Report on the NTCIR-12 MedNLPDoc Task Results

Dina Vishnyakova
Division of Medical Information Sciences, University Hospitals of Geneva
4, rue Gabrielle-Perret-Gentil
CH-1211, Geneva, Switzerland
+41223790810
dina.vishnyakova@hcuge.ch

Christophe Gaudet-Blavigniac
Division of Medical Information Sciences, University Hospitals of Geneva
4, rue Gabrielle-Perret-Gentil
CH-1211, Geneva, Switzerland
+41223790815
Christophe.gaudet-blavigniac@hcuge.ch

Selen Bozkurt
Department of Biostatistics and Medical Informatics
Dumlupınar Blv. 07058 Akdeniz University, Antalya, Turkey
+905322231883
selenb@gmail.com

David-Zacharie Issom
Division of Medical Information Sciences, University Hospitals of Geneva
4, rue Gabrielle-Perret-Gentil
CH-1211, Geneva, Switzerland
+41223790816
david.issom@hcuge.ch

Renat Vishnyakov
Mathematics and Computer Science department, Institute of Management and Social-Informational Technologies
5, Zipovskaya
350010, Krasnodar, Russia
+79996302050
renat.vishnyakov@mail.ru

Christian Lovis
Division of Medical Information Sciences, University Hospitals of Geneva
4, rue Gabrielle-Perret-Gentil
CH-1211, Geneva, Switzerland
+41223726201
christian.lovis@hcuge.ch

ABSTRACT
The reuse of clinical data for the research environment is becoming one of the important tasks in medical informatics. The automatic assignment of the medical codes to the pre-identified concepts is turning to the Sisyphean task. For the MedNLP task in NTCIR-12 a new approach to automatically enrich the dictionary using online data is proposed. We have developed a text-mining system able to treat medical textual data represented in Japanese language and assign ICD-10 codes with English descriptors to the identified concepts. There are three main parts in the functionality of the system: 1) English version of ICD-10-based dictionary, 2) Wikipedia-based synonyms 3) statistical translation tools such Yandex and Google Translate APIs. This report presents the description of the system and the achieved results on the MedNLPDoc test data. Additionally, we provide an ICD assignation frequency in University Hospitals of Geneva.

General Terms
Algorithms, Standardization, Languages, Theory, Legal Aspects

Keywords
Medical records, electronic medical records (EMR), named entity recognition (NER), shared task and evaluation, ICD-10

Participant ID: iSIMED

Subtask: Phenotyping task

1. INTRODUCTION
Electronic health records (EHRs) are the essential part of the health care system. However still most of medical information is available as free text, which thwart their use in computerized applications [1]. In order to use EHR data or other electronic health data for research purposes, information it contains must be extracted and formatted. Mainly the information, which is used for billing purposes, is well structured and coded. The codes representing billing information [2] usually are assigned and verified by the specialists. But the rest of the data such as health status, history of the patient and etc. remain as free text. To date, there is no perfect solution (with 100% accuracy) for the automatic transformation of the textual concepts to the standardized codes.

It should be also noticed that in order to develop natural language processing (NLP) tools applicable to the medical or clinical domain it is necessary to have an access to data such as admission or discharge letters, radiology reports and procedure reports. Since all these data contain sensitive information regarding patients or care providers, it becomes challenging to get the access to them. Usually it is required to pass all procedures (in order to get approval from the ethical committee), which differs from country to country. Consequently there is no text corpus containing textual medical records available for free access for the NLP researchers. Thus each clinical site develops its own NLP systems and techniques on mapping data from the text to the standardized codes. The biomedical corpora available for the NLP community by Text Retrieval Conference (TREC) [3], BioCreaTive [4] or Conference and Labs of the Evaluation Forum (CLEF) [5], mainly consist of scientific literature extracted from the electronic libraries such as MEDLINE. These corpora doesn't represent the style of medical reports and records [6], mainly due to the style of language which is used for scientific documents and is not applicable to the medical reports. Moreover, it is not rare when clinicians have problems in understanding the jargon of other professional groups [7]. One more challenge when dealing with medical reports is that the style of textual information in these reports depends on the individual physician or care provider. They don't use the descriptors provided by medical thesauri, instead they may use abbreviations or conversational equivalents. All these make it hard to identify medical concepts in the text.

Additionally it arises such challenge as identification of non-local dependencies in the text. Last but not least is the everlasting problem of NLP researchers - dealing with negations [8].

