cancer that is sebacious (like a gland) rather than just originating in or located in glandular tissue, and such cancers are not represented correctly at present.

3.2. Rewriting at paragraph level

The first step in analysing a new medical narrative is to identify in text the features described in the Data profiling section. This identification is supported by a pre-processing stage consisting of syntactica parsing, sentence and clause splitting, RST parsing and medical term identification. The result of the initial analysis phase is a list of feature scores associated to each text segment (in our case, paragraph).

When the system examines a basket of feature scores and determines whether the whole segment needs to be rewritten or not it needs some information about the domain of the scores for each feature in order to combine them all together. One option would be to hardcode a threshold for each feature and reject segments of the text where some hardcoded threshold of feature failures were found to occur. However, such a system seems unsatisfactorily arbitrary in design and would not be portable, and any changes to the set of features being used would require manual configuration changes to the system.

If the system is to determine the thresholding automatically it needs to know the directionality of each feature (in other

words whether a low or high score is desired) and to understand something about the distribution of scores for each feature across the target corpus and the text under analysis.

Source	A	B	C	D
Segment	0.44	1,239	4.0	-3
Target Corpus	0.32	290	8.6	11

Table 2: A Sample Feature Set

Table 2 lists some scores for a set of imagined features A-D for some segment of the text under analysis and for the target corpus. Directionality notwithstanding, the system still needs to determine how to react when the segment outscores the target corpus for certain features, and in particular how to combine this information. How does the difference in scores for feature A compare to the different for feature B? Which, if any, is problematic?

We cannot simply rely on ratios since the distribution of scores for each feature may not be linear. Instead the system takes the maximum, minimum and mean score for each feature from the target corpus and fits them to a 2-degree Lagrange interpolating polynomial adjusted to the x-axis such that the corpus minimum has an x-value of 0.0, the mean has an x-value of 0.5 and the maximum an x-value of 1.0. It then treats the feature value for the segment as a y-value and uses the polynomial to recover the smoothed x-value in the range 0.0-1.0. This process ensures that the range of each feature score is the same, but within that range the shape approximates the original distribution. We can then compare the segment feature vector to the target corpus feature vector using the Dice metric (twice the inner product over the sum of the Euclidean norms) or by measuring Euclidean distance (the root of the sum of the squares of the differences), in either case measuring only distances in the direction that is of interest.

The result of the technicality assesment phase is a list of segments that require rewriting, together with the features that contributed to the technicality of the document. The next phase uses these features and their scores as parameters for the text rewriting process. On attempting to rewrite text at paragraph level based on a set of technical features we are faced with two major challenges.

First, some of the features we use are qualitative rather than quantitative. This is the case of, for example, readability scores, which are based on a combination of factors such as sentence length and word complexity which may not individually be problematic. Our challenge is to decompose composite measures to identify the exact cause of readability obtrusiveness and rewrite the text accordingly.

The second challenge is attempting to modify the text simultaneously based on constraints that interact at multiple levels. Most document features are not independent. Therefore, the rewriting suggestions the system provides may themselves have an unwanted impact on the rewritten text, leading to a circular process for the end-user. For example, the document structure and the syntactic structure are closely inter-dependent: changes at syntactic level could have a dramatic impact on the rhetorical structure of a text. In attempting to rewrite a text segment deemed too

technical due to its rhetorical structure we may introduce complex syntactic structures that render the text too technical according to its syntactic features. Similarly, trying to minimise the frequency of, for example, nominalisations in text could result in an increase in sentence length, which in turn may lead to increased readability scores.

One solution to this problem is to adopt an iterative process of simplification, where several passes of simplification are performed on the text until all the technicality scores are below the intended technicality level. However, this method does not guarantee a correct output, due to the circularity problem we mentioned. We are looking into statistical methods of inferring rewriting rules using the model of patient language described in section 2..

4. Related work

Comparing technical and lay medical language with a view to making medical literature more accesible to patients is one of the fastly emerging areas of research, with various projects performing comparisons between expert and lay texts. Closer to our work is the research described in (Leroy et al., 2006), that describes an experiment which analyses the readability of medical documents and compares it with patient blogs. Their conclusion is that readability is a combination of lexical and syntactical constraints and argue that word rewriting could significantly improve the readability of some medical documents which deal with easier topics, but may not be sufficient for documents discussing topics which are harder to understand.

There is a large body of literature dealing with the generation of textual reports for patients. Projects such as MIGRAINE (Buchanan et al., 1992) and PIGLIT (Binsted et al., 1995) generate personalised explanations of migraine and diabetes using information extracted from the patient's electronic medical record. Our work is substantially different from these project in the fact that we are aiming at making the actual electronic patient record available in a patient friendly format, instead of using it as a source for parameterising the personalisation process.

There are various text simplification projects that are aimed at people with learning difficulties or some medical conditions or people with low readability levels. We use a similar overall architecture as described in (Siddharthan, 2002). Our rewriting system differs in the fact that we only attempt partial regeneration, therefore a complete model of the input document is not required. We also rely on a more comprehensive set of rewriting rules which are drawn from a model of patient language.

The medical concept interpretation using a description logic based ontology is similar to the method described in (Baud et al., 1997). However, whilst their method is aimed at regenerating medical concepts into natural language for the purpose of ontology checking, we aim at generating completely new textual representations and also linking them back in text.

