Some Remarks on Automatic Semantic Annotation of a Medical Corpus

Agnieszka Mykowiecka, Małgorzata Marciniak

Institute of Computer Science, Polish Academy of Sciences,
J. K. Ordona 21, 01-237 Warsaw, Poland
agn,mm@ipipan.waw.pl

Abstract. In this paper we present arguments that elaborating a rule based information extraction system is a good starting point for obtaining a semantic annotated corpus of medical data. Our claim is supported by evaluation results of the automatic annotation of a corpus containing hospital discharge reports of diabetic patients.

1 Introduction

Many current methods of recognizing various types of information included within natural language texts are based on statistical and machine learning approaches. Such applications need specially prepared domain data for training and testing. Clinical texts are hard to obtain because of privacy laws, in particular, none of the Polish corpora include this type of texts. Corpora available during the past decade more often contain biomedical than clinical texts (e.g. corpora described in [1]). Recently, creating corpora containing clinical data has started to attract much more attention, e.g. (Cincinnati Pediatric Corpus http://computationalmedicine.org/cincinnati-pediatric-corpus-available, [2], or data collected within Informatics for Integrating Biology and the Bedside (i2b2) https://www.i2b2.org/NLP/DataSets/Main.php). This year, the Text REtrieval Conference (TREC) added the Medical Records Track devoted to exploring methods for searching unstructured information in patient medical records. In nearly all existing resources, semantical annotation is absent or very limited. The one of the few exceptions is CLEF [10] which contains cancer patient records annotated with information about clinical relations, entities, and temporal information.

There are two main approaches to the task of annotating new linguistic data – manual annotation, and manual correction of automatically assigned labels. The traditional annotation methodology consists in preparing and accepting annotation guidelines, annotating every text by at least two annotators and finally, resolving differences by a third experienced annotator. This approach, applied to part-of-speech annotation, is described in [9], semantic manual annotation is described in [10] or [12]. Manual annotation is a time-consuming and expensive process, moreover manual work is error-prone. Manually constructed data are very hard to extend and modify – every change imposes extra effort for checking the consistency of the result. Therefore, providing automatic methods to

facilitate the task is very important. Automatic annotation is much faster and although it also does not guarantee complete correctness, the cost of correcting already labeled data is lower than the cost of entirely manual annotation. Automatic annotation of data was applied in the MUCHMORE project [13]. The methods described in [11] can support automatic annotation of textual contents with SNOMED concepts.

A good starting point for automatic annotation are methods of Information Extraction (see [6]) based on regular expressions and lexicons (e.g., [3]), which do not require annotated corpora as machine learning techniques do. In this paper we discuss the results of annotating a corpus of Polish diabetic records with a set of complex semantic labels consisting of about 50 attributes. For this task we reused an already existing rule based IE extraction system. In section 2 we present the method used to create the annotated corpus and the methodology accepted for the evaluation process. Then, in section 3 we describe the results obtained. The paper concludes with a discussion of the evaluation results.

2 Method

2.1 Data description

The corpus consists of 460 hospital discharge reports of diabetic patients, collected from 2001 to 2006 in one of the hospitals in Warsaw. Each document is about 1.5 – 2.5 pages long and written in MS Word. The documents are converted into plain text files to facilitate their linguistic analysis and corpus construction. As the data include information serving identification purposes (names and addresses) they were substituted with symbolic codes before making the documents accessible for analysis. The anonymization task was performed in order to make the data available for scientific purposes.

The entire dataset contains about 1,800,000 characters in more than 450,000 tokens, out of which 55% are words, abbreviations and acronyms, while 45% are numbers, punctuation marks and other symbols.

2.2 Automatic annotation process

In contrast to many annotated text corpora which were built by manually assigning labels to appropriate text fragments, we decided to adopt an existing IE system [8] for the task. However, after inspecting the IE system's results it turned out that they do not contain all the information needed. For the IE system, the main goal was to find out whether a particular piece of information is present in an analyzed text, while the task of text annotation requires identifying the boundaries of text fragments which are to be assigned a given label. To solve the problem, the idea of combining two extraction grammars was introduced. On the basis of the existing grammar a simplified version, consisting of a subset of the original rules, was created. The final information associating text fragments with semantic labels is the effect of a comparison of the results of these correlated IE grammars. The limits of text fragments representing attribute values

are recognized in the simplified grammar, while their correctness is justified by more complex grammar rules which describe the contexts in which a particular phrase has a desired meaning. Thus, the annotation process (described in detail in [4]) consists of the following steps:

- parsing the text with the existing full extraction grammar,
- parsing the entire text using the simplified grammar,
- removing unnecessary information from the output of both grammars,
- comparing and combining the results – only structures that are represented in both results are represented in the final corpus data together with information on boundaries of the entire phrase and its subphrases,
- combining the semantic information with morphological information (see [5]) to create a set of corpus XML files,
- manual correction of annotations.

