De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields

Hercules Dalianis
Department of Computer and Systems Sciences
DSV/KTH-Stockholm University
Forum 100, 164 40 Kista, Sweden
hercules@dsv.su.se

Sumithra Velupillai
Department of Computer and Systems Sciences
DSV/KTH-Stockholm University
Forum 100, 164 40 Kista, Sweden
sumithra@dsv.su.se

Abstract

In order to perform research on the information contained in Electronic Patient Records (EPRs), access to the data itself is needed. This is often very difficult due to confidentiality regulations. The data sets need to be fully de-identified before they can be distributed to researchers. De-identification is a difficult task where the definitions of annotation classes are not self-evident. We present work on the creation of two refined variants of a manually annotated Gold standard for de-identification, one created automatically, and one created through discussions among the annotators. These are used for the training and evaluation of an automatic system based on the Conditional Random Fields algorithm. Evaluating with four-fold cross-validation on sets of around 4-6 000 annotation instances, we obtained very promising results for both Gold Standards; F-score around 0.80 for a number of experiments, with higher results for certain annotation classes. Moreover, 49 false positives that were verified true positives were found by the system but missed by the annotators. Our intention is to make this Gold standard available for other research groups in the future. Despite being slightly more time-consuming we believe the manual consensus gold standard is the most valuable for further research. We also propose a set of annotation classes to be used for similar de-identification tasks.

1 Introduction

Health related texts and specifically Electronic Patient Records (EPRs) are an abundant source of valuable information for both clinicians, computer scientists and linguists. Text mining tools, for instance, could be developed by computer scientists for the exploration of such information rich resources. Clinicians could use these text mining tools both on individual patient cases as well as on whole EPR corpora, to find previously unknown information. Moreover, linguists could use such resources to make interesting stylistic and empirical analyses on EPR language.

We have access to a very large EPR corpus, the Stockholm EPR Corpus, containing clinical texts written in Swedish (Dalianis et al., 2009). The Stockholm EPR Corpus contains over one million patient records from over 2 000 clinics. We strive to make this corpus available for a larger research community encompassing researchers in both computational linguistics and medical informatics as well as to practicing clinicians.

In order to develop methods that exploit the vast amount of information contained in EPRs, researchers need to be able to access the data itself. This is often difficult, as such data sources are often restricted due to confidentiality reasons and the like. EPR corpora contain information that can reveal the identity of the patients and hence sensitive information about the individual patient. To remove the information that can identify the individ-

ual patient one needs to de-identify the patient records.

De-identification is an extremely important and difficult task, and many questions arise. What constitutes identifiable information? How much information can be removed (or replaced), ensuring patient integrity and still keeping important information? Moreover, manually de-identifying large resources such as the Stockholm EPR corpus in its entirety manually is not feasible, therefore automatic methods are needed. For the evaluation and training of automatic systems, manually annotated Gold standards are needed. One issue that arises is how large training set does a trainable system require in order to obtain high results? Furthermore, an interesting question to analyse is whether the merging of conceptually similar annotation classes will increase results.

In this paper we describe work on de-identification of Swedish EPRs. We have two aims; (1) refining an existing manually annotated Gold standard for de-identification purposes, one automatically refined and the other (semi-)manually refined, and (2) initiating experiments on using these refinements to evaluate an automatic machine learning system based on the Conditional Random Fields (CRF) algorithm. We have analysed the annotation classes used for de-identification and identified issues that are complicated and need further refinement.

2 Previous research

Using manually annotated resources for Natural Language Processing (NLP) and Information Access (IA) research is very common. Such resources are useful for at least two purposes; for empirical studies on the topic the annotations cover, and for developing and evaluating computational models. It is, however, time-consuming and costly to create such resources. Moreover, for the resources to be useful in an automated system, the annotations must be well-defined and reliable. For an annotated resource to be considered reliable, one must ensure that the annotations have high agreement among the annotators (Artstein and Poesio, 2008).

