Lessons from Developing an Annotated Corpus of Patient Histories

Thomas Brox Røst
Department of Computer and Information Science
Norwegian University of Science and Technology
NO-7491 Trondheim, Norway
brox@idi.ntnu.no

Ola Huseth
Department of Language and Communication Studies
Norwegian University of Science and Technology
NO-7491 Trondheim, Norway
olahu@hf.ntnu.no

Øystein Nytrø
Department of Computer and Information Science
Norwegian University of Science and Technology
NO-7491 Trondheim, Norway
nytroe@idi.ntnu.no

Anders Grimsmo
Department of Community Medicine and General Practice
Norwegian University of Science and Technology
NO-7491 Trondheim, Norway
anders.grimsmo@ntnu.no

We have developed a tool for annotation of electronic health record (EHR) data. Currently we are in the process of manually annotating a corpus of Norwegian general practitioners' EHRs with mainly linguistic information. The purpose of this project is to attain a linguistically annotated corpus of patient histories from general practice. This corpus will be put to future

Journal of Computing Science and Engineering, Vol. 2, No. 2, June 2008, Pages 162-179.

use in medical language processing and information extraction applications. The paper outlines some of our practical experiences from developing such a corpus and, in particular, the effects of semi-automated annotation. We have also done some preliminary experiments with part-of-speech tagging based on our corpus. The results indicate that relevant training data from the clinical domain gives better results for the tagging task in this domain than training the tagger on a corpus from a more general domain. We are planning to expand the corpus annotations with medical information at a later stage.

1. INTRODUCTION

Medical treatment and other contacts with a patient are always documented with a narrative text in the health record describing the encounter. Over time, the amount of information grows, making it difficult to make sense of the patient's treatment history. A lot of effort has been put into the application of natural language processing (NLP) methods as a way of extracting relevant information–be it clinical findings, adverse drug events, or sensitive information– from the health record.

For our research, we take a particular interest in the domain of primary care. NLP applied on primary care patient records is to a large extent uncharted territory, possibly for practical reasons: Data from hospital patient records are easier to gain access to for researchers within e.g. university clinics. In many cases, this is not a problem. For research where NLP is used as a means toward improving treatment of particular medical conditions–such as tuberculosis and breast cancer–the use of specialized patient records is a necessity. From such sources we may learn a lot about the application of health care in particular–but not in general. For most people their primary point of contact with health services is their primary physician, whose job is not only to manage the health of the patient but also to manage interactions with other health care actors. In sum, the information recorded by the primary physician gives the most complete picture of the lifecycle of a patient. This longitudinal information is, from our point of view, a necessary prerequisite for learning how illness is documented; knowledge that, in turn, may help us create patient record systems that are better adapted towards managing patient treatment.

Through our collaboration with the Norwegian EHR Research Centre[1] (NSEP) we have access to a large data set from a Norwegian primary care practice. NSEP is a multidisciplinary research community at the Norwegian University of Science and Technology. The centre is involved in different research projects regarding development, use and usefulness of electronic patient records. Our long-term goal is to use the data set as a basis for developing tools and techniques that enable easier access to data hidden within the patient record–in particular through automated structuring of patient record narrative. NLP applied on narrative in the patient record can be supported by an annotated corpus that is representative for the given domain. Such a corpus will give us training and evaluation data for construction of automated tools. A primary research area is to uncover the intentionalities behind different parts of the consultation note. Through analysis of the syntactic and grammatical characteristics of the narrative we seek to find function and structure in texts where the structure is not explicitly stated.

[1]http://www.nsep.no

This paper describes our efforts in the development of a linguistically annotated corpus that will form a basis for future research efforts. The intended outcome differs from development of other medical corpora in three ways. First, our focus is on data from primary care. Second, we are interested in complete patient histories that may span several years of treatment, as compared to isolated incidents. Third, our approach towards making sense of such patient histories will be primarily shallow and data driven; that is, in the spirit of keeping things simple and at the same time realizing robust applications that can handle the kind of noisy, ungrammatical narrative typically found in primary care health records, we forgo the use of traditional rule-based NLP methods.

We give particular emphasis to our experiences from developing such a corpus and some preliminary results indicating how a domain-specific corpus fares against a more general corpus depending on corpus size. Developing corpora of this size can be time consuming. One of our goals has therefore been to automate parts of the annotation process without sacrificing annotation quality.

2. BACKGROUND AND MOTIVATION

The electronic health record is the main tool for recording and communicating information about the medical care process. It is, however, a tool that in many instances fails to deliver to its full potential:

- Navigating the health record and retrieving relevant information gets increasingly difficult as the amount of information grows, almost to the point of being inaccessible through simple browsing and searching.
- Electronic health record systems in use today are only semi-structured. Physicians still document clinical encounters in the traditional written narrative. There are perfectly valid reasons for doing so, such as the semantic richness of narratives and the problems associated with structured data entry [Walsh 2004]. As noted by Powsner [Powsner et al. 1998], clinicians "value the ability of flowing prose to paint an evocative clinical picture." Well-structured and rich representations of health record narrative are often lacking, thus reducing the health record's utility for clinical, administrative and research purposes. Nonetheless, a lot of the information surrounding the encounter note is available in a structured, easily accessible and sometimes standardized format. This information is, however, insufficient for providing a complete view of the patient's state and history. Thus, finding ways of extracting relevant information from encounter notes is a useful research goal.
- The lack of structure in health record narrative makes it difficult to do research on their use and content. In order to develop electronic health records better suited for both clinical practice and research purposes, one needs a clearer picture of actual usage and documentation patterns. Knowing the hows and whys of clinical documentation might be essential for providing improved ways of documenting clinical practice.