2.3 Annotated data

Within the semantic annotation layer about 50 simple attributes, 11 complex structures and 3 lists types are defined. Below, they are described in the same groups as the evaluation of the annotation given in the table (1).

- Identification of a patient's visit in hospital: visit identification number and information if it is a main document or a continuation; the date of the document; dates when the hospitalization took place.
- Patient information: a structure with the patient's identifier, sex and simple attributes representing age, height, weight (in numbers or words) and BMI.
- Data about diabetes (in some cases grouped in a *feature_l_str* structure), e.g.: type (D_TYPE); if the illness is balanced (D_CONTROL); when diabetes was first diagnosed (expressed as absolute or relative date); reasons of hospitalization (as a list of attributes); and results of basic tests: HbA1c, acetone, LDL, levels of microalbuminury and creatinine.
- Complications, other illnesses including autoimmunology and accompanying illnesses, which may be correlated with diabetes.
- Diabetes treatment described by: *insulin_treat_str* that contains insulin type and its doses; description of continuous insulin infusion therapy (*ins_inf_treat*); description of oral medications; information that insulin therapy was started. The applied therapy is sometimes given as a list of information that is represented by a *cure_l_str* list of attributes.
- Diet description represented by *diet_str* that contains information on type of diet (DIET_TYPE), and structures describing how many calories are recommended and a similar structure representing numbers of meals.
- Information on therapy given in text form, e.g.: patient's EDUCATION, DIET_OBSERVEing, THERAPY_MODIFFication, SELF_MONITORING.

Some of the attributes have values representing dates, e.g. HOSPIT_STRucture has two substructures describing the beginning and the end of a visit in hospital

(H_FROM and H_TO). To correctly label these attributes it is necessary to recognize the different formats of dates and the appropriate contexts indicating the meaning of a date. Dates are also recognized in case of DOCument_BEGining, and for representing date when diabetes was first diagnosed.

Most attributes representing results of tests have numbers as values. They are usually attached to short phrases consisting of an introductory phrase indicating a type of a test and its value, sometimes after one of the following characters: '=, :, -'. Values can also be given in brackets. Only the results of LDL cholesterol levels need a wide context, because they are represented in a table form together with other test results. This explains the average length of 27 tokens of a phrase representing *lipid_str* indicating the context of the LDL attribute.

Some attributes, having boolean values, label relatively short phrases like results of acetone tests. For example, a negative value is attached to the following strings: *ac. (-*, *ac. -*, *ac. /-*, *ac. nieobecny* 'absent', *bez acetonurii* 'without acetone in urine', *ustąpiła acetonuria* or *ustąpienie acetonurii* 'acetone in urine subsided'. Boolean values also have attributes that are represented by many different, sometimes long, phrases. For example, the information if a therapy of diabetes was modified or not is represented in the test set after correction by 23 different phrases of average length of 4.3 tokens.

Attributes of the last group have many values of different types. For example attribute COMPlication has 17 different values. It is usually attached to a short phrase (avg. 2.2 tokens) representing just the complication name. Longer phrases (avg. 5 tokens) represent the opposite information (N_COMP) when a particular complication was not diagnosed or there are no complications. These phrases have to contain a phrase like: *nie wykryto* 'not diagnosed'.

3 Results

In the corpus consisting of 460 patient records, 66165 occurrences of simple attributes were labeled. To check the quality of the results, manual verification of a randomly selected 10% of the corpus (46 records, 46439 tokens) was done by two annotators, who were given the following guidelines:

- Structures should be assigned to continuous phrases. i.e. to all tokens between the first and last tokens of the phrase.
- Boundaries of a phrase to which a label is assigned are determined on the basis of sets of words that may start and end the phrase.
- In case of phrases that represent information that should be taken into account, but were not predicted by the grammar designer, annotators have to rely on their own opinion which words belong to such a phrase. If it is possible, similar rules to those described in the guidelines should be applied.
- Annotators have to point out information that is understandable to human readers, so phrases with spelling errors should be annotated.