The de-identification task is very similar to the Named Entity Recognition (NER) task, which has been successfully used for NLP and IA systems. There exist quite a few resources that have been annotated for these purposes, such as for the MUC conferences (Message Understanding Conferences, Grishman and Sundheim (1996)). However, as pointed out by Fort et al. (2009), the fundamental question of defining which annotations such systems should be able to handle, and how the annotators interpret these definitions, is often not addressed. For de-identification, defining the identifiable instances and their scope is a very important issue. In Table 1, an overview of the annotation classes used for de-identification tasks are shown. As can be seen, several different ways of defining identifiable instances in EPRs have been employed by different research groups. Moreover, de-identifying clinical corpora pose specific problems, as such corpora have properties that differ from other types of text, mainly in grammaticality and in levels of noise. Wang (2009) describes work on annotating clinical corpora for Named Entities. Although the work is not intended for the purpose of de-identification, similarities in the annotation task for such language use is presented. For instance, problematic aspects such as variants in the representation of entities are discussed.

Automatic de-identification systems are mainly of two types; rule- and dictionary based or based on machine learning algorithms. There exist many different de-identifiers developed for English clinical text, for example, rule-based systems such as the Scrub system (Sweeney, 1996), the de-identification software engine described in Gupta et al. (2004), and De-id (Neamatullah et al., 2008). De-id is evaluated on a gold standard of 1 836 nursing notes containing 300 000 tokens. For other languages, rule based de-identification systems have also been developed, for instance Medina for French (Grouin et al., 2009) and the Kokkinakis and Thurin (2007) system for Swedish. Statistical or machine learning based de-identification systems for English include Stat De-id (Uzuner et al., 2008). Seven de-identification systems (including one rule-based system and Stat De-id) are described in Uzuner et al. (2007). These systems are used in the i2b2 challenge which consists of 889 discharge letters containing 470 000 tokens for training and 140 000 tokens for testing respectively. The training corpus contained 14 000 annotation instances distributed over eight annotation classes. One of the highest performing systems in Uzuner et al. (2007) used the machine learning

algorithm Conditional Random Fields (CRF), obtaining an F-score > 0.95.

Different machine learning algorithms are better suited for different classification problems. In both Uzuner et al. (2008) and Li et al. (2008) Conditional Random Fields (CRF) and Support Vector Machines (SVM) are compared for the task of classifying entities in clinical text. In Li et al. (2008) both algorithms are applied on a small subset of clinical text written in English. The algorithms were trained on 1 265 annotations and evaluated on 292 annotations. The results show that CRF outperforms SVM for these types of classification tasks, producing an F-score of 0.86 and 0.64 respectively.

However, all systems mentioned above use resources that are annotated with different annotation classes, in many cases with different granularity (see Table 1), and results from the different de-identification systems are therefore difficult to compare. Moreover, the resources are gathered from different types of clinical corpora (discharge letters, pathology reports, etc.), and both language use and number of identifiable instances may differ greatly, which makes results even more difficult to compare across systems. Also, portability to other languages is difficult to ensure, as language differences may affect system performance.

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Annotation classes	First Name	Last Name	Name	Clinician Name	Proxy/Relative Name	Patient Name	Patient Name initial	Hospital	Disease	Pharma/ names	Measure	Organization	Adress	Location	Employer	Job title	Phone Number	e-mail address	Social security number	Medical record number	ID	Account number	Date	Year	Age	Age over 89	Ethinicty/ Holiday	Undefined
Gupta et al. 2004			X		X								X		X	X	X	X	X	X		X	X					
Uzuner et al. 2007				X		X		X						X			X	X			X		X					
Kokkinakis & Thurin 2007 (Swedish)			X						X	X	X	X		X									X					
Neamatullah et al. 2008				X	X	X	X							X			X							X	X	X	X	X
Grouin et al 2009 (French)	X	X												X									X		X			

Table 1. An overview of annotation classes used for de-identification by different research groups on clinical corpora.

3 Refinement of a Gold Standard

We have previously started the process of creating a Gold standard for de-identification of the Stockholm EPR Corpus. Three annotators annotated 100 patient records containing both free text and structured information, encompassing a total of 380 000 tokens. Identifiable instances were defined for the 18 Protected Health Information (PHI) classes given in HIPAA (2003), with some changes. In total 38 annotation classes were used, including four nested classes and some additional classes. The creation of the Gold standard, the annotation guidelines and the resulting set of annotation classes is described in Velupillai et al. (2009).

The average Inter-Annotator Agreement (IAA) result for all instances of the annotation classes on the Gold standard was 0.65 F-score. Some classes showed higher agreement than others, and the total number of annotations differed between the annotators. The approach taken for the creation of the Gold standard was deliberately coarse and loosely defined, for the purpose of getting an initial idea of what type of identifiable instances the EPRs actually contain. The Gold standard has been further analysed in the work presented here, and used for the creation of two refined consensus sets.