The shortcomings of electronic health records and the intractability of clinical narrative have triggered substantial research efforts into providing structure and accessibility where there is none. Specifically, steps have been taken in the application of natural language processing and data mining methodologies to clinical documentation. The purpose is to automate querying and information extraction from clinical narrative. Finding ways of unlocking the information within clinical documentation would benefit both clinical practice and health record research.

2.1 Medical Language Processing

The clinical data available in coded format is not sufficient to fully communicate the patient's true state and progress [Iezzoni 1997; Spyns 1996]. While health care institutions increasingly store medical information in an electronic format the frequent use of narrative makes them inaccessible to large scale or automated analysis [Hripcsak et al. 2003].

Simple text searches in the electronic health record can prove effective in detecting concepts of interest [Giuse and Mickish 1996; Goldman et al. 1999; Honigman et al. 2001] but suffer from serious shortcomings. Some of the problems with keyword detection are negated words, different ways of expressing the same concept, ambiguity resolving, and interpreting the context in which concepts occur. Thus, this simple approach will often lead to many false positives and poor specificity [Murffu et al. 2003].

Several studies have shown that medical language processing (MLP)–natural language processing (NLP) applied in the medical domain–can achieve much higher accuracy than simple concept detection techniques [Hripcsak et al. 1995; Hripcsak et al. 1998]. Notably, with respect to sensitivity (recall), specificity and positive predictive value (precision), the performance of some systems is shown to be indistinguishable from physician performance [Hripcsak et al. 1995] and superior to other methods [Fiszman et al. 1999].

The text in EHRs is often fragmented and quite often plainly ungrammatical. This state of affairs requires robust NLP analyzing methodology. Traditional deep linguistic grammars often have a problem with getting broad enough coverage and thus lack robustness. Parsing with such grammars also tends to be time-consuming and inefficient, and the output is often highly ambiguous. Shallow NLP processing can to some extent solve these issues, but the pay-off is less informative output. Shallow techniques can also be used as preprocessing modules in deep grammars, to help resolve some ambiguities early. Machine learning techniques can be applied to shallow processing, but requires training data. Both data-driven and rule-based methods require at least some annotated data for evaluation purposes. Examples of shallow NLP techniques include:

- **Part-of-speech (POS) tagging:** POS tagging is the process of assigning each word in a text with its correct part-of-speech tag in the relevant context. There are two basic tasks: choosing the correct tag for ambiguous known words, and assigning tags to unknown words. Part-of-speech tagging is widely used in other disambiguation tasks (e.g. speech recognition) and as a first step towards richer

syntactic structures.

- **Noun phrase chunking:** Noun phrases (NPs) often carry the most interesting pieces of information in running texts. The set of NPs comprises subsets like proper names (John, Jane), location names (London, Europe), other proper nouns (Viagra, Losec), common nouns (medicine, pain), phrases (no pain, improved general condition), and so on. NPs are often the target items for search in texts, and NP chunkers are thus useful tools in information retrieval.
- **Shallow parsing:** Shallow parsing (often called partial parsing or chunking) assigns some syntactic structure to sentences. Instead of hierarchical, nested structures, a shallow parser identifies chunks or phrases that are contiguous and non-overlapping. Shallow parsers can be used for preprocessing in deep grammars, or as basis for robust semantic analysis.

2.2 The Health Record in Primary Care

Most MLP research on health records concerns itself with specialist documentation that originates from within hospitals, while the domain of primary care has been largely overlooked. We believe there are particular traits of the Norwegian health care system that makes research on primary care health records attractive. Our data originates from areas with a low rate of migration, ensuring the availability of comprehensive and long-term patient histories. Also, the list patient system reform of 2001 established that each patient should have a single responsible primary care physician. Some of the goals were to ensure continuity, reduce patients' switching between different physicians, and to strengthen the gatekeeper role of the primary care physician. While the reform inflicted some short-term loss of continuity, it is still regarded to have been a success [Bakken 2006]. The most recent evaluation stresses the need for research on primary care practice: Research from specialist care is not immediately transferable to primary care and there is thus a need to initiate more research on clinical practice in primary care.

Moreover, Norway's early adoption of electronic communications between primary and specialist care implies that the patient histories are supplemented with additional data, such as hospital discharge notes and communications with social services. This further enhances the uniqueness and completeness of available data. Finally, electronic health records have been in common use in Norwegian primary care since the early 1990s. Accordingly, the opportunity exists to follow patient histories across a considerable time span. These characteristics, combined with the lack of previous research, make a strong case for focusing on primary care health records.

2.3 Patient Histories

A lot of research on the application of NLP on patient documentation focuses on single notes and narratives. Given that disease and its treatment can be complex and long-term, these are just brief glimpses that might say little or nothing about the bigger picture. There is a distinct lack of research that considers how a health record note exists within a context and that it has a past and a future: The observations,

interventions and outcomes of previous treatment and the (often implicit) purposes and expectations of upcoming care, both in the short and long term.

For these reasons we have ensured that the corpus consists of full patient histories where the included patients can be followed over time. Moreover, links to the data structure in the originating EPR have been preserved. This makes it possible to trace the additional information associated with each consultation note, such as classification codes, prescriptions and lab results. Previous research has indicated that consultation notes in the primary care patient record can not always be interpreted through the narrative alone; the accompanying information fills in details that can not be inferred from the text on its own [Røst et al. 2007]. Intuitively, this makes sense: The text should not repeat what has already been stated through structured data entry but, if necessary, rather supply a story that motivates for e.g. cessations, ordering of lab tests, or the issuing of medical certificates. It can thus be argued that automated processing of patient record narrative should also take the additional structured information into account in order to increase the probability of making the correct inferences from the text.