The results of the manual corrections of the system's output made by the two annotators were then compared and the agreed version was accepted as a Gold-standard version. The final number of differences between the automatically

obtained annotation and the Gold-standard concerned 596 token labels (1.3%). Human corrections mainly concerned the addition of new labels (79 structures – 554 tokens). Deletion of mistakenly recognized structures were much less frequent (4 structures – 20 tokens); very few changes concerned only the boundaries or the name of the structure. 283 corrections were proposed consistently by both annotators. Kappa coefficient for annotators agreement counted for all word-label pairs was equal to 0.976 if empty labels were counted (for total 46439 occurrences) and 0.966 when they were ignored (9031 occurrences). The agreement between the corrected version and the automatically annotated set was equal to 0.94. Inter-annotator agreement counted only for structures beginnings (3308) was equal to 0.976.

The corrected results were compared with the automatically annotated data. In general, the verification of 9057 non empty labels showed that automatic annotation achieved an accuracy equal to 0.987, precision – 0.995, recall – 0.936 and 0.966 f-measure value. Precision was equal to 1.00 for all attributes but DOC_DAT and COMP and for *dose_str* and *insulin_treat_str* structures. Recall and F-measures values for all attributes and structures which occurred in the evaluation set are given in table 1. Errors can be classified into 3 groups:

- Omissions and mistakes of the system: *dieta cukrzycowa wysokobiałkowa 188 kcal, 3 posiłki* 'diabetic high protein diet 1800 kcal, 3 meals' – we did not recognize a diet of type 'diabetic and high protein'; the system did not label information on an obesity of a patient, when it was expressed in Latin 'obesitas' instead of Polish 'otyłość'.
- Spelling errors or punctuation errors in the original data in words that are crucial for rules: *wlew podsttawowy* instead of *podstawowy* 'base infusion'; pRetinopathia; *masa ciała103* 'weight103'.
- Information represented by phrases not predicted by the extraction grammars, or difficult to label by the system because of ambiguity (examples discussed in section 4).

As evaluations based on verifying system output can be biased towards types of phrases which are recognized by the system and may result in the omission of other types of phrases which represent the same information, we performed a second type of evaluation. We manually compared the automatically generated annotation with a manual annotation which was done without seeing the system results. For this purpose, 5 discharge records randomly selected from the Gold-standard subcorpus were annotated manually. It took a well trained person 250 minutes (correction of automatic annotation took less than 1 hour) and the F-measure of the results in comparison to the Gold-standard annotation was equal to 0.86. Kappa coefficient between manually obtained annotation and the corrected system output was equal to 0.87 when all word-label pairs were counted and 0.82 for structures' beginnings. Lower coefficient value was due to annotator inattention resulted in omission of information or indicating an inappropriate text fragment to a label. The agreement between the corrected version and the automatically annotated set was equal to 0.94.

