The Evaluation of Semantic Tools to Support Physicians in the Extraction of Diagnosis Codes

Regina Geierhofer and Andreas Holzinger

Institute for Medical Informatics, Statistics & Documentation (IMI)
Research Unit HCI4MED
Medical University Graz, A-8036 Graz, Austria
regina.geierhofer@meduni-graz.at,
andreas.holzinger@meduni-graz.at

Abstract. Over the past few years the extraction of medical information from German medical reports by means of semantic approaches and algorithms has been an increasing area of research. Currently, several tools are available that aim to support the physician in different ways. We developed a method to evaluate these tools in their ability to extract information from large amounts of data. We tested two off-the-shelf tools that worked in a background mode. We found that the field of quality management made it necessary that these large amounts of data could be background or batch processed. Additionally, we developed a metric, based on the semantic distance of the ICD codes, in order to improve the comparison of the accuracy of the codes suggested by the tools. The results of our evaluation showed that, at present, the tools are capable of supporting inexperienced physicians, however are still not sophisticated enough to work without human interaction.

Keywords: Human Language Analysis and Natural Language Processing, Evaluation, Semantics.

1 Introduction and Motivation for Research

The coding of diagnoses and procedures in many countries is obligatory in medical practise, because it provides basic information used for the financing and controlling of health care institutions [LKF], [DRG], [1]. In medicine, it is necessary to categorize free text reports for further processing [2], [3], [4], the decision on which code is most accurate is done by the physician. Recently many tools have been developed in order to support physicians and facilitate decision support [5]. Many doctors regard these tools as simply *diagnosis browsers,* which consist of little more than a search engine; few of these tools provide any more functionality. Consequently, these tools are not reliable enough to prepare medicals texts for coding on their own, or reliable enough for the automatic background coding of text.

This is despite the fact that several of these programs are based on network-like structures and are, in principle, capable of analyzing text in a far more sophisticated way that just substring matching [6].

A. Holzinger (Ed.): USAB 2007, LNCS 4799, pp. 403–408, 2007.

Our goal was to develop methods to evaluate such tools with regard their accurateness in the **automatic** analysis of medical texts and their ability at mapping these texts to medical codes. The fact that these texts could be background processed was useful for the following reasons; 1) there were large quantities of reports, 2) it made the performance comparison easier, and 3) because the tools all had different ways in which they interacted with the end user. This interaction can interfere with the accurateness of the final results/codes. Sometimes, users are not able to find a correct code at all, especially true if there are inaccuracies in the knowledgebase or in the filtering/ranking algorithm.

2 Methods and Materials

We used two types of coded reference sets to evaluate two separate off-the-shelf tools. Despite the fact that we would have preferred to have evaluated more tools, only two managed to satisfy our preconditions, specifically: 1) to work with German texts, 2) be capable of background processing, and 3) be available free of charge. However, in order to develop a suitable method of evaluation, it was not necessary to have more than two tools available. Both tools differed in their underlying philosophy in many respects, however, the following were relevant to us:

Different goals: Tool 1 only extracts short phrases from text, sufficient enough to choose a correct diagnosis code, while Tool 2 analyzes the entire text and codes the medical information using semantic axes, analogous to coding in SNOMED [7] .

Different fields: Tool 1 was designed for interactive use as an expert system. It works with a structure based on decision trees; each node is a term or concept. Goal-oriented questions are asked if the extracted information is too limited, according to the rules stored in its structure, in order to make the necessary decisions to reach a leaf (a diagnosis code). These questions form part of the tool's results if used in batch mode. As input it expects short phrases, such a physician's typical diagnosis, or a discharge letter diagnosis.

Tool 2 uses a semantic network. Its focus is not limited to the coding of diagnoses; it provides a knowledge base used by many applications, each with different goals. Goal-oriented questions are not a basic feature of this tool. The tool returns more than one code, both in interactive and in background mode. In principle, medical texts of any length are suitable as input.

Definition of success: In contrast to Tool 2, the first tool has an internal definition of the degree of its success (i.e. it extracts sufficient concepts or terms to reach an unambiguous diagnosis code).

Due to the differences and restrictions of the tools, we chose two types of reference sets: the ICD WHO 2005 descriptions, and sample diagnosis descriptions extracted from various discharge letters.

Scenario 1: ICD descriptions:
The main advantage of using the ICD descriptions is that you have, by definition, an unambiguous code for each description. We used the ICD WHO 2005 German

descriptions that can be downloaded from the German Institute of Medical Documentation and Information (DIMDI). We also allowed alternative descriptions for certain codes (see Table 1).