In the framework of the MedNLPDoc task of NTCIR-12, we have developed a text mining system, which accepts medical records in Japanese language as input and returns the assigned identifications (codes) of International Codes for Diseases (ICD-10). The main idea is to assign ICD codes not to the original text but to the version translated in English [13-14]. The descriptors of the ICD-10 codes used in assigning process are also in English. The development of the proposed system consists of three main modules: 1) Dictionary enrichment 2) Documents translation (Japanese to English) 3) Codes assignment.

2. Data and Methods

2.1 Train Set

The committee of NTCIR-12 has provided a set of 200 medical records in XML format. Each record contains meta-data/attributes such as gender, age and medical text. The latter is represented as a text tag with 30 variations of types; see Table 2.1. The survey of the text showed that coded diagnosis was mainly assigned to the text types: **現病歴, 現往歴, 現病歴** and in the untagged text. The distribution of texts per patient ID is shown in Figure 1.

It should be noticed that 161 patient records contained medical notes with no attribute information (e.g. xml-tag "text" had no attributes).

Table 2.1 Some of text types in training set and its translation into English.

Japanese original	English - Google translation	English -Yandex translation
現往歴	Current 往歴	Current traffic history
入院時現症	Admission the current disease	Hospitalization at the time of the disease
現病歴	History of present illness	Current medical history
入院後経過	Elapsed after admission	Hospitalization after
家族歴	Family history	Family history
現在の愁訴	The current complaint	Current chief complaint
家族歴・既往歴	Family history, medical history	Family history・history
月経歴	Menstrual history	Menstrual history
手術所見	Operative findings	Surgical findings
術後経過	Postoperative course	The postoperative
既往歴	medical history	History
手術	Surgery	Surgery
検査所見	Laboratory findings	Inspection findings
現病歴	History of present illness	Current medical history
出生時検査所見	At birth laboratory findings	At birth findings
主訴	Chief complaint	Chief complaints
4歳時	When 4-year-old	4 years of age
2005年1月5日	January 5, 2005	In 2005 1 month, 5 days

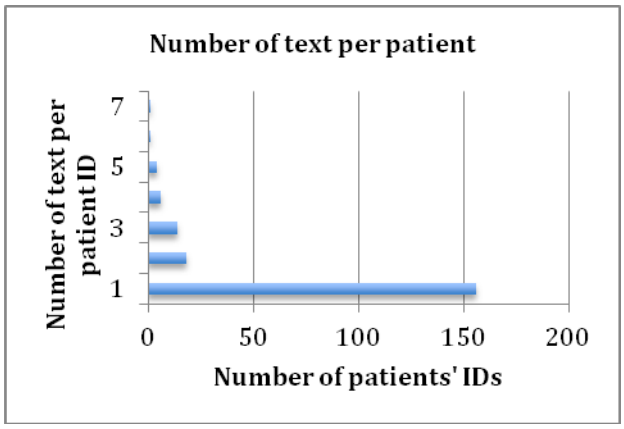

Figure 1 Distribution of text per patient ID in the training set

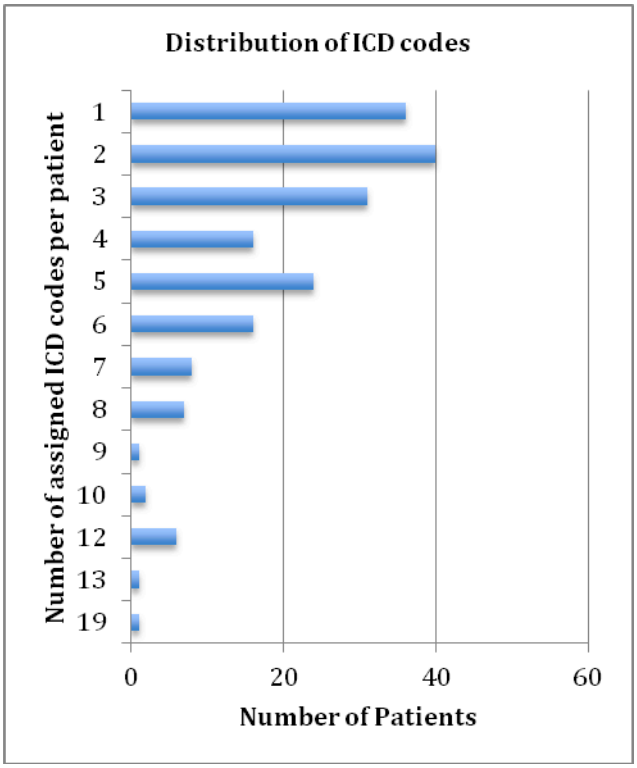

Figure 2 Distribution of assigned codes per patient ID in the training set.