Table 1. Semantic label diversity and label verification

structure/attribute	numb of occurr. G-std.	numb of occurr. rule based	F-measure	recall	phrase types	avg phrase length	words total
administrative information							
DOC_BEG	46	46	1	1	1	2	92
DOC_DAT	37	38	0.99	1	1	1	298
id_str	46	45	0.99	0.98	2	4	183
ID	46	45	0.99	0.98	–	–	–
CONT	45	45	1	1	–	–	–
hospit_str	46	43	0.97	0.93	7	18	831
H_FROM	46	43	0.97	0.93	2	7	323
H_TO	46	43	0.97	0.93	2	7	323
EPIKRYZA_BEG	46	46	1	1	1	1	46
recommendation_str	44	44	1	1	13	5.6	248
RECOMMEND_BEG	44	44	1	1	1	2	88
basic patient data							
id_pat_str	46	45	0.99	1	11	4	191
id_pat_sex	46	45	0.99	1	3	2.2	100
ID_PAT	46	45	0.99	1	–	–	–
ID_P_SEX	46	45	0.99	1	–	–	–
ID_AGE	46	45	0.99	1	6	2	91
W_IN_WORDS	6	5	0.91	0.83	4	1	6
WEIGHT	40	39	0.99	0.97	6	3.5	138
BMI	33	33	1	1	3	3.6	119
HEIGHT	40	39	0.99	0.97	3	2	80
basic diabetes data							
D_CONTROLL	30	27	0.95	0.90	18	1.8	55
FROM_IN_W	1	0	–	–	1	2	2
HBA1C	59	54	0.96	0.92	8	5	299
ACET_D	42	42	1	1	4	2	85
creatinin_str	43	41	0.98	0.95	7	4.4	191
microalbuminury_str	13	12	0.96	0.92	6	5	65
lipid_str	31	27	0.93	0.87	6	27	834
LDL1	31	27	0.93	0.87	3	2.3	78
feature_l_str	91	91	1	1	59	5.7	518
COMP	5	5	1	1	4	1.6	8
D_CONTROLL	34	34	1	1	9	1.2	40
D_TREAT	24	24	1	1	6	2	49
D_TYPE	70	70	1	1	2	2	139
FROM_IN_W	10	10	1	1	5	1	10
RELATIVE_DATA	19	18	0.97	0.95	7	3	58
W_IN_WORDS	10	10	1	1	4		10
reason_l_str	30	27	0.95	0.90	27	12.3	370
D_CONTROLL	40	37	0.85	0.95	19	2.3	94
KETO_D	2	2	1	1	2	1	2
KWAS_D	1	1	1	1	1	2	2
RELATIVE_DATA	1	1	1	1	1	4	4
SELF_MONITORING	1	1	1	1	1	1	4
complication and acc diseases							
ACC_DISEASE	48	48	1	1	3	1	48
COMP	134	132	0.97	0.96	49	2.2	294
N_COMP	27	15	0.71	0.56	11	5	134
therapy							
insulin_treat_str	446	444	0.99	0.99	103	5.7	2531
I_TYPE	439	436	0.99	0.99	23	1.7	746
dose_str	441	440	0.99	0.99	8	3.1	1363
corr_str	2	1	0.67	0.5	2	6	12
DOSE_MODIFF	2	1	0.67	0.5	1	1	2
THERAPY_MODIFF	2	1	0.67	0.5	1	2.5	5
diet_str	47	44	0.97	0.94	29	7.8	366
DIET_TYPE	47	44	0.97	0.94	4	2.1	100
cal_str	47	44	0.97	0.94	6	2.8	131
CAL_MIN	47	44	0.97	0.94	–	–	–
meals_str	45	41	0.95	0.91	8	2.2	99
MEALS_MIN	45	41	0.95	0.91	–	–	–
ORAL_TREAT	63	63	1	1	18	1.2	75
I_THERAPY_BEG	4	1	0.40	0.25	4	5.3	21
THERAPY_MODIFF	24	19	0.88	0.79	23	4.3	103
DOSE_MODIFF	9	8	0.94	0.89	6	3.3	30
DIET_CORRECTION	2	2	1	1	2	3	6
SELF_MONITORING	0	2	–	–	3	1	3
EDUCATION	27	25	0.96	0.93	20	8	215

4 Discussion and Conclusions

Standard information given as numbers or dates is often easy to recognize automatically by any rule based system. The vast majority of such data is labeled correctly, yet sometimes there are problems as a result of unpredicted long phrases representing the desired information. These errors should be corrected during manual verification of the corpus.

For example, the phrase *HbA1c przy przyjęciu do Kliniki wynosiło 7,8%* 'HbA1c level at the time of admission to hospital was 7.8%' contains information that is usually represented by 'HbA1c = 7,8%'. As rule based systems are greedy, rules have to be relaxed carefully. For example if we allow several tokens between the introductory string *HbA1c* and a number in a rule assigning the HBA1C attribute, it may recognize another number as the value (for *HbA1C 9 %, HbA1 11,3 %* the value assigned would be 11.3%). It is possible to relax the extraction grammar rules by imposing restrictions on tokens that appear between the 'HbA1c' token and its value, e.g. a word which has the base form *przyjęcie* 'admission'.

The second reason for attribute omission is paraphrasing. Natural language allows us to express the same information in many ways. Thus, it is extremely difficult to write a system that correctly recognizes all possible phrases. For instance, in the interpretation of the following phrase: *pacjentka z cukrzycą typu 1 została przyjęta do Kliniki z powodu chwiejnego przebiegu choroby* 'patient with diabetes type 1 was hospitalized in the Clinic because of the unstable course of illness' it is necessary to know that the illness mentioned in the second part of the sentence refers to diabetes, to recognize the reason for hospitalization.