For the research described in this paper, the additional structured information will not be put to use or described further–though we expect to base future research on a combination of language processing applied on the narrative and traditional data mining applied on structured data in the corpus.

2.4 Structuring Health Record Narrative

A primary motivation for building this corpus was to use it as a basis for long-term research on techniques that make information in the EHR more readily accessible to its users. In practice, this implies finding structure in the free-text narrative which constitutes a major part of the information content in the EHR. The availability of linguistically annotated texts is necessary when progressing from simple, lexical approaches to parsers that can infer the grammatical structure of health record narrative. In comparable medical language processing research, this kind of information can e.g. help finding qualifiers and modifiers such as negation and adjectives for clinical findings [Hripcsak et al. 1995].

A second approach is to use the availability of linguistic information to build richer representations of health record narrative for data mining and classification purposes. We have previously attempted to classify consultation notes into their corresponding ICPC classification code [Røst et al. 2006]. This is, from a practical point of view, a feasible task: We are training classifiers to classify text based on a gold standard that is already present in the health record. However, this does not tell us anything new about the content. A far more interesting challenge is to use classification to help reveal any hidden structures that are not readily available to us.

Sharda et al. did a study where it was shown that restructuring narrative in clinical discharge summaries lead to an improved recall rate when test subjects where tasked with verbalizing their assessment of each summary [2006]. This indicates that the application of structure can be useful on an intra-document level. We believe that restructurings can also prove useful on inter-document levels; that

is, on full patient histories. Prior to creating this corpus we tried using text compression algorithms on patient history consultation notes in order to detect novelties or anomalies–novelties here being text fragments that stands out from a lexical point of view [Edsberg 2007]. The motivation for this research was to highlight consultations in the patient history that may be more noteworthy than the rest; this from the observation that a considerable part of disease treatment in primary care is documented in a very homogenous manner [Røst et al. 2007].

We intend to continue this line of research with the annotated corpus. A first application area will be an attempt to classify sentences within the consultation note according to their function and intentionality. The initial classification task will be to classify sentences according to the SOAP format, as proposed by Lawrence Weed in 1969 [Weed 1969]. SOAP is closely related to the advent and rise of the problem oriented medical record and suggests that a consultation note should be structured according to four categories: Subjective observations, Objective observations, Assessment and Plan. Through enhancing the corpus with SOAP annotations, the intended outcome is to learn if there are patterns in the way these categories are used throughout the treatment of a disease. The linguistic annotations of the corpus will help establish if there are lexical, syntactic and grammatical features that enable classifiers to differentiate between these categories.

In general, methods of automating the application of structure to unstructured health record narrative should be a benefit. Not only do physicians prefer reading standardized documents [Walraven et al. 1999], but structuring also improves the completeness and accuracy of clinical narrative [Johnson et al. 2008]. Structured data entry will typically prove more time-consuming than free text information entry [MacDonald 1997] while at the same time losing out on the innate ability of text to evoke a more complete picture of the patient's situation. Johnson et al. [2008] describes how "structured data entry can be quite slow when events are broad in scope and exhibit high variation." This is exactly the kind of situation one finds in primary care, motivating for finding ways of applying structure to primary care health records without sacrificing the convenience of free-text narrative.

3. RELATED WORK

A growing interest in data-driven natural language processing research has lead to the development of annotated corpora for testing and training of computational models for language applications. Common annotation categories are part-of-speech tags, base forms, phrasal categories and syntactic tree structures. The Penn Treebank [Marcus et al. 1994] has become the de facto standard corpus for evaluation of part-of-speech tagging for English, and, as the name implies, also contains syntactic tree structures. For German, the Negra corpus [Skut et al. 1993] and the TIGER Treebank [Brants et al. 2002] are similar resources, though smaller in size. Large scale annotated corpora are more common for well-used languages, but these kinds of resources also exist for some smaller languages, like the Stockholm-Umeå corpus of Swedish [Ejerhed et al. 1992]. Most of the corpora are built from newspaper texts, which are widely available in large quantities.

In later years, some effort has been put into research on domain adaptation of general language models into the medical domain. Much of the focus has been on part-of-speech tagging. Campbell and Johnson [2001] concluded that there is a syntactic difference between the medical domain and the more general newspaper domain, and that the availability of relevant medical training data gives significantly better results than just adopting a general language model for classification. However, Hahn and Wermter [2004] concluded that "off-the-shelf NLP-tools can be applied to MLP in a straightforward way". The two experiments were performed on different languages (English and German), different tag sets, different part-of-speech taggers and different training and test data from different medical sub-domains, and as such might not be directly comparable.

Pakhomov et al. [2006] discuss the discrepancies between these two studies, and suggest that the richer inflectional morphology of German is one reason for the divergence. The TnT tagger [Brants 2000] used by Hahn and Wermter [Hahn and Wermter 2004] has been shown to do better for German than English when it comes to unknown words, due to its reliance on suffix analysis for classification of unknown words. This strategy is particularly suitable for inflectional rich languages like German. Pakhomov et al. show similar results as Campbell and Johnson [Campbell and Johnson 2001], even though the tagger, the medical sub-domain and the evaluation methodology differ. While the two other reports focus on the impact of syntactic similarities or differences between the training domain and the target domain, Pakhomov et al. note that the amount and types of unknown words in the test corpus also contribute to a substantial degree.

4. DATA

Our data set has been collected from a rural Norwegian general practice center and encompasses all recorded activities in their electronic health record system from November 1991 until October 2006. In total, there are more than 616,000 consultations and 12,000 patients. The population in proximity to the medical center has remained reasonably stable over the years, giving a mix of both longer and shorter patient histories. In addition, the Norwegian list patient system ensures that patient histories are fairly complete for interventions involving general practice.

When selecting patient histories for annotation we did not use e.g. history length, disease type or consultation note size as a selection criterion but rather chose random histories. Over time, this should give us a corpus that is fairly representative for the different types of diseases found in the general population. Note that a patient history is defined as all available encounters for a given individual.

5. ANNOTATION PROCESS

Manual annotation of large corpora requires considerable effort. Streamlining the tools and techniques used in the annotation process may help towards reducing the overall workload. For our project, this involved making sure that the annotation tool had the best trade-off between automation and manual labor in terms of minimizing annotation errors. In addition, we needed to make sure that our initial linguistic

annotation would be suitable for and compatible with future annotation efforts in which the same corpus is annotated from a medical point of view.

5.1 Categories and Tag Sets

So far, our efforts have been directed towards annotation of linguistic information, but we are planning to enrich the corpus with medical annotations at a later stage. For the linguistic annotations, the main motivation is to create relevant training and test data for data-driven natural language processing. For that purpose we have annotated the data with base forms, part-of-speech tags and simple phrasal tags. The set of POS tags (Table I) is about half the size of tag sets used for other languages, e.g. the Penn Treebank tag set. This is partly because we have access to a POS tagged corpus of Norwegian newspaper text that uses this set. Having the same tag set on our data gives opportunities for comparative studies and other evaluations. Having a smaller tag set also simplifies the job of the human annotator. The phrasal tag set follows the IOB format of Ramshaw and Marcus [1995], where I is used for words inside a chunk and O is used for words outside. A word is tagged as being the first word of a phrase by adding the suffix -B to the phrase name, e.g. NP-B for the first word of a noun phrase. Words other than the first one in a phrase get the suffix -I. Words that are outside of any phrase we are interested in is given the tag O. We assume 5 phrasal categories: NP, VP, PP, AP and AdvP, which gives a tag set of 10 phrasal tags (Table I), since we do not use the tag PP-I (every preposition is annotated as PP-B, and any other words in the prepositional phrase are members of some other phrase, e.g. an NP).

In addition to the linguistically motivated annotation categories, the human annotator has the possibility of marking a word as sensitive. This information may be used to

Table I. Tag sets.

Phrase Tags	POS Tags	POS tag explanation
AdvP-B	adj	adjective
AdvP-I	adv	adverb
AP-B	det	determiner
AP-I	infm	infinitive marker
NP-B	interj	interjection
NP-I	konj	conjunction
O	noun	noun
PP-B	noun_prop	proper noun
VP-B	PAR	parenthesis etc.
VP-I	prep	preposition
	pron	pronoun
	subj	subjunction
	tall	number
	verb	verb
	.	sentence-final punctuation
	,	mid-sentence punctuation

help develop automated de-identification tools. She also has the option of marking a word as unsure, which means it will be controlled later by either a medical expert or a linguist. In practice, medical terms and abbreviations were the most common sources of uncertainty for our human annotator, but some syntactic constructions were problematic as well.

5.2 Part-of-Speech Tagger

Studies [Marcus et al. 1994] have shown that manually editing the output from an automatic part-of-speech tagging process, rather than annotating from scratch, can be approximately twice as fast, as well as reducing error rates. This motivated us to make a part-of-speech tagger an integrated part of the annotation tool.

We have previously developed a part-of-speech tagger [Huseth 2005], and this has been integrated with the annotation tool. It suggests POS tags for every word to be annotated. The human annotator then only has to correct the errors. The POS tagger is based on the theory of Hidden Markov Models (HMM) [Rabiner 1989], a probabilistic machine learning technique. In lack of more appropriate corpora, the tagger was originally trained on a corpus of Norwegian newspaper texts with approximately 100,000 words.

The problem of probabilistic part-of-speech tagging can be formulated as finding the most probable tag sequence T given a word sequence W: $\text{argmax}_T P(T|W)$. By Bayes' theorem, the sequence $P(T|W)$ is equivalent to $P(T)P(W|T)$. Our tagger uses supervised training to train a trigram tag model. The trigrams are used to estimate the prior probability $P(T)$, such that $P(T) = \prod_{i=1}^{n} P(t_i|t_{i-2}, t_{i-1})$. The trigram model is smoothed with linear interpolation. The likelihood $P(W|T)$ is estimated directly, without smoothing. This means we are assuming that trigrams not seen in the training data cannot exist. This is a rather strong assumption, but empirically gives better results than any of the smoothing techniques we have tested. For unknown words, the likelihood is estimated from a weighted model of suffixes from the training data. In addition, capitalization and numbers are used to adjust the probabilities for an unknown word having proper noun or number as its tag. As noted by Johannesen and Hauglin [Johannessen and Hauglin 1998], compounding "is extremely productive in Norwegian, and it is futile to ever hope for a lexicon (dictionary) that will contain all or even most of the compounds that occur in actual texts". Unlike many compounds in e.g. English, Norwegian compounds are orthographically realized as single words. The POS tag for a compound is determined by its final part, so we have included a module for the identification of compounds. If the last basic word of a compound is in our training data, the observation probability distribution of that word is used.

5.3 Automation and Incremental Training

As the quality of the automatic annotation improves, the speed of manual annotation can be expected to increase. We thus wanted to benefit from already annotated data from the relevant domain. We do so by making sure the tagger is incrementally trained. This means that each manually tagged sentence is added to the training

data of the POS tagger as it becomes available. The tagger's output is thus expected to improve as the amount of annotated data increases. This way of training the tagger means that probabilities have to be computed for every sentence to be tagged, instead of doing all probability calculations in advance. This makes the tagger slightly slower, but the decrease in speed is barely noticeable when tagging single sentences. On a side note, the annotator working on our data mentioned that as the quality of the automatic annotation improved, she was more inclined to trust the suggestions, which could lead to her becoming less critical and alert when annotating.

Some automatization was involved for base form and phrase tag selection, tokenization and sentence splitting as well:

- **Base forms:** For each word, a set of possible base forms were inferred from the NorKompLeks computational lexicon [Nordgård 2000]. As the amount of annotated data grew, the suggested base forms were sorted according to their overall frequency in the annotated corpus with the most probable base form as the primary suggestion. If no possible base forms were found, either from NorKompLeks or the corpus, the word itself was suggested. Accordingly, the amount of effort needed for base form modifications decreased over time.

- **Phrase tags:** A set of static rules was used to suggest phrase tags based on the POS tags suggested by the tagger. For instance, words with the *prep* tag were automatically assigned the *PP-B* phrase tag, words with the *noun_prop* tag were given the NP-B phrase tag, and so on. As the tagger accuracy increased, so did the phrase tag suggestion.

- **Tokenization and sentence splitting:** Our data included a lot of domain-specific constructs–e.g. blood pressure measurements, lab results and diagnosis codes–that were not properly handled by standard whitespace tokenization algorithms. Special rules to handle these exceptions were applied. Also, the use of abbreviations was very common; more so than in typical corpora. These would cause a lot of unwanted sentence splitting and would have to be dealt with. Our approach was to allow the human annotator to designate abbreviations as she went along and then allow the tokenizer and sentence splitter to make use of this knowledge.

5.4 Annotation Tool

A number of existing annotation and markup tools were evaluated and found insufficient for a number of reasons: The tight coupling between the part-of-speech tagger and the annotation tool; the dual purpose of the annotation process; the need for manual preprocessing of texts in order to ensure correct tokenization and sentence splitting; the need to preserve and take advantage of the original database structure; the focus on patient histories rather than individual notes; the need to integrate data from disparate sources; the need to search annotated texts based on a number of different criteria. Ultimately, we ended up developing our own annotation tool that would satisfy these requirements. The tool was developed using the Python programming language and the Django web framework, since the characteristics of

the tool closely resembles those of a typical database-backed application.

There are three main components to the annotation tool:

- **Sentence and tokenization review:** We quickly found that the encounter notes would have to undergo some manual preprocessing before the annotation. E.g., for ease of annotation it is useful to have the note split into proper sentences. However, the common use of abbreviations and sometimes haphazard use of punctuation often required us to make slight modifications to the original texts, such as adding missing punctuation. A general principle was to modify only when absolutely necessary so as to not deviate too far from the original text.
- **Annotation:** The encounter note is presented to the annotator as separate sentences (Figure 1). The full text of both the original and the edited note is available for reference. Each sentence is shown vertically, with columns containing respectively the word, base form, POS tag, phrase tag, sensitivity and whether or not the annotator is unsure of how to annotate a word. Annotation categories can be navigated using either the keyboard or the mouse. In practice, keyboard navigation turned out to be the most convenient and efficient option. It is worth noting that our annotator preferred annotating a sentence per annotation category–that is, to first annotate all base forms, then all POS tags, and so on–rather than annotating all categories for one word at a time.
- **Search and batch modification:** In order to ensure consistency we occasionally need to review and modify previous annotations. To do so we implemented a custom search interface (Figure 2) that would allow us to perform complex queries across multiple annotation categories and to make batch modifications to the results.

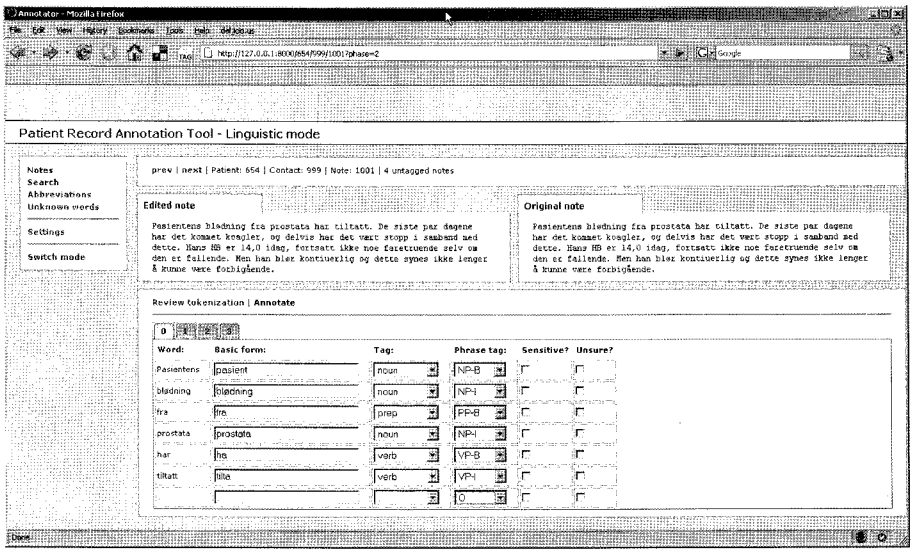


Figure 1. Annotation view.