3.1 Automatic Consensus Gold Standard

Our first approach to refine the Gold standard was to automatically create a union of all three annota-

tion sets. One requirement for evaluating a de-identification system is that high recall is preferable over high precision, therefore we took the union of all annotations. Whenever there was a mismatch found, majority decision was prioritized. If two annotations covered almost the same instance, the longest instance span was chosen.

Moreover, as many classes were mismatched, a semi-automatic decision on resolving these discrepancies was made (if it could not be resolved by majority decision). For example, if an instance was annotated only by two annotators, and one annotator annotated the instance as *Clinician_First_Name* and the other as *First_Name*, the instance was annotated as *Clinician_First_Name*. Rules for resolving such cases were written after analyzing common mismatches for all annotation classes. All instances that were annotated only by one annotator were also included in the final set of annotation instances. This process resulted in a total amount of 6 170 annotation instances.

As many of the annotation classes are conceptually similar, several variants of merging similar (and removing some infrequent) classes were also made. This was done in order to evaluate whether the automatic classifier would perform better on more general, merged annotation classes.

3.2 Manual Consensus Gold Standard

By creating pairwise matrices covering the total amount of annotations for each annotator, as well as an agreement table (Di Eugenio and Glass, 2004), covering all annotated instances and their number of assigned class judgments, a better overview of the class distributions, annotation instances and annotator judgements was obtained. In total, over 7 000 instances were annotated. However, the total amount of annotations per annotator could differ with over 1 000 instances. Many of these differences were due to boundary discrepancies and class mismatches.

In general, the distribution of annotation instances was very skewed. The annotation class *Health_Care_Unit* contained, by far, the largest amount of annotation instances. Some of the HIPAA classes were not present at all in the data set, such as *Social_Security_Number* and *Medical_Record_Number*. Only 28 of the defined 38 annotation classes were used for annotation. IAA was highest for the Name classes, see Velupillai et al. (2009).

The analysis of the pairwise matrices and the agreement tables resulted in the identification of some differences in the interpretation of the guidelines. In particular, the use of the annotation class *Health_Care_Unit* differed greatly with a very low IAA, see Velupillai et al. (2009). These discrepancies were discussed jointly by the group of annotators and resulted in a more refined set of guidelines.

The main changes to the guidelines were the following:

- An instance should never be tokenized by the annotator. For example, *34-årig* (Aged 34) should be annotated in its entirety.
- The *Relation* and *Ethnicity* classes were deleted. The annotators judged that these classes did not pose a high risk of identifying individual patients.
- The classes *Street_Address*, *Town*, *Municipality*, *Country* and *Organization* were merged into the more general class *Location*. Many of these classes were confused in the individual sets of annotations but covered the same instances. Moreover, the largest possible span should always be used for such instances. An address such as *Storgatan 1, 114 44 Stockholm* should be annotated in its entirety.
- Dates should never include weekdays. The division between *Date_Part* and *Full_Date* should be kept.
- Health care units should be annotated with the largest possible span, and should only be annotated if they denote a specific unit. General units that are not identifiable in themselves should not be annotated.

As stated above, the class *Health_Care_Unit* was the most problematic. In the EPRs, these instances could be mentioned in a variety of ways. Moreover, in the Stockholm area, many health care units have names that include their location. *Karolinska Universitetssjukhuset* (Karolinska University Hospital), for example, is located both in Huddinge and Solna, and the respective locations are included in their names. In the EPRs, these hospitals (and clinics within these hospitals) could be mentioned as for example:

Karolinska Univ. Sjukh, Huddinge

Karolinska/Huddinge

Avd. 11 på Karolinska

Moreover, in some cases, the hospital was referred to as *Karolinska i Solna* (Karolinska in Solna), where *Solna* in this case denotes a *Location*. Following the new guidelines, the longest span possible should always cover the instance, but only if the referred unit was specific. A general unit such as *Geriatriken* (the Geriatrics department) should not be annotated if it was not specified by its hospital. The definition of a general unit has, however, not been specified in detail but is left to be judged by the annotators. Such instances may still be a source of error. A new, refined Gold standard was created semi-automatically after resolving these differences. Many annotations in the initial Gold standard did not conform to the new guidelines (weekdays annotated as *Date_Part* and generic health care units for instance) and were deleted. This resulted in a total amount of 4 423 annotation instances.