Another example that is easy for a human annotator but caused problems in automatic annotation was when context was disregarded for a test result. We assume that phrases like *cukrzyca typu 2* 'diabetes type 2' indicate type of patient's diabetes. But for the following phrase *pacjent obciążony rodzinnie, mama i babcia z cukrzycą typu 2* 'patient with family burden, mother and grandmother with diabetes type 2' this is not true. Another difficult example is the phrase *dawka dodatkowa 21.00 - 2j. Humalog* 'additional dose 21.00 - 2j Humalog', where the string '21.00' was not recognized as a time description but as a dose.

The biggest problem for automatic rule based semantic annotation stems from phrases that require a very wide context. For example, it is impossible to correctly interpret the following phrase: *Wprowadzono intensywną insulinoterapię* 'Intensive insulin therapy was introduced.' This phrase is a candidate for I_THERAPY_BEG indicating the introduction of insulin into a patient's therapy. Unfortunately, from this phrase we do not know if the verb 'introduce' refers to 'insulin' or to the word 'intensive' — a feature of the therapy. This problem could be resolved only by a human annotator (and not always) after an analysis of other information in the document. For example, if there is information on newly diagnosed diabetes or previous oral therapy, the phrase shall be labeled with the I_THERAPY_BEG attribute, whereas if there is information that patient was cured with continuous insulin infusion therapy, the phrase shall not be labeled with the I_THERAPY_BEG attribute.

The semantic annotation of text corpora is domain and application related. As for a new purpose, a new annotation is usually necessary, all methods of increasing the efficiency of the annotation procedure are very desirable. In the paper we presented the evaluation results of a corpus annotation obtained using IE grammars. The results turned out to be of a quality good enough for statistical purposes [7]. The advantage of designing an IE system instead of preparing only guidelines for manual annotation is its flexibility – the set of rules may be changed and a slightly different resource with a high degree of consistency can be produced, whilst changing the manually annotated resource is more error-prone and time consuming.

References

1. Cohen, K.B., Fox, L., Ogren, P.V., Hunter, L.: Corpus design for biomedical natural language processing. In: ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. pp. 38–45. Detroit (2005), http://www.aclweb.org/anthology/W/W05/W05-1306
2. Dalianis, H., Hassel, M., Velupillai, S.: The Stockholm EPR corpus - characteristics and some initial findings. to be published. In: in Proceedings of the 14th International Symposium for Health Information Management Research. pp. 14–16 (2009)
3. Gold, S., Elhadad, N., Zhu, X., Cimino, J.J., Hripcsak, G.: Extracting Structured Medication Event Information from Discharge Summaries. In: AMIA Annual Symposium Proceedings. p. 237–241 (2008)
4. Marciniak, M., Mykowiecka, A.: Construction of a medical corpus based on information extraction results. Control and Cybernetics (in print) (2011)
5. Marciniak, M., Mykowiecka, A.: Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish. In: Proceedings of BioNLP 2011 (2011)
6. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.: Extracting information from textual documents in the electronic health record: A review of recent research. IMIA Yearbook 2008: Access to Health Information pp. 128–144 (2008)
7. Mykowiecka, A., Marciniak, M.: Automatic semantic labeling of medical texts with feature structures. In: Text, Speech and Dialogue. Proceedings of the TSD 2011, Plzen, Czech Republic, 2011, LNAI, Springer (2011, accepted for publication)
8. Mykowiecka, A., Marciniak, M., Kupść, A.: Rule-based information extraction from patients' clinical data. Journal of Biomedical Informatics 42, 923–936 (2009)
9. Pakhomova, S.V., Codenb, A., Chutea, C.G.: Developing a corpus of clinical notes manually annotated for part-of-speech. International Journal of Medical Informatics 75, 418–429 (2006)
10. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., Setzer, A.: Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics 42(5), 950–966 (2009)
11. Ruch, P., Gobeill, J., Lovis, C., Geissbühler, A.: Automatic medical encoding with SNOMED categories. BMC Medical Informatics and Decision Making 8 (2011)
12. South, B.R., Jones, M., Garvin, J., Samore, M.H., Chapman, W.W., Gundlapalli, A.V.: Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics, 10 (2009)
13. Vintar, S., Buitelaar, P., Ripplinger, B., Sacaleanu, B., Raileanu, D., Prescher, D.: An efficient and flexible format for linguistic and semantic annotation. In: In Third International Language Resources and Evaluation Conference, Las (2002)