

# CUNI at the ShARe/CLEF eHealth Evaluation Lab 2014

Shadi Saleh and Pavel Pecina

Institute of Formal and Applied Linguistics  
Charles University in Prague  
Czech Republic  
{saleh,pecina}@ufal.mff.cuni.cz

**Abstract.** This report describes the participation of the team of Charles University in Prague at the ShARe/CLEF eHealth Evaluation Lab in 2014. We took part in Task 3 (User-Centered Health Information Retrieval) and its both subtasks (monolingual and multilingual retrieval). Our system was based on the Terrier platform and its implementation of the Hiemstra retrieval model. We experimented with several methods for data cleaning and automatic spelling correction of query terms. For data cleaning, the most effective method was to employ simple HTML markup removal. The more advanced cleaning methods which remove boilerplate decreased retrieval performance. Automatic correction of spelling errors performed on the English queries for the monolingual task proved to be efficient and led to our best P@10 score equal to 0.5360. In the multilingual retrieval task, we employed the Khresmoi medical translation system developed at the Charles University in Prague and translated the source queries from Czech, German, and French to English and employed the same retrieval system as for the monolingual task. The cross-lingual retrieval performance measured by P@10 relative to the scores obtained in the monolingual task ranged between 80% and 90% depending on the source language of the queries.

**Keywords:** multilingual information retrieval, data cleaning, machine translation, spelling correction

## 1 Introduction

The digital medical content available on-line has grown rapidly in recent years. This increase has a potential to improve user experience with Web medical information retrieval (IR) systems which are more and more often used to consult users' health related issues. Recently, Fox [2] reported that about 80% of Internet users in the U.S. look for health information on-line and this number is expected to grow in future.

In this report, we describe our participation at the the ShARe/CLEF eHealth Evaluation Lab 2014, Task 3 [3] which focus on developing methods and data resources for evaluation of IR from the perspective of patients.

Our system is built on Terrier [7] and employs its implementation of the Hiemstra retrieval model. The main contribution of our participation is the examination of several methods for cleaning the document collection (provided as raw documents with HTML markup) and automatic correction of spelling errors in query terms and handling unknown words.

In the remainder of the paper, we review the task specification, describe the test collection and queries, our experiments, their results and conclude with the main findings of this work.

## 2 Task description

The goal of Task 3 in ShARe/CLEF eHealth Evaluation Lab 2014 is to design an IR system which returns a ranked list of medical documents (English web pages) from the provided test collection as a response to patients' queries. The task is split into two tasks:

- **Task 3a** is a standard TREC-style text IR task<sup>1</sup>. The participants had to develop monolingual retrieval techniques for a set of English queries and return the top 1,000 relevant documents from the collection for each query. They could submit up to seven ranked runs: Run 1 as a baseline, Runs 2–4 for experiments exploiting discharge summaries provided for each query, and Runs 5–7 for experiments not using the discharge summaries (see Section 3.2).
- **Task 3b** extends Task 3a by providing the queries in Czech, German, and French. The participants were asked to use these queries to retrieve relevant documents from the same collection as in Task 3a. They were allowed to submit up to seven ranked runs for each language using same restrictions as in Task 3a. No restrictions were put on techniques for translating the queries to English.

## 3 Data

### 3.1 Document Collection

The document collection for Task 3 consists of automatically crawled pages from various medical web sites, including pages certified by the Health On the Net<sup>2</sup> and other well-known medical web sites and databases. The collection was provided by the Khresmoi project<sup>3</sup> and covers a broad set of medical topics. For details, see [5].

The collection contains a total of 1,104,298 web pages. We excluded a small portion of the pages because of no or un-readable content (382 pages contained a Flash-related error message, and 658 pages were unreadable binary files).

---

<sup>1</sup> <http://trec.nist.gov/>

<sup>2</sup> <http://www.hon.ch/>

<sup>3</sup> <http://khresmoi.eu/>

---

```

<topic>
  <id>qtest2014.47</id>
  <discharge_summary>
    22821-026994-DISCHARGE.SUMMARY.txt
  </discharge_summary>
  <title>
    tretament for subarachnoid hemorrhage
  </title>
  <desc>
    What are the treatments for subarachnoid hemorrhage?
  </desc>
  <narr>
    Relevant documents should contain information on the treatment
    for subarachnoid hemorrhage.
  </narr>
  <profile>
    This 36 year old male patient does not remember how he was treated
    in the hospital. Now he wants to know about the care for
    subarachnoid hemorrhage patients.
  </profile>
</topic>

```

---

**Fig. 1.** An example of a test query. Note the spelling error in the title.

### 3.2 Queries

Two sets of queries were provided for Task 3 [3]. The training set of 5 queries and their matching relevance assessments and a test set of 50 queries for the main evaluation. All queries were provided in English (for Task 3a) and in Czech, German, and French (for Task 3b).