5. Conclusions & Future work

This paper described an ongoing project which addresses the problem of rewriting the narrative part of Electronic Patient Records into patient friendly language. The rewrit-

ing rules are based on a language model consisting of textual features extracted from a corpus-based comparison of expert and lay medical documents. The rewiting process works in two steps: first, we translate medical concepts into lay alternatices and second, we identify the paragraph-level features which contribute to the technicality of the document.

The project is still in its initial phases of development, therefore a comprehensive user-based evaluation is not at this time available. In some preliminary experiments we investigated whether the technicality asessment process correctly identifies the segments of text that need rewriting and the features that contribute to the technicality of the segment. Five users with little medical knowledge and similar readability levels were asked to read a set of 10 medical narratives containing 27 paragraphs, and to mark the text segments that require rewriting. The segment identification method proved succesful, with 26/27 (96.3%) segments marked correctly are requiring/not requiring explanation. Asessing whether the features that contributed to the segment technicality were correct proved more challenging. Since we couldn't present the users with the output of the text rewriting system, we asked them to identify the changes that would make the text more readable. This however did not lead to any conclusive results, since the range of possible changes was too extensive to be able to obtain a consensus between users. Some of the changes they suggested matched some of our simpler system's predictions (for example, sentence length reduction and explanation of medical terms), however more complex required changes were not consistently identified.

Our current and future work will concentrate on designing rewriting rules that combine technicality features at paragraph level, a research problem which is both novel and could impact on the design of text simplification systems in general. In the future we intend to also consider rewriting parameters that take into account user-specific readability characteristics which could be inferred from each individual EPR.

This project could have an impact on the distribution of patient friendly translations of EPRsin several ways:

- as a complete simplification tool which produces patient friendly text,

- as an aid to medical professionals that release EPRsto patients, by highlighting segments of text that need explanations and providing rewriting suggestions

- as a support tool for patients who require various levels of rewriting

6. References

R.H. Baud, J.M. Rodrigues, J.C. Wagner, A.M. Rassinoux, C. Lovis, P. Rush, B. Trombert-Paviot, and J.R. Scherrer. 1997. Validation of concepts representation using natural language generation. In *AMIA Annual Fall Symposium*, pages 841–847.

K. Binsted, A. Cawsey, and R.B. Jones. 1995. Generating personalised patient information using the medical record. In *Proceedings of Artificial Intelligence in Medicine Europe*, Pavia, Italy.

B.G. Buchanan, J.D. Moore, D.E. Forsythe, G. Carenini, and S. Ohlsson. 1992. Involving patients in health care: explanation in the clinical setting. In *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care (SCAMC'92)*, pages 510–512.

Kathi Canese. 2003. New Entrez Database: MeSH. *NLM Technical Bulletin*, March-April.

D. Detmer and P. Singleton. 2004. The informed patient. Technical Report TIP-2, Judge Institute of Management, University of Cambridge, Cambridge.

D. Hardcastle and C. Hallett. 2007. Exploring the use of nlp in the disclosure of electronic patient records. In *Proceedings of BioNLP'07*, Prague.

Dirk Hüske-Kraus. 2003. Suregen-2: A shell system for the generation of clinical documents. In *Proceedings of EACL'03*, pages 215–218.

R. B. Jones, J. Pearson, A. J. Cawsey, D. Bental, A. Barrett, J. White, C. A. White, and W. H. Gilmour. 2006. Effect of different forms of information produced for cancer patients on their use of the information, social support, and anxiety: randomised trial. *BMJ*, 332:942–948, April.

D. Kokkinakis, M. Toporowska Gronostaj, C. Hallett, and D. Hardcastle. 2007. Lexical parameters, based on corpus analysis of english and swedish cancer data, of relevance for nlg. In *Proceedings of NODALIDA'07*.

Gondy Leroy, Evren Eryilmaz, and Benjamin T. Laroya. 2006. Health information text characteristics. In *Proceedings of the AMIA Annual Symposium*, pages 479–483.

Kathleen R. McKeown, Desmond Jordan, Steven Feiner, Jame Shaw, Elizabeth Chen, Shabina Ahmad, Andre Kushniruk, and Vimla Patel. 2002. A study of communication in the cardiac surgery intensive care unit and its implications for automated briefing. In *Proc. of the American Medical Informatics Association 2000 Symposium*.

L. M. Osman, M. I. Abdalla, J. A. G. Beattie, S. J. Ross, I. T. Russell, J. A. Friend, J. S. Legge, and J. G. Douglas. 1994. Reducing hospital admission through computer supported education for asthma patients. *BMJ*, 308:568–571.

Alan Rector, Jeremy Rogers, Adel Taweel, David Ingram, Dipak Kalra, Jo Milan, Robert Gaizauskas, Mark Hepple, Donia Scott, and Richard Power. 2003. Clef - joining up healthcare with clinical and post-genomic research. In *Second UK E-Science "All Hands Meeting"*, Nottingham, UK.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *Proceedings of the Language Engineering Conference 2002 (LEC 2002)*, pages 64–71.

Olof Torgersson and Göran Falkman. 2002. Using text generation to access clinical data in a variety of contexts. In G. Surján, R. Engelbrecht, and P. McNair, editors, *Health Data in the Information Society. Proceedings of MIE2002*, volume 90 of *Studies in Health Technology and Informatics*, pages 460–465. IOS Press.