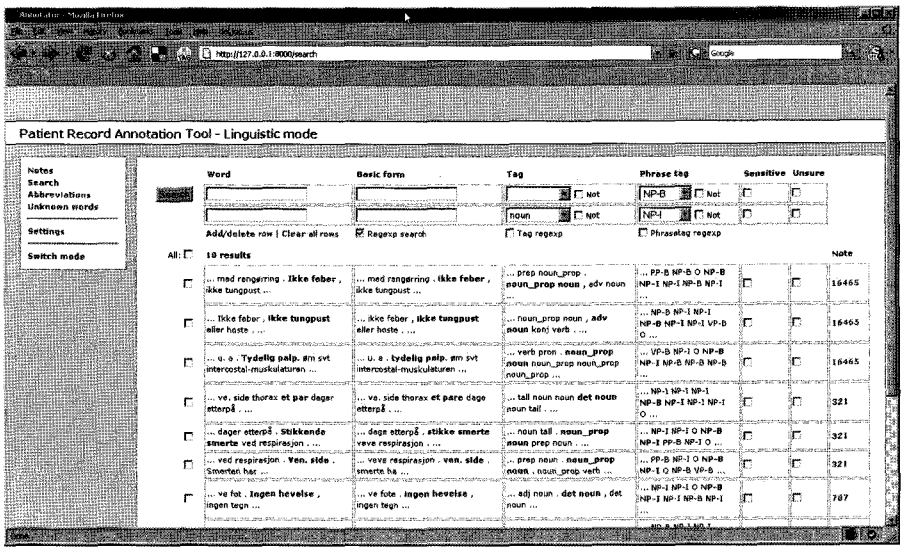


Figure 2. Search view.

6. PRELIMINARY RESULTS

We have done an initial evaluation of the impact of relevant training data on the accuracy of part-of-speech tagging in the medical domain, experimenting with our newly annotated corpus of Norwegian primary care health records as training and test data. Similar experiments have been done before, but then for other languages and other medical subdomains.

We have previously evaluated our part-of-speech tagger on two corpora, The English Penn Treebank corpus [Marcus et al. 1994], and a Norwegian corpus of newspaper text. The Penn Treebank corpus consists of approximately 1,200,000 words of newspaper text. The Norwegian corpus has approximately 100,000 words. The tag set for the Penn treebank corpus has 43 tags, while the Norwegian tag set has 20 tags.

We evaluated the tagger on both corpora by 10-fold cross validation. This means dividing the data into 10 parts, doing 10 iterations where 9/10 of the data is used for training and 1/10 for testing at each iteration. This ensures that test data is unseen by the tagger at every turn, but gives a sufficient amount of test data. The average accuracy of the 10 iterations for the Penn treebank test was 96.10 %. In comparison, Brants [2000] reports an accuracy of 96.70 % for his TnT tagger in a similar study, using the same data and the same test method. For the 10-fold cross validation of the Norwegian corpus, the accuracy was 95.04 %.

We did several experiments where we tested the effect of introducing relevant training data when doing automatic part-of-speech tagging of Norwegian health record data.

- **Experiment 1:** We tested the tagger, trained on the Norwegian newstext corpus mentioned above, on 74,000 words of our newly annotated health record data.

We also did a 10-fold cross validation of the health record data. Thus, the first test uses a general language model on the classification of texts from the health record domain, while the second test only uses in-domain training and test data.

- **Experiment 2:** We set aside 10,000 words from the annotated health records for testing, and used the remaining 64,000 words for training. We started with the newstext corpus as training data, and incrementally added portions of the health record training data to this. The first iteration added 1,000 words of health record data, and this amount increased by 1,000 words for each iteration until all 64,000 words were used for the final iteration. For each iteration the tagger was tested on the 10,000 word test set. The results are presented in Figure 3.
- **Experiment 3:** We did a similar experiment as the previous one, but did not include the newstext corpus in the training data. The size of the training data (taken from the annotated health records) was incrementally increased by 1,000 words, and at every turn tested on the 10,000 word test set. The results are presented in Figure 3.

For the first experiment, the accuracy when we trained the tagger on newstext data and tested on health record data was 76.87 %. When we did a cross-validation of the health record data, however, the accuracy increased to 94.60 %. Even if the evaluation methodology differs for the two tests, the total test data is the same in both.

Experiments 2 and 3 are summarized and compared in Figure 3. From Experiment 2, it is worth noting that by adding only 1,000 words of health record data to the newspaper training data, the accuracy increased from 76.87 % to 84.46 %. It is also interesting that the first iteration in Experiment 3, with only 1,000 words of training data in total (from the health record data) outperforms the test where 100,000 words

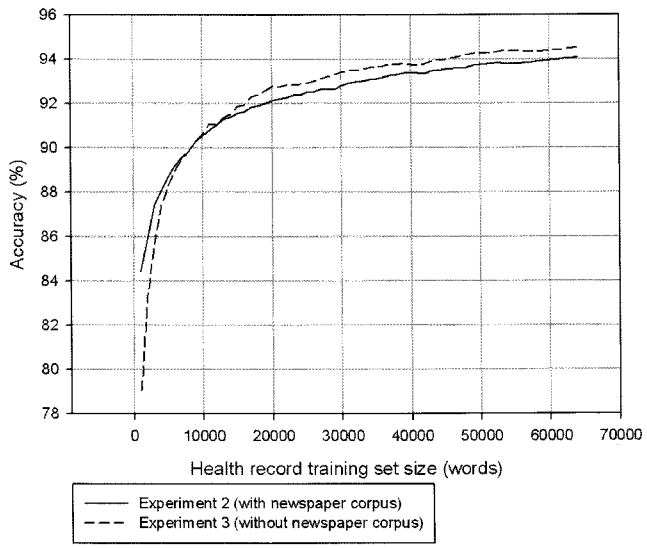


Figure 3. Results, experiments 2 and 3.

of training data from the newstext corpus is used, with 79.05 % contra 76.84 % These tests can not, however, be directly compared, as they use 10,000 and 74,000 words as test data respectively, but still indicate that very little training data from the medical domain is needed to improve part-of-speech tagging in this domain.