4 Using the Consensus Gold Standards with a CRF Classifier

We have used the two created Consensus Gold standards to train and evaluate a Conditional Random Fields (CRF) classifier. As discussed above, such classifiers have shown very promising results for de-identification classification tasks.

We have used the Stanford Named Entity Recognizer (Finkel et al., 2005), using the default settings for all experiments.

	Exact matches						
	Class	**Relevant**	**Retrieved**	**Corpus**	**Precision**	**Recall**	**F-score**
	Account_Number	0	0	1	NaN	0.000000	NaN
AGE	**Age**	40	45	54	0.888889	0.740741	0.808081
AGE	**Age_Over_89**	0	0	3	NaN	0.000000	NaN
	Biometric_Identifier	0	0	4	NaN	0.000000	NaN
DATE	**Date_Part**	629	682	843	0.922287	0.746145	0.824918
DATE	**Full_Date**	338	427	503	0.791569	0.671968	0.726882
DATE	**Year**	15	22	62	0.681818	0.241935	0.357143
PERSON NAME	**First_Name**	0	0	13	NaN	0.000000	NaN
PERSON NAME	**Last_Name**	0	0	5	NaN	0.000000	NaN
PERSON NAME	**Patient_First_Name**	6	14	75	0.428571	0.080000	0.134831
PERSON NAME	**Patient_Last_Name**	0	0	3	NaN	0.000000	NaN
PERSON NAME	**Relative_First_Name**	67	72	128	0.930556	0.523438	0.670000
PERSON NAME	**Relative_Last_Name**	8	8	23	1.000000	0.347826	0.516129
PERSON NAME	**Clinician_First_Name**	541	612	735	0.883987	0.736054	0.803267
PERSON NAME	**Clinician_Last_Name**	706	764	901	0.924084	0.783574	0.848048
LOCATION	**Location**	0	0	3	NaN	0.000000	NaN
LOCATION	**Country**	5	7	27	0.714286	0.185185	0.294118
LOCATION	**Municipality**	3	14	34	0.214286	0.088235	0.125000
LOCATION	**Organization**	0	1	60	0.000000	0.000000	NaN
LOCATION	**Street_Address**	0	0	12	NaN	0.000000	NaN
LOCATION	**Town**	7	12	52	0.583333	0.134615	0.218750
LOCATION	**Health_Care_Unit**	910	1 162	1 747	0.783133	0.520893	0.625645
	Device_Identifier_and_Serial_Number	0	0	6	NaN	0.000000	NaN
	Ethnicity	0	0	9	NaN	0.000000	NaN
PHONE-NUMBER	**Fax_Number**	0	0	5	NaN	0.000000	NaN
PHONE-NUMBER	**Phone_Number**	56	65	130	0.861538	0.430769	0.574359
	Relation	458	473	714	0.968288	0.641457	0.771693
	Uncertain	0	1	18	0.000000	0.000000	NaN
	Total	3789	4 381	6 170	0.864871	0.614100	0.718226

Table 2. Results on the initial experiment using all 28 classes from the automatic Consensus Gold standard, giving results on exact matches. The different divisions show which classes have been merged for the remaining six experiments on the automatic Consensus Gold standard. Same- and similarly-marked annotation classes were merged.

All experiments have been evaluated with four-fold cross-validation (Kohavi 1995), where the total set has been split into four equally sized subsets used for training and evaluation.

Seven experiments using the automatic Consensus Gold standard were reported, each with different mergings of the annotation classes into more general classes and two experiments using the manual Consensus Gold standard, one evaluated with ten-fold cross-validation. No nested annotation classes were used.

4.1 Results: Automatic Consensus Gold Standard

In the initial experiment, all 28 annotation classes are used in the classifier (see Table 2). Some annotation classes contained very few annotations. Four-fold cross-validation on the 28 annotation classes, 6 170 annotation instances and 380 000 tokens in total took around 8 hours to execute on a server with Dual CPU Quad Core Intel Xeon E5405, 2.0 GHz with total 8 kernels and 16 Gb RAM.