Table1. Alternative descriptions for ICD-Code D68.0

D68.0 Willebrand-Jürgens-Syndrom	**D68.0 Von Willebrand's disease**
Angiohämophilie	Angiohaemophilia
Faktor-VIII-Mangel mit Störung der Gefäßendothelfunktion	Factor VIII deficiency with vascular defect
Vaskuläre Hämophilie	Vascular haemophilia

If a code's description was not self-documenting, it was replenished according to the rules of the WHO. This task should be easy to compute, due to the fact that these texts already form parts of each tool's respective knowledgebase. Tool 1, however, found that the test data was not suitable; consequently we could not gather any results for scenario 1.

Scenario 2: Diagnosis description:
In this scenario the input consists of several thousand diagnosis descriptions coded by physicians from various medical disciplines. In this scenario, the physician's coding is used for comparison. Their codes, however, cannot be used as a reference value in the same manner as the ICD codes since it is possible that the code chosen by the physician is not the most appropriate code available for the medical text. Therefore, the ICD codes derived from the medical documentation were treated as if they were the results of a third tool.

Scenario 1 and Scenario 2:
In both scenarios we categorized each result as "precise" if the first 4 digits of any of the ICD codes returned by the tools matched the codes provided by us. Discrepancies in the ICD code's 5th digit were not considered, as the WHO itself does not utilize a fifth digit and because the 5th digit differs between the German, Swiss and Austrian editions. The result was considered "imprecise" if and when only the first 3 digits matched the correct classification code for the disease in question. We classified a result as "false" if none of the returned codes (maximum of 10 hits per description) was at least imprecise.

Semantic distance:
Simple string matching provides us a first impression of the quality of the initial results. However, it considers the hierarchical structure of the ICD only; it completely ignores the semantic structure that the ICD provides and the medical closeness of the described disease patterns. This semantic structuring considers the fact that related diseases correspond to codes in various chapters of the ICD.

In the ICD the WHO provides information where these related diseases and codes may be found. We analyzed this structure, converted it to a more suitable form, and

developed a metric. Subsequently, we used an adaptation of the a-star-algorithm to calculate their semantic distance [8].

Based on the categorization method mentioned above we refined the evaluation results using the semantic distances. To exemplify the approach, consider the case of a patient who is allergic to her eye shadow.

One possible coding from a medical point of view is

H01.1 Noninfectious dermatoses of eyelid `Dermatitis: allergic another`
L23.2 Allergic contact dermatitis due to cosmetics.

The ICD defines H01.1 as being directly related to L23. Without any weight at the edges of the graph, the semantic distance is 2 and is equal to the distance between L23.2 and L20-L30 Dermatitis and Eczema. The hierarchical distance of H01.1 and L23.2 would be 7 and equal to the distance between H01.1 and F20.2 Catatonic schizophrenia.

3 Results

Scenario 1:
We were only able to test some versions of the second tool, due to the reasons mentioned above. The results vary from between 84.2% and 95.27%.

Table 2. Results of the evaluation of Tool 2 as per Scenario 1

		Worst version		Best version	
precise		**50928**	**84.20%**	**57623**	**95.27%**
	First suggested code	46316		43851	
	Suggested code #2-10	4612		13772	
imprecise		**1757**	**2.90%**	**850**	**1.41%**
	First suggested code	279		315	
	Suggested code #2-10	1478		535	
false		**5114**	**8.46%**	**1963**	**3.25%**
	First suggested code	665		497	
	Suggested code 2-10	4449		1466	
	all false	1898		637	
	Suggested code <#10	2551		829	
no code		**2683**	**4.44%**	**46**	**0.08%**

Scenario 2:
Applying the same evaluation approach and using the physician's coding as a reference set, we got the following results:

Table 3. Results from the Tools from Scenario 1

	Tool 1	Tool 2
Precise	57%	57%
Imprecise	7%	4%
False	21%	27%
no code	15%	12%

It was not possible to compare the results in the granularity above, because the first tool only returns a single code or none at all. To be as fair as possible, we considered only the first hit returned by the second tool. This is, in part, responsible for the noticeable decrease in Tool 1's precise results.

Deficiencies and peculiarities of the tools were, in some cases, responsible for both tools returning different results or returning results that did not match the physician's coding. Sometimes the knowledge base was not complete; in other cases the processing was stopped too early, and in some cases the direction of the interpretation of the text lead to differing results. Text which could not be unambiguously interpreted due to a lack of information was yet another reason. Unambiguousness, is not normally a problem for a physician as they have more information available to them than the tools. The tools can only work on a given phrase, and have no other information available. Because a physician's coding could be incorrect for a particular text, it cannot be used as an absolute reference such as an ICD description or a gold standard. Consequently we built a number of sets of codes that matched either (a) the codes supplied by the other tool or (b) a physician's coding. Additionally, we used our implementation of a semantic distance metric to further refine the evaluation results. The results are presented in table 4.

Table 4. Results for the 3468 diagnosis descriptions mentioned above

	# codes	Median of the semantic distance (tool 1: physician)	median of the semantic distance (tool 2:physician)
tool 1=tool 2 = physician	1587	1,2	0,6
(tool 1 = tool 2) <> physician (3 digit)	152	6,3	6,6
(tool 2 = physician) <> tool 1 (3 digit)	521	6,7	1,3
(tool 1 = physician) <> tool 2 (3 digit)	641	2,7	7,4
tool 1 <> tool 2 <> physician (3 digit) but all return a code	183	6,3	7,3
tool 1 <> tool 2 <> physician (3 digit) one tool returns no code	247	5,76	7,4
(tool 1 = tool 2 = no code) <> physician	137	---	---

In cases where both tools suggested the same ICD code, these codes were more accurate than the physician's code. After manually checking all of the codes which had a short semantic distance between them, we corrected the results. After correction, the tools had a rate of 70.24% and 76.87% respectively (codes which were either *precise* or *imprecis*e).

4 Conclusion and Future Work

At present, current tools used for the extraction of diagnoses codes are able to produce practical suggestions for diagnoses, especially if the input is short enough. As an interactive support tool for unskilled or inexperienced (novice) physicians, the benefit to the end user can be reasonable. We think that within a few years the use of such tools could be used to facilitate quality control significantly. At present it makes sense to develop methods to evaluate these tools with a systematic and technological approach. These methods also enable to deal with realistic magnitudes of text. Also of importance is the fact that this evaluation method measures objectively. We discovered that it is not easy to evaluate tools objectively against each other when they work in batch mode, due to their different approaches and, most of all, their various interactive designs. It will also be necessary to focus more on large text passages rather than short phrases, and to refine our way of measuring semantic distances (by using weights for the edges, for example). Adjustments, using other semantic interpretations of the ICD (such as SNOMED or UMLS), are also planned.

References

1. Stausberg, J., Koch, D., Ingenerf, J., Betzler, M.: Comparing paper-based with electronic patient records: Lessons learned during a study on diagnosis and procedure codes. Journal of the American Medical Informatics Association 10(5), 470–477 (2003)
2. Holzinger, A., Geierhofer, R., Errath, M.: Semantische Informationsextraktion in medizinischen Informationssystemen. Informatik Spektrum 30(2), 69–78 (2007)
3. Ruch, P., Baud, R., Geissbuhler, A.: Learning-free text categorization. In: Dojat, M., Keravnou, E.T., Barahona, P. (eds.) AIME 2003. LNCS (LNAI), vol. 2780, pp. 199–208. Springer, Heidelberg (2003)
4. Geierhofer, R., Holzinger, A.: Creating an Annotated Set of Medical Reports to Evaluate Information Retrieval Techniques. In: SEMANTICS 2007, Graz, Austria, September 5-7, 2007, pp. 331–339 (2007)
5. Holzinger, A., Geierhofer, R., Errath, M.: Semantic Information in Medical Information Systems - from Data and Information to Knowledge: Facing Information Overload. In: Proceedings of I-MEDIA 2007 and I-SEMANTICS 2007, pp. 323–330 (2007)
6. Matykiewicz, P., Duch, W., Pestian, J.: Nonambiguous concept mapping in medical domain, In: Artificial Intelligence and Soft Computing. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 941–950. Springer, Heidelberg (2006)
7. Schulz, S., Hanser, S., Hahn, U., Rogers, J.: The semantics of procedures and diseases in SNOMED (R) CT. Methods of Information in Medicine 45(4), 354–358 (2006)
8. Senvar, M., Bener, A.: Matchmaking of semantic web services using semantic-distance information. In: Yakhno, T., Neuhold, E.J. (eds.) ADVIS 2006. LNCS, vol. 4243, pp. 177–186. Springer, Heidelberg (2006)