From Figure 3, we can also see that the curves cross when we get approximately 10,000 words of health record domain training data. At that point, the newspaper text does not contribute to the improvement of tagging accuracy any longer, but just acts as noise. The difference between the two curves is fairly stable at around 0.4-0.5 percentage points from around 17,000 words and above, so the negative effect of the newstexts seems to be stable and not very severe. At 64,000 words of health record training data, the accuracy is 94.09 % with the newstext data and 94.54 % without it.

7. CONCLUSIONS

Our informal experience from usage of the annotation tool was that automation efforts and incremental learning did, on the whole, benefit the annotation process. Over time, automation reduces the human annotator's role to that of verifying the suggested annotations and to correct the diminishing number of errors. The presence of a search and batch modification interface was of great help during the initial annotation phase when the annotator and the supervisors were still working out how to annotate ambiguous syntactic constructions. Search and modification was not, in fact, one of the original suggested features of the tool but rather an afterthought as we realized the need to ensure consistency through verification of previous annotations.

Our findings from Norwegian part-of-speech tagging of medical text seem to be in accordance with similar studies for English. However, our initial, unadapted tagger performs worse than the same initial studies for English show. One possible explanation might be that we have less general training data to begin with. Other possible explanations are language differences, or differences in documentation practice. We use data from primary care health records, which is often written by the doctor during consultations. Specialist care health records, on the other hand, are often dictated and written at a later stage. This difference may lead to the specialist care record being less prone to grammatical and spelling errors, and thus being more similar to the newspaper texts that the tagger is initially trained on.

As we get more data, the results are comparable to those of Pakhomov et al. [2006]. As Figure 3 shows, adding more training data gives better results. And as we are continuing the manual annotation, our accuracy can be expected to improve further.

It is worth noting that our experiments differ slightly from Pakhomov et al. in terms of training data, test data, medical sub-domains, language, size of accessible data, tag sets and evaluation methodologies and results should, accordingly, be interpreted with caution. As an example, Pakhomov et al. uses a slightly modified version of 10-fold cross-validation, where the corpus is divided in 10 chunks of 2/10 of the corpus size instead of the standard 1/10 division that we used. The overall accuracy of Pakhomov et al.'s experiment when using training data from the medical

domain is 94.69 %, and the average accuracy on the same test data with training data from the newstext domain is 89.79 %. In relation to this last test, Pakhomov et al. also divided the test data according to medical sub-domains, and the results for the different domains varied from 74.70 % (the domain of current medications) to 92.62 % (the domain of family history). Our own accuracy for this kind of test was 76.87 %, and is as such comparable to what Pakhomov et al. reports for some sub-domains, although not for the average. It should also be noted that the size of the test data from the different sub-domains varied from 392 to 43,633 tokens, which could also affect the results.

8. FUTURE WORK

We are still working on the manual annotations, and aim towards at least 100,000 words annotated with both linguistic and medical information. More data will make more extensive evaluations of e.g. POS tagging possible. We are also planning formal evaluations of the impact of semiautomatic methods on manual annotation, regarding quality and speed of the annotation process.

The development of an NP chunker and a shallow parser are natural next steps, using output from our part-of-speech tagger, trained on health record data, as input.

Furthermore, the corpus will be enriched with medical annotations in order to approach specific medical information extraction challenges.

9. ACKNOWLEDGEMENTS

The authors wish to thank The Norwegian EHR Research Centre (NSEP) for financing the annotation project. We also thank Torbjørn Nordgård for comments and suggestions.

REFERENCES

BAKKEN, C. 2006. Fastlegeordningen en suksess. *Tidskr Nor Lægeforen*, 126(6):814.

BRANTS, S., S. DIPPER, S. HANSEN, W. LEZIUS, AND G. SMITH. 2002. The TIGER Treebank. In *Workshop on Treebanks and Linguistic Theories (TLT)*, Sozopol.

BRANTS, T. 2000. TnT - a statistical part-of-speech tagger. In *NAACL/ANLP*.

CAMPBELL, D. AND S. JOHNSON. 2001. Comparing syntactic complexity in medical and non-medical corpora. *Proc AMIA Annu Fall Symp*, pages 90–94.

EDSBERG, O., Y. NYTRØ, AND T. B. RØST. 2007. Novelty detection in patient histories: Experiments with measures based on text compression. In *7th International Symposium on Intelligent Data Analysis*, Ljubljana, Slovenia.

EJERHED, E., G. KÄLLGREN, O. WENNSTEDT, AND M. ÅSTRÖM. 1992. The linguistic annotation system of the stockholm-umeå corpus project. Technical report, Umeå University.

FISZMAN, M., W. CHAPMAN, S. EVANS, AND P. HAUG. 1999. Automatic identification of pneumonia related concepts on chest x-ray reports. In *AMIA Symp*, pages 67–71.

GIUSE, D. AND A. MICKISH. 1996. Increasing the availability of the computerized patient record. In *AMIA Fall Symp*, pages 633–637.

GOLDMAN, J. A., W. W. CHU, D. S. PARKER, AND R. M. GOLDMAN. 1999. Term domain distribution analysis: a data mining tool for text databases. *Methods Inf Med*, 38(2):96–101. Journal Article.

HAHN, U. AND J. WERMTER. 2004. High-performance tagging on medical texts. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 973,

Geneva, Switzerland.

HONIGMAN, B., P. LIGHT, R. M. PULLING, AND D. W. BATES. 2001. A computerized method for identifying incidents associated with adverse drug events in outpatients. *Int J Med Inform*, 61(1):21–32. Journal Article.

HRIPCSAK, G., S. BAKKEN, P. STETSON, AND V. PATEL. 2003. Mining complex clinical data for patient safety research: a framework for event discovery. *Journal of Biomedical Informatics*, 36(1/2):120–130.