By consecutively merging conceptually similar annotation classes, we tried to examine whether the classifier would increase the recall results as well as the overall performance. In the final experiments, all annotation instances are merged into one general PHI class. In Table 3 we see that, for all experiments, precision is very high when looking at the results for both exact and partial matches. Recall, on the other hand, is much lower for exact matches than for partial matches. For de-identification purposes high recall is preferable over high precision, since it is more important to ensure a minimal risk of identification possibilities rather than ensuring trustworthiness of identified instances. Merging all annotation classes into one, general PHI class gives the highest F-score for partial matches. However, for exact matches, experiments 3 and 4 (using 16 or 13 annotation classes, respectively) give the best results.

It seems that the drop in performance for exact matches between experiment 4 and 5 mainly originates in a heavy overgeneration of names, where First_Name and Last_Name are grouped in the more generic class Name. However, looking at partial matches, the drop is not as dramatic, which indicates that there is some boundary problem here which might be due to initials or titles.

One conclusion is that conceptually similar annotation classes can be merged successfully, but not into too general classes. The amount of training instances for each class naturally affect results.

					Exact matching			Partial matching		
Experiment	Annotation classes	Relevant	Retrieved	Corpus	Precision	Recall	F-score	Precision	Recall	F-score
1	28	3789	4 381	6 170	0.864871	0.614100	0.718226	0.910709	0.645112	0.755240
2	22	3910	4 439	6 171	0.880829	0.633609	0.737041	0.925873	0.664288	0.773565
3	16	4013	4 501	6 160	0.891580	0.651461	0.752837	0.939449	0.683472	0.791273
4	13	4016	4 498	6 135	**0.892841**	**0.654605**	**0.755384**	0.940548	0.686649	0.793790
5	9	3147	4 050	5 654	0.777037	0.556597	0.648599	0.933757	0.673440	0.782517
6	7	3150	4 053	5 642	0.777202	0.558313	0.649819	0.933730	0.675193	0.783689
7	1	3040	4 177	5 453	0.727795	0.557491	0.631360	**0.937036**	**0.714620**	**0.810852**

Table 3. Results on all seven experiments using the automatic Consensus Gold standard. The total average over all classes is given. For each new experiment, conceptually similar annotation classes are merged into a more general annotation class. The final experiment shows the results of merging all annotation classes into one general PHI class.

4.2 Results: Manual Consensus Gold Standard

The manual Consensus Gold standard contained a smaller total amount of annotation instances. When using this set in the CRF classifier, we merged all name classes into the generic *First_Name* and *Last_Name* respectively. We also merged *Age* and *Age_Over_89* into one generic *Age* class. The annotation classes *Full_Date, Date_Part, Health_Care_Unit, Location,* and *Phone_Number* were also used. In Table 4, the results on using this set are given. We see that the overall results are similar to the results on using the automatic Consensus Gold standard. However, given the smaller total amount of annotation instances, these results may be interpreted as being a bit better. In particu-

lar, the classes *Date_Part* and *Phone_Number* show much better results on the manual Consensus Gold standard (compare with Table 1). The results for the classes *Health_Care_Unit* and *Location* are, for all experiments, relatively low. This is probably due to the ambiguous nature of many of the instances (i.e. *Huddinge* as a *Location* or *Health_Care_Unit*). Moreover, as can be seen in Table 1, the initial classes *Street_Address*, *Town*, *Municipality*, and *Organization*, that have been merged in the manual Consensus Gold standard had few instances respectively. With more training instances, these results might improve. In Table 5 we see the results on evaluating the CRF classifier (using the manual Consensus Gold standard) with ten-fold cross-evaluation. It is clear that the overall results improve considerably when providing more training data.

4.3 Results: General Discussion

As stated above, it is difficult to compare these results to previous research due to differences in corpora, annotation classes, evaluation methods and also language. However, given the small size of the corpus, we believe that our results are very promising. The lower results for the *Location* and *Health_Care_Units* classes can be compared with the results for the competing systems described in Uzuner et al. (2007), where the results for these classes are consistently lower for all systems. Also, the generally high results for classes covering patients and clinicians can be compared to our high results on the name classes.

Noteable are the general results for exact and partial matches. Naturally, the overall results for exact matches are generally lower, but the differences are not as drastic for the Manual Consensus Gold standard. This indicates that boundaries are difficult to identify for de-identification instances, which was also concluded during the discussions among the annotators, especially for the *Health_Care_Unit* class, and dealt with for the creation of this refined set.

When using manually annotated resources for training and evaluation, it is also interesting to scrutinize the resulting false positives from the classifier. In the experiment using the manual Consensus Gold standard, the Stanford NER also discovered in total 178 false positives, where 49 were actual true positives from the annotation classes. *First_Name*, *Last_Name*, *Location*, *Health_Care_Unit*, *Date_Part* and *Full_Date*. Clearly, the human annotators missed out on identifiable information!

				Exact matches			Partial matches		
Class	**Relevant**	**Retrieved**	**Corpus**	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Age	37	41	56	0.902439	0.660714	0.762887	0.947735	0.719388	0.817923
Date_Part	605	643	710	0.940902	0.852113	0.894309	0.943721	0.854829	0.897078
Full_Date	332	427	500	0.777518	0.664000	0.716289	**0.922879**	**0.791425**	**0.852112**
First_Name	676	718	923	0.941504	0.732394	0.823888	0.944290	0.733686	0.825771
Last_Name	715	757	929	**0.944518**	**0.769645**	**0.848161**	0.953967	0.776767	0.856296
Health_Care_Unit	455	594	1021	0.765993	0.445642	0.563467	0.903736	0.499340	0.643260
Location	44	65	148	0.676923	0.297297	0.413146	0.720513	0.311562	0.435015
Phone_Number	65	70	136	0.928571	0.477941	0.631068	0.949649	0.490950	0.647273
Total	2929	3315	4423	0.883560	0.662220	0.757043	0.932133	0.692842	0.794869

Table 4. Results on the manual Consensus Gold standard using four-fold cross-evaluation.

				Exact matches			Partial matches		
Class	**Relevant**	**Retrieved**	**Corpus**	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Age	37	45	56	0.822222	0.660714	0.732673	0.904762	0.778061	0.836642
Date_Part	617	654	710	0.943425	0.869014	0.904692	0.946196	0.871730	0.907438
Full_Date	342	426	500	0.802817	0.684000	0.738661	**0.931665**	**0.802106**	**0.862045**
First_Name	713	749	923	0.951936	0.772481	0.852871	0.954606	0.773772	0.854729
Last_Name	777	816	928	**0.952206**	**0.837284**	**0.891055**	0.961653	0.845484	0.899835
Health_Care_Unit	559	689	1021	0.811321	0.547502	0.653801	0.921497	0.608116	0.732705
Location	54	73	148	0.739726	0.364865	0.488688	0.778539	0.379129	0.509933
Phone_Number	80	86	135	0.930233	0.592593	0.723982	0.954195	0.613105	0.746535
Total	3179	3538	4421	0.898530	0.719068	0.798844	0.941190	0.751441	0.835680

Table 5. Results on the manual Consensus Gold standard using ten-fold cross-evaluation.

The automatic consensus took around one and half working week of implementation including some manual work and the manual consensus took around two and a half weeks of work including some programming. The advantage of the manual consensus creation was having control over the full process while in the automatic consensus previous errors and discrepancies were not handled.

5 Conclusions

Fully de-identified EPR corpora are very important resources for the research community. With these, development of new methods for exploiting and exploring the valuable information contained in such data sets is possible. Moreover, it enables researchers to compare and evaluate findings in a more reliable manner. We have refined an existing de-identification Gold standard into two Consensus Gold standards. The refined Consensus Gold standards have been used in a CRF classifier with promising results. The automatic Consensus Gold standard has resulted in a larger set of annotation instances, where discrepancies have been resolved semi-automatically. The creation of this set required less cost in time, but contains more noise.

The manual Consensus Gold standard is the result of discussions within the group of annotators, and a new set of guidelines has been developed for future similar annotation tasks. In the end, the group of annotators settled for using the following annotation classes in the future: *Age, First_Name, Last_Name* (these are further refined for *Patient*, *Clinician* and *Relative*)*, Date_Part, Full_Date. Location Health_Care_Unit, Phone_Number, E-mail_address* and *Social_Security_Number*. This set has passed through two iterations of thorough reviews, and our intention is to make this set available for a broader group of researchers in the future

By merging conceptually similar annotation classes, it is possible to automatically refine an existing Gold standard with somewhat inconsistent annotations and improve results, but better results, both for exact and partial matches, are obtained by systematically identifying inconsistencies (through analysis and discussions) and refining the annotations thereafter. For this work, we conclude that, despite the slightly more costly procedure of refining an existing set of de-identification annotations manually, the resulting set is more reliable for further research.

However, as the size of the corpus is relatively small, and the amount of instances for some annotation classes is very low, more training material would be needed in order to produce more stable results. Some annotation classes such as social security number and patient names will probably be very scarce in the EPRs. We will therefore need other approaches to capture these annotation instances. One possibility is to use a rule- and dictionary based method for de-identification of such instances.

In our experiments with CRF we have used the default settings of Stanford CRF for all experiments. Further analysis on and evaluation of useful and extended features as well as weighting schemes for this specific classification problem is needed.

Defining annotation classes for de-identification is difficult. Moreover, EPRs contain a language use which is very noisy and rich in variations of expressions. Such properties makes clear definitions on boundaries and coverage of annotation classes complicated. We have further outlined the criteria needed for the creation of an annotated EPR corpus for de-identification, but many questions may still arise in the future.

We believe that the resulting set of annotation classes obtained after discussions is useful for similar tasks, as it covers the most important identifiable instances. However, even if it would be possible to guarantee optimal performance for these classes, it is impossible to ensure that no individual patient can be identified remaining from the information contained in an EPR.

Acknowledgments

Many thanks to Andreas Amsenius for fantastic programming and for creating a good atmosphere, and to Gunnar Nilsson at Karolinska Institutet for exhaustive annotations and invaluable comments on the consensus discussion. Finally, thank you Martin Hassel for giving observant and important comments on the article! Many thanks also to the anonymous reviewer who gave us invaluable comments on this article.

References

Ron Artstein and Massimo Poesio 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics.* 34, 4 (Dec. 2008), 555-596. DOI= http://dx.doi.org/10.1162/coli.07-034-R2

Hercules Dalianis, Martin Hassel and Sumithra Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of the 14th International Symposium for Health Information Management Research*, Kalmar, Sweden, 14-16 October, 2009, pp 243-249.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Jenny R. Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* 363-370, Ann Arbor, Michigan, June 25 - 30, 2005.

Karen Fort, Maud Ehrmann and Adeline Nazarenko. 2009. Towards a Methodology for Named Entities Annotation. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009*, 142–145, Suntec, Singapore, 6-7 August 2009.

HIPAA 2003, Health Insurance Portability and Accountability (HIPAA), Privacy Rule and Public Health Guidance, 2003. From CD Cand the U.S. Department of Health and Human Services, April11, 2003. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm (accessed October 11, 2009).

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference–6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING),* Copenhagen, Denmark, 1996.

Cyril Grouin, Arnaud Rosier, Olivier Dameron and Pierre Zweigenbaum. 2009. Testing Tactics to localize de-identification. In *Proceedings of 22nd Conference of the European Federation for Medical Informatics (MIE2009)*, Sarajevo, Bosnia and Herzegovina, 2009.

Dilip Gupta, Melissa Saul and John Gilbertson. 2004. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *American Journal of Clinical Pathology* 2004 Feb;121(2):176-86.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selecption. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, edited by Chris. S. Mellish, pp. 1137-1143. San Francisco, CA: Morgan Kaufmann.

Dimitrios Kokkinakis and Anders Thurin. 2007. Identification of Entity References in Hospital Discharge Letters. In *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*. University of Tartu, Tartu, 2007.

Dingcheng Li, Karin Kipper-Schuler and Guergana Savova. 2008. Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. In *BioNLP2008: Current Trends in Biomedical Natural Language Processing*, 94–95, Columbus, Ohio, USA, June 2008.

Ishna M Neamatullah, Margaret Douglass, Li-wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark and Gari D. Clifford. 2008. Automated De-identification of Free Text Medical Records. *BMC Medical Informatics and Decision Making 2008*, 8: 32, doi:10.1186/1472-6947-8-32.

Latanya Sweeney. 1996. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In *Proceedings of The AMIA Annual. Fall Symposium 1996*, 333-337.

Özlem Uzuner, Yuan Luo, and Peter Szolovits: 2007. Evaluating the State-of-the-art in Automatic De-identification. *Journal of the American Medical Informatics Association*, September 2007,14(5): 550-563

Özlem Uzuner, Tawanda C. Sibanda, Yuan Luo and Peter Szolovits 2008. A De-identifier for Medical Discharge Summaries. *Journal of Artificial Intelligence in Medicine*, Jan;42(1):13-35.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel and Gunnar Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics* (2009), doi:10.1016/j.ijmedinf.2009.04.005

Yefeng Wang. 2009. Annotating and Recognising Named Entities in Clinical Notes. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, August, Suntec, Singapore, 2009