The English queries were constructed by medical professionals from the main disorder diagnosed in discharge summaries of real patients (i.e. documents containing a summary of important information from their entire hospitalization) provided for Task 2. Then, the queries were translated to Czech, German, and French by medical professionals and reviewed. Each query description consists of the following fields:

- **title:** text of the query,
- **description:** longer description of what the query means,
- **narrative:** expected content of the relevant documents,
- **profile:** main information on the patient (age, gender, condition),
- **discharge\_summary:** ID of the matching discharge summary.

An example of a query is given in Figure 1 and some basic statistics associated with the query sets are shown in Table 1.

**Table 1.** Statistics of the query sets: number of queries, average number of tokens in the titles, descriptions, narratives, and profiles, and total number of relevant documents.

query set	size	title	description	narrative	profile	relevant docs
training	5	3.6	10.6	12.8	52.4	134
test	50	4.3	8.9	12.4	33.4	3,209

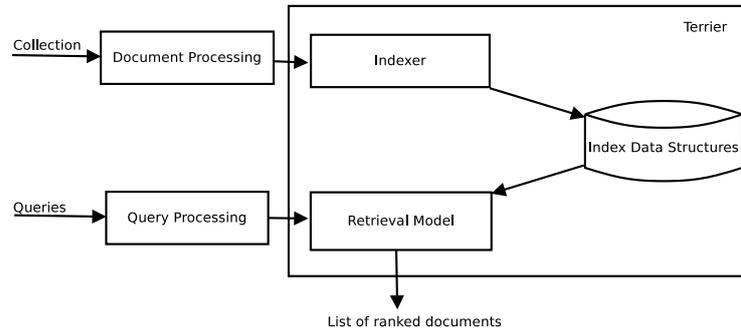


Fig. 2. System architecture overview.

## 4 System description

Our system consists of three main components: a document processing component, a query processing component, and a search engine (see Figure 2). First, the collection is processed and data to be indexed is extracted from each document in the collection. Second, the search engine is employed to index the data. Third, each query is processed (eventually translated), enters the search engine which retrieves the top 1,000 ranked documents based on a retrieval model and its parameters.

The main evaluation metrics for Task 3 is precision at top 10 ranked documents (P@10), however, we also present results of other well known metrics implemented in the standard *trec\_eval* tool:<sup>4</sup> such as precision at top 5 ranked documents (P@5), Normalized Discount Cumulative Gain at top 5 and 10 ranked documents (NDCG@5, NDCG@10), Mean Average Precision (MAP), precision after R documents have been retrieved where R is the number of known relevant documents (Rprec), binary preference (bpref), and the number of relevant documents (rel.ret). In the remainder of this section, we describe our retrieval system in more detail.

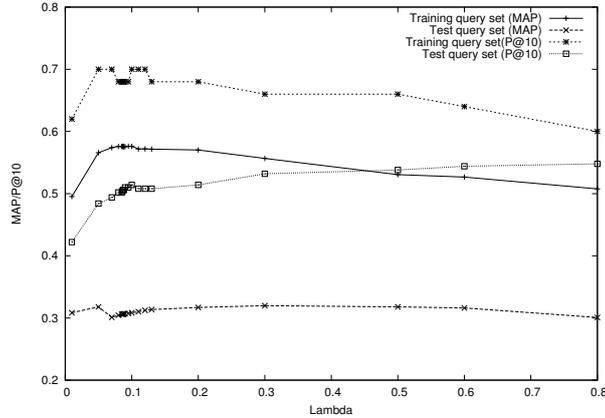
### 4.1 Retrieval model

We employ Terrier 3.5 [7] as the search engine for indexing and retrieval. The retrieval model is the standard Hiemstra language model [4] as implemented in Terrier, where given a query  $Q$  and its terms  $Q = (t_1, t_2, \dots, t_n)$ , each document  $D$  in a collection  $C$  is scored using the following formula:

$$P(D, Q) = P(D) \cdot \prod_{i=1}^n ((1 - \lambda_i) P(t_i|C) + \lambda_i P(t_i|D)),$$

where  $P(D)$  is the prior probability of  $D$  to be relevant estimated as by summing up frequencies of query terms in the document  $D$  over their frequencies in the