HRIPCSAK, G., C. FRIEDMAN, P. O. ALDERSON, W. DUMOUCHEL, S. B. JOHNSON, AND P. D. CLAYTON. 1995. Unlocking clinical data from narrative reports: A study of natural language processing. *Ann Intern Med*, 122(9):681–688.

HRIPCSAK, G., G. KUPERMAN, AND C. FRIEDMAN. 1998. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods of Information in Medicine*, 37(1):1–7.

HUSETH, O. 2005. Automatisk ordklassetagging og grafem-fonemoversettelse med skjulte markovmodeller.

IEZZONI, L. 1997. Assessing quality using administrative data. *Ann Intern Med*, 127(8 Pt 2):666–674.

JOHANNESSEN, J. M. B. AND H. HAUGLIN. 1998. An automatic analysis of norwegian compounds. In *16th Scandinavian conference of linguistics*, Turku/Åbo.

JOHNSON, S. B., S. BAKKEN, D. DINE, S. HYUN, E. MENDONÇA, F. MORRISON, T. BRIGHT, T. VAN VLECK, J. WRENN, AND P. STETSON. 2008. An electronic health record based on structured narrative. *J Am Med Inform Assoc*, 15(1):54–64.

MACDONALD, C. J. 1997. The barriers to electronic medical record systems and how to overcome them. *J Am Med Inform Assoc*, 4(3):213–221.

MARCUS, M. P., B. SANTORINI, AND M. A. MARCINKIEWICZ. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

MURFF, H. J., A. J. FORSTER, J. F. PETERSON, J. M. FISKIO, H. L. HEIMAN, AND D. W. BATES. 2003. Electronically screening discharge summaries for adverse medical events. *J Am Med Inform Assoc*, 10(4):339–350. Evaluation Studies Journal Article.

NORDGÅRD, T. 2000. Norkompleks. a norwegian computational lexicon. In *COMLEX-2000*, Patras, Greece.

PAKHOMOV, S. V., A. CODEN, AND C. G. CHUTE. 2006. Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, 75(6):418–429.

POWSNER, S. M., J. C. WYATT, AND P. WRIGHT. 1998. Opportunities for and challenges of computerisation. *Lancet*, 352(9140):1617–1622.

RABINER, L. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

RAMSHAW, L. AND M. MARCUS. 1995. Text chunking using transformation-based learning. In D. Y. Church and Kenneth, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94. Association for Computational Linguistics, Somerset, New Jersey.

RØST, T. B., O. EDSBERG, A. GRIMSMO, and y. NYTRØ. 2007. Comparing medical code usage with the compression-based dissimilarity measure. In *12th World Congress on Health (Medical) Informatics - Building Sustainable Health Systems*, Brisbane, Australia.

RØST, T. B., Y. NYTRØ, AND A. GRIMSMO. 2006. Classifying encounter notes in the primary care patient record. In B. Stein and O. Kao, editors, *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, volume 205, pages 1–5, Riva del Garda, Italy, CEUR-WS.

SHARDA, P., A. K. DAS, T. A. COHEN, AND V. L. PATEL. 2006. Customizing clinical narratives for the electronic medical record interface using cognitive methods. *Int J Med Inform*, 75(5):346–368.

SKUT, W., T. BRANTS, B. KRENN, AND H. USZKOREIT. 1993. A linguistically interpreted corpus

of german newspaper text. In *1st Conference on Linguistic Resources*, Dictionnaires electroniques et analyse automatique de textes: le systeme INTEX, pages 705–712, Granada, M. Silberztein.

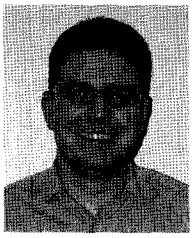
SPYNS, P. 1996. Natural language processing in medicine: an overview. *Methods Inf Med*, 35(4-5):285–301, Journal Article Review.

VAN WALRAVEN, C., A. LAUPACIS, R. SETH, AND G. WELLS. 1999. Dictated versus database-generated discharge summaries: a randomized clinical trial. *CMAJ*, 160(3):319–326.

WALSH, S. H. 2004. The clinician's perspective on electronic health records and how they can affect patient care. *BMJ*, 328:1184–1187.

WEED, L. L. 1969. Medical Records, *Medical Education and Patient Care. The Problem-Oriented Record as a Basic Tool.* Case Western Reserve University Press, Cleveland.

Thomas Brox Røst graduated with a Master of Science degree from the Department of Computer and Information Science at the Norwegian University of Science and Technology in 2003. He is currently a research fellow at the same department, working towards a degree in Health Informatics. His main research focus is on the use of natural language processing and text mining techniques for querying and extracting information from free-text in the general practice patient record.



Ola Huseth received the Master of Philosophy degree in linguistics from the Department of Language and Communication Studies at the Norwegian University of Science and Technology (NTNU) in 2005. He has since then worked on research projects at NTNU on machine translation and health informatics, with a focus on shallow computational linguistic methods. He is currently employed at ErgoGroup AS as a consultant in the field of Business Intelligence.

Øystein Nytrø is associate professor of computer science at the Department of Computer and Information Science at the Norwegian University of Science and Technology. He is head of the Program for Health Informatics. Analysis, representation and modeling of care and patient processes are among his main research interests.

Anders Grimsmo Professor in community medicine. General Manager at the Norwegian Electronic Health Record Research Centre (Norwegian University of Science and Technology, Trondheim, Norway). His research work has evolved from the domains of health promotion and prevention with a focus on chronic care to health informatics. Among others he is member of the National ICT steering committee for ICT projects in Specialist and Hospital Health Care Services and the Ministry of Health's advisory group on Health Informatics.