In total there were 772 ICD codes provided along with the patient data. Among these codes, 553 codes are unique and 220 are duplicates. It should be noticed that in the training set the provided ICD codes had neither information on provenance, e.g. nor the tagged text/factoids that are linked to the provided codes.

The number of the provided ICD codes per patient varies from 1 to 19. The distribution of the ICD codes was shown in Figure 2.

2.2 Dictionary

The ICD-10 concepts in English, of the year 2015 version, were used for the assignment task. It consists of 91737 items (codes/descriptors). The first three characters of the code represent the heading of a category of diagnosis. For example W61 has the description "*Contact with birds (domestic) (wild)*" while W6142XA has the description "*Struck by turkey, initial encounter*". Characters 4 to 6 denote more precise diagnosis and the 7th character shows more detailed precisions such as "*initial encounter*" or "*sequela*". The character X, in the example above, is used as a placeholder in order to make the position of the 7th character constant.

2.3 Diagnosis entity recognizer

The description field in ICD-10 dictionary is the diagnosis. Some recurrent terms such as "other" or "unspecified" are met often in the description field. It is useful to note that the description field is often not representative as a term that clinicians use in a clinical record. For example *Streptococcus pneumoniae* is often named *pneumococcus* by clinicians and *Miliaria rubra* is named *Heat rash*. In order to identify ICD-10 CM codes in free text, it is necessary first to link the informal terms to their corresponding ICD-10 CM descriptors/codes. Therefore, an automatic query system is built for the crowdsourcing resource (Wikipedia). The system aims at bridging the descriptors from ICD-10 to real clinical records. It returns the synonyms (if available) and links them to the descriptors of ICD-10 CM codes. The idea is not to use Wikipedia to make a diagnosis but rather to use its summarized simplified knowledge [9-11]. For instance, the article returned on the tuberculosis query contains in the first section *"Tuberculosis, MTB, or TB (short for tubercle bacillus), in the past also called phthisis, phthisis pulmonalis, or consumption, is a widespread, infectious disease caused by various strains of mycobacteria, usually Mycobacterium tuberculosis."* This sentence contains three synonyms following the "*also called*" expression. After the survey of the information returned from Wikipedia we observed that potential synonyms in the Wikipedia.org content appeared in the first 1300 characters of the page[1]. Another observation is that the returned article often contains common classification information such as ICD-9, OMIM code, DiseaseDB or MeSH terms.

Final thesaurus based on ICD-10 consisted of 3 types of descriptors: 1) original descriptors 2) modified descriptors and 3) synonyms. The modified descriptions were created according to the following rules:

- All occurrences of "Other", "Other specified", "unspecified", "Unspecified" were removed;
- The plural form such as "infections" was changed to the singular - "infection";
- For every occurrences of the character "/" that didn't concern vertebra (example: "C3/C4") or physical quantities (example: "mg/100 ml") or karyotype (46 XX/XY) were generated 2 descriptors, one for each side of the slash.
- For every occurrence of brackets "[]" in descriptors, except "[left]" and "[right]" we directly added the term to synonyms and removed it from the description.

The regular expressions were built out of these 3 types of descriptors for the dictionary-based match.

Since the provided training set is in Japanese and the descriptors and synonyms returned by Wikipedia are in English, it became necessary to translate medical texts in English. For this purpose we used both APIs provided by Google and Yandex[2] for automatic translation.

In the translated version of the training set it was observed that medical texts often contains medical abbreviations. Thus, we have extracted the list of medical abbreviations[3] from the Wikipedia and expanded the abbreviations in the text.

While the task of the MedNLPDoc was focused on identifying ICD concepts relevant to the recent/final diagnosis, the text types, which are not referring to the current medical history (current disease, current symptoms), were ignored.

2.4 Results and Discussions

Overall 91806 times Wikipedia.org was queried. The synonyms were retrieved for 3.35% (3073) of ICD10 concepts. Additionally, 1903 MeSH codes, 510 OMIM codes and 1775 DiseaseDB codes were retrieved from the training set. Since heading terms are more generic concepts, statistics were calculated based on this category of terms. For the 1905 queries based on heading terms retrieved 456 articles. Synonyms for 355 heading terms were extracted from these articles.

Evidently, the retrieved IDs from other sources (MeSH and DiseaseDB), at the best-case scenario, cover over 60% of ICD concepts with synonyms. If to consider that it is already less than 3% of all diagnosis concepts, then the usefulness of the mapping retrieved from the Wikipedia is arguable. On the contrary, we have retrieved all ICD10 codes, which were assigned for the period of one year in University Hospitals of Geneva (HUG). This was done during the mapping tool assessment, specially designed for the EHR4CR4 project. It is important to note that HUG uses French version of ICD-10 (CIM-10 in French) terminologies for coding diagnosis. Results showed that over 4399 unique ICD-10 codes were assigned in HUG, 4244 of them were in the dictionary and 898 have synonyms retrieved from Wikipedia. In the 50 most frequent diagnoses (ICD-10 codes) in HUG, 48 of them were in the dictionary, 10 of them have synonyms, retrieved from Wikipedia. Thus, the usage of the Wikipedia as an external resource for mapping concepts is rather rational, since it covers most frequent diagnosis.

[1] [https://en.wikipedia.org/w/api.php?%20format=xml&action=query&titles=[ICD10_descriptor] &prop=revisions&rvprop=content] This query returns a content, which is then reduced to a 1300 character string.

[2] https://tech.yandex.com/translate/

[3] http://www.cancerindex.org/medterm/medtm15.htm#section2

[4] http://www.ehr4cr.eu/ Electronic Health Records for the Clinical Research

Table 2.2 Result of the system obtained with the test set. Here capital letters P, R and F refer to Precision, Recall and F-score accordingly. The lower case letters s, m and p refer to SURE, MAJOR and POSSIBLE.

Run	Ps	Rs	Fs	Pm	Rm	Fm	Pp	Rp	Fp
1	0.02	0.01	0.013	0.033	0.021	0.026	0.045	0.016	0.023
2	0.087	0.087	0.087	0.132	0.087	0.105	0.151	0.057	0.083
3	0.058	0.087	0.07	0.092	0.093	0.092	0.11	0.063	0.08

For the MedNLPDoc task we have submitted three runs:

- Run#1 - the test set was translated with Google translation API.
- Run#2 - the test set was translated with Yandex translation API
- Run#3 - conjunction of ICD codes assigned to text translated by Yandex and Google.

Table 2.3 The comparison of the best run "Best_2" achieved by the system with the runs submitted by other participants of the MedDocNLP task.

Run	Ps	Rs	Fs	Pm	Rm	Fm	Pp	Rp	Fp
Best_2	0.087	0.087	0.087	0.132	0.087	0.105	0.151	0.057	0.083
B	0.209	0.364	0.266	0.361	0.363	0.362	0.42	0.23	0.297
C	0.423	0.295	0.348	0.597	0.239	0.341	0.681	0.145	0.239
D	0.237	0.223	0.23	0.313	0.168	0.219	0.374	0.109	0.169
E	0.316	0.353	0.334	0.524	0.338	0.411	0.6	0.217	0.319
F_1	0.018	0.064	0.028	0.032	0.072	0.044	0.044	0.05	0.047
F_2	0.065	0.042	0.051	0.096	0.044	0.06	0.166	0.038	0.062
F_3	0.086	0.04	0.054	0.12	0.039	0.058	0.199	0.032	0.055
G_1	0.265	0.253	0.259	0.391	0.229	0.289	0.483	0.149	0.228
G_2	0.267	0.256	0.261	0.393	0.232	0.291	0.484	0.15	0.229
G_3	0.223	0.47	0.303	0.402	0.487	0.44	0.48	0.318	0.382
H	0.173	0.388	0.235	0.314	0.408	0.354	0.37	0.265	0.309

The results achieved by the system are represented in Table 2.2. The organizers of the MedDocNLP contest provided the explanation of the metrics used for the evaluation such as SURE, MAJOR and POSSIBLE in their report. It is obvious that the evaluation results for the codes assignment on the text translated by Yandex API is at least 4 times better. These results completely depend on the quality of the translation. The combination of run#1 and run#2 did not improve results for the SURE category, but it improved Recall for the Major and POSSIBLE categories.

Table 2.3 shows the results achieved by different participants of the task. This table shows that all teams achieved F scores lower than 0.5. It can be explained by the prevalence of the Precision scores over Recall ones.

3. Conclusion

We have shown the analysis of the medical records set and ICD-10 thesaurus. This analysis reveals that the terms used for the diagnosis are not the same as for the description of diseases and diagnosis used by treating physicians. It also explains the results achieved by our team. One of the main challenges was to deal with the text in Japanese. We should notice that even if the texts were in English, it would be still difficult to achieve good evaluation results. It is not rare when the diagnosis concepts are represented implicitly in the text. Thus to use only dictionary-based methods with some ad-hoc techniques is not optimal when it comes to ICD concepts identification.

In the field of the biomedical natural language processing (bioNLP) exist many tools such as named-entity recognizers for diseases, drugs or genes. However, detecting disease in text is not a trivial task. Since diagnosis is not a disease, the characteristics of diagnosis are less clear and encompass multiple concepts broader than only diseases. As the ICD codes are used mainly for the billing purposes the construction of concepts and its logic is not aligned with the physicians who is responsible for the diagnoses. The language of physicians is live. It is not rare when one can use informal terms or abbreviations. Additionally, the ICD codes extracted from the HUG showed that in real-world hospital barely 5% of overall ICD-10 concepts are used. It also showed that there are some differences in ICD codes depending on the language. For instance we have identified some codes, which exist only in the French version. Consequently, it is possible to assume that it is the same situation with the Japanese version of ICD-10 codes.

The use of external resources such as Wikipedia.org in automatic diagnosis assignment showed that it could be useful for both: 1) enriching the dictionary and 2) mapping the ICD-10 codes to the existing resources such as MeSH, DiseasesDB and etc.

Last but not least, it is very challenging to develop effective encoding tools without access to the real data. Besides, the statistical methods for assigning the codes require the annotated corpus for training. Thus the provided medical records are neither representative as the real medical texts nor optimal for the training purposes. Mainly it is due to the size of training set and the lack of ICD codes provenance, but still they provide an idea of what medical coding is.

4. REFERENCES

[1] Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." Nature Reviews Genetics 13.6 (2012): 395-405.

[2] Green, D. "New ICD-10 coding: documentation to provide better care, support more accurate billing." The American nurse 46.1 (2013): 4-4.

[3] Voorhees, Ellen M., and William R. Hersh. "Overview of the TREC 2012 Medical Records Track." TREC. 2012.

[4] Arighi, Cecilia N., et al. "Natural language processing pipelines to annotate BioC collections with an application to the NCBI disease corpus." *Database*2014 (2014).

[5] Goeuriot, Lorraine, et al. "Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval." *CLEF 2014 Online Working Notes* 1180 (2014): 43-61.

[6] Richesson, Rachel L., Kin Wah Fung, and Olivier Bodenreider. "Coverage of Rare Disease Names in Clinical Coding Systems and Ontologies and Implications for Electronic Health Records-Based Research." ICBO. 2014.

[7] Howard, Tera, Kara L. Jacobson, and Sunil Kripalani. "Doctor talk: physicians' use of clear verbal communication." *Journal of health communication* 18.8 (2013): 991-1001.

[8] Furst, Elizabeth, and Jennifer Lynn Swindell. "AUTOMATIC MEDICAL CODING SYSTEM AND METHOD." U.S. Patent No. 20,150,379,241. 31 Dec. 2015.

[9] Khare, Ritu, et al. "Crowdsourcing in biomedicine: challenges and opportunities." *Briefings in bioinformatics* (2015): bbv021.

[10] Khare, Ritu, et al. "Scaling drug indication curation through crowdsourcing."*Database* 2015 (2015): bav016.

[11] Burger, John D., et al. "Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing." *Database* 2014 (2014): bau094.

[12] Hailu, Negacy D., K. Bretonnel Cohen, and Lawrence E. Hunter. "Ontology Translation: A Case Study on Translating the Gene Ontology from English to German." *Natural Language Processing and Information Systems*. Springer International Publishing, 2014. 33-38.

[13] Krasnova, T. I., and I. S. Vanushin. "Machine translation error analysis." *Young*9 (2015): 89.

[14] Hailu, Negacy D., K. Bretonnel Cohen, and Lawrence E. Hunter. "Ontology Translation: A Case Study on Translating the Gene Ontology from English to German." *Natural Language Processing and Information Systems*. Springer International Publishing, 2014. 33-38.