<sup>4</sup> [http://trec.nist.gov/trec\\_eval/](http://trec.nist.gov/trec_eval/)



**Fig. 3.** Tuning the lambda parameter of the Hiemstra model on the training data (by MAP) and the scores obtained on the test data. The plot also contains the results for P@10.

collection  $C$ .  $P(t_i|C)$  and  $P(t_i|D)$  are probabilities of  $t_i$  in the collection  $C$  and document  $D$ , respectively. They are estimated by maximum likelihood estimation using frequencies of the term  $t_i$  in the collection  $C$  and document  $D$ , respectively.  $\lambda_i$  is a linear interpolation coefficient reflecting the overall importance of the term  $t_i$ . In our system, we used the same value  $\lambda$  for all the terms and tune it on the training query set by grid search to maximize MAP (see Figure 3). The highest MAP value was achieved with  $\lambda=0.087$  which is used in all our experiments. After releasing the relevance assessments of the test queries, we measured the effect of  $\lambda$  on the test set performance and the results are shown in Figure 3 too. The figure also contains test and training curves for P@10, the official measures for Task 3 in 2014, which was announced together with the evaluation results.

We also perform Pseudo Relevance Feedback (PRF) implemented in Terrier as *query expansion* which modifies (expands) a given query by adding the most informative terms from top retrieved documents and performs the retrieval again with the expanded query. We use Terrier's implementation for Bo1 (Bose-Einstein 1) from the Divergence From Randomness framework [1]. We expanded each query by taking ten highest scored terms from three top ranked documents. These values achieved the best results measured on the training set, although they were lower than the results without PRF.

## 4.2 Document processing

The documents in the collection are provided as raw web pages including all the HTML markup and eventually also CSS style definitions and Javascript code which should be removed before indexing. We employed three data cleaning methods and evaluate their effect on the retrieval quality measured on the training queries.

**Table 2.** Collection size (in MB and millions of tokens) and average document length (in tokens) after applying the different cleaning methods.

method	total size (MB)	%	total length (mil. tokens)	%	avg length (tokens)
<i>none</i>	41,628	100.00	–	–	–
HTML-Strip	6,821	16.38	1,006	100.00	911
Boilerpipe	3,248	7.80	423	42.11	383
JusText	2,853	6.85	452	44.93	409

**Table 3.** The effect of the different data cleaning methods on the retrieval results using the training queries, and the results before cleaning.

run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	Rprec	bpref	rel_ret
<i>none</i>	0.5200	0.6200	0.4758	0.5643	0.5019	<b>0.5006</b>	0.6542	127
HTML-Strip	<b>0.7200</b>	<b>0.6800</b>	<b>0.7030</b>	<b>0.6844</b>	<b>0.5757</b>	0.4833	<b>0.6597</b>	<b>130</b>
Boilerpipe	0.6400	0.5200	0.6767	0.5793	0.4601	0.4665	0.5800	121
JusText	0.5600	0.5000	0.5785	0.5315	0.3061	0.2983	0.5087	99

First, we simply removed all markup, style definitions, and script code by the a Perl module HTML-Strip<sup>5</sup> (but keep meta keywords and meta description tags). This reduces the total size of the collection from 41,628 MB to 6,821 MB, which is about 16% of the original size and average document length is 911 tokens (words and punctuation marks).

Although the size reduction is very substantial, the resulting documents still contained a lot of noise (such as web page menus, navigation bars, various headers and footers), which is likely not to be relevant to the main content of the page. Such noise is often called boilerplate. We used two methods to remove it: Boilerpipe [6] reduced the total number of tokens in the collection by additional 58% (the average document length is 383 tokens) and JusText [9] by 55% (the average document length is 409 tokens).

More details from the data cleaning phase are provided in Table 2. Table 3 then reports the IR results obtained by the Hiemstra model using the training set and the collection processed by the three methods compared with the case where no cleaning was performed at all. Surprisingly, the most effective method is the simple HTML-Strip tool. The two other methods are probably too aggressive and remove some relevant material important for IR. In all the following experiments, the collection is cleaned by HTML-Strip.

### 4.3 Query processing

For Task 3a, the queries entering the search engine of our system are constructed from the title and narrative description fields. For Task 3b, we translated the queries to English by the Khresmoi translator described in [8] which is tuned

<sup>5</sup> <http://search.cpan.org/dist/HTML-Strip/Strip.pm>

**Table 4.** Task 3a: Monolingual IR performance on test queries.

run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	Rprec	bpref	rel_ret
RUN1	0.5240	0.5060	0.5353	0.5189	0.3064	0.3538	0.4586	<b>2562</b>
RUN5	<b>0.5320</b>	<b>0.5360</b>	<b>0.5449</b>	<b>0.5408</b>	<b>0.3134</b>	<b>0.3599</b>	<b>0.4697</b>	2556
RUN6	0.5080	0.5320	0.5310	0.5395	0.2100	0.2317	0.3984	1832
RUN7	0.5120	0.4660	0.5333	0.4878	0.1845	0.2141	0.3934	1676

specifically to translate user queries from the medical domain [10]. For comparison, we also provide results obtained by translating the queries using on-line translators Google Translate<sup>6</sup> and Bing Translator<sup>7</sup>. In the baseline experiment, we take the title terms as they are. As an additional query processing step, we attempt to handle words which are unknown.

There are three types for unknown words in the queries. The first (and frequent) type is made of words with spelling errors. Such errors could be automatically and corrected. The second type of unknown words is made of words, which are correct in the source language, but they are out-of-vocabulary (OOV) for the translation system and thus remain untranslated. Such words could be modified/replaced by known words (e.g. morphological variant) before translation or translated ex-post by a dictionary look-up or another translation system. The third type is made of query terms which are correct (and correctly translated) but they do not appear in the test collection and are not indexed. In such a case, there is no straightforward and easy solution how to deal with them (possibly they could be replaced by a synonym or another related words).

**Spelling correction** The queries for Task 3 were written by medical professionals, but this does not guarantee that they do not contain spelling errors, see e.g. the query in Figure 1, where the word *tretament* contains a spelling error. To detect and correct the spelling errors we employed an on-line English medical dictionary MedlinePlus<sup>8</sup>. This dictionary provides a definition for correct medical words and for those which are not correct, it offers a possible correction.

We automated the process and for each term in the title and narrative of the English queries, we check whether the word exists or not. If the response is "404 not found", we parse the page to get the closest word.

**Two steps translation** After translation, our spell checking script reported some unknown words which left untranslated by the Khresmoi system. We passed all such words to Google Translate and obtained their translation which replaced the untranslated forms.

<sup>6</sup> <http://translate.google.com/>

<sup>7</sup> <http://www.bing.com/translator/>

<sup>8</sup> <http://www.nlm.nih.gov/medlineplus/>

**Table 5.** Task 3b: Cross-lingual IR performance using the test queries in Czech.

run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	Rprec	bpref	rel_ret
RUN1	0.4400	0.4340	0.4361	0.4335	0.2151	0.2696	0.3863	1965
RUN5	<b>0.4920</b>	<b>0.4880</b>	0.4830	<b>0.4810</b>	<b>0.2399</b>	<b>0.2950</b>	<b>0.4245</b>	<b>2112</b>
RUN6	0.4680	0.4560	<b>0.4928</b>	0.4746	0.1573	0.1984	0.3458	1591
RUN7	0.3360	0.3020	0.3534	0.3213	0.1095	0.1482	0.2982	1186
RUN1 <sub>G</sub>	<b>0.5347</b>	<b>0.5061</b>	<b>0.5200</b>	<b>0.5065</b>	<b>0.2814</b>	<b>0.3230</b>	<b>0.4504</b>	<b>2324</b>
RUN1 <sub>B</sub>	0.4980	0.5020	0.4877	0.4948	0.2603	0.3138	0.4463	2293

**Table 6.** Task 3b: Cross-lingual IR performance using the test queries in German.

run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	Rprec	bpref	rel_ret
RUN1	0.3760	0.3920	0.3561	0.3681	0.1834	0.2283	0.3359	1806
RUN5	<b>0.4160</b>	<b>0.4280</b>	0.3963	<b>0.4058</b>	<b>0.2014</b>	<b>0.2463</b>	<b>0.3629</b>	<b>1935</b>
RUN6	0.3880	0.3820	<b>0.4125</b>	0.4024	0.1348	0.1671	0.3054	1517
RUN7	0.3520	0.3200	0.3590	0.3330	0.1308	0.1593	0.3433	1556
RUN1 <sub>G</sub>	<b>0.4583</b>	<b>0.4583</b>	<b>0.4491</b>	<b>0.4521</b>	<b>0.2559</b>	<b>0.3092</b>	<b>0.4445</b>	<b>2298</b>
RUN1 <sub>B</sub>	0.4375	0.4229	0.4323	0.4238	0.2317	0.2814	0.4169	2133

**Table 7.** Task 3b: Cross-lingual IR performance using the test queries in French.

run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	Rprec	bpref	rel_ret
RUN1	0.4640	0.4720	0.4611	0.4675	0.2344	0.2887	0.4153	2056
RUN5	<b>0.4840</b>	<b>0.4840</b>	0.4766	<b>0.4776</b>	<b>0.2398</b>	<b>0.2939</b>	<b>0.4191</b>	<b>2064</b>
RUN6	0.4600	0.4560	<b>0.4772</b>	0.4699	0.1703	0.2055	0.3447	1531
RUN7	0.3520	0.3240	0.3759	0.3520	0.1300	0.1716	0.3209	1313
RUN1 <sub>G</sub>	0.5160	0.5200	0.5001	0.5088	0.2780	0.3286	0.4565	2421
RUN1 <sub>B</sub>	<b>0.5469</b>	<b>0.5265</b>	<b>0.5331</b>	<b>0.5242</b>	<b>0.2904</b>	<b>0.3313</b>	<b>0.4751</b>	<b>2449</b>

## 5 Experiments and results

### 5.1 Task 3a: monolingual IR

We submitted 4 runs (RUN1, RUN5, RUN6 and RUN7). We did not submit RUN2–4 because we did not use the discharge summaries in our experiments. In all the submitted runs, we apply the Hiemstra retrieval mode with the tuned parameter value on the test collection processed by the HTML-Strip. The submitted runs employ the techniques discussed in the previous section as follows:

**RUN1** exploits queries constructed from the titles only without any processing.

**RUN5** extends RUN1 by conducting spelling correction on query titles.

**RUN6** extends RUN5 by applying PRF (query expansion).

**RUN7** extends RUN6 by adding queries from both of titles and narrative tags.

The results of our runs submitted to Task 3a are summarized in Table 4. The only improvement was achieved by RUN5 implementing spelling correction of English. We found 11 misspelled words in English test queries which affected 7 queries in total. Neither query expansion using PRF in RUN5 nor adding additional query terms from the narrative fields bring any improvement.

**Table 8.** Comparison of the P@10 scores achieved in the cross-lingual runs with the scores from the corresponding monolingual runs.

run ID	English	Czech		German		French	
	P@10	P@10	%	P@10	%	P@10	%
RUN1	0.5060	0.4340	85.77	0.3920	77.47	0.4720	93.28
RUN5	<b>0.5360</b>	<b>0.4880</b>	<b>91.04</b>	<b>0.4280</b>	<b>79.85</b>	<b>0.4840</b>	<b>90.29</b>
RUN6	0.5320	0.4560	85.71	0.3820	71.80	0.4560	85.71
RUN7	0.4660	0.3020	64.80	0.3200	68.66	0.3240	69.52

## 5.2 Task 3b: multilingual IR

In the second part of Task 3, we apply the previous runs on translated queries using the same setup. But we handle OOV problem in RUN5 not by spelling correction as we do in task 3a. We found 7 untranslated words from Czech test queries, 5 words French, and 12 from German, which were post-translated.

The best P@10 is achieved by Czech IR using RUN5 as shown in Table 5. Solving the OOV issue in Czech queries enhances the results by 5.4%, 1.2% in French IR and 3.6% in German IR, while PRF in all multilingual runs does not help. It might have happened because of complex morphological forms for medical terms. Also the usage of narrative tags does not improve the results.

Unofficially, we also show results for queries have been translated using Google Translate (See RUN1<sub>G</sub>) and Bing Translator (See RUN1<sub>B</sub>). Google Translate does better than Bing Translator on Czech and German, while Bing Translator performs better than Google Translate when the source language is French. However, both these services outperform the Khresmoi translator on this test set.

Table 8 compares our results in Task 3a and Task 3b. The best relative scores are obtained by translation from Czech, which gives 91.04% of the best monolingual results (RUN5). For translation from French, the relative performance is similar (90.29%) but for German, it is only 79.85%. Here the problem is probably in German compound words, which are difficult to translate.

## 5.3 Conclusion and future work

We have described our participation in ShARe/CLEF 2014 eHealth Evaluation Lab Task 3 in its two subtask. Our system was based on the Terrier platform and its implementation of the Hiemstra retrieval model. We experimented with several methods for data cleaning in the test collection and domain-specific language processing (e.g., correction of spelling errors) and found that the optimal cleaning method is a simple removal of HTML markup. In future, we would like to examine query expansion techniques based on the UMLS [11] thesaurus and extend our work on spelling correction to languages other than English.

## Acknowledgments

This work was funded by the Czech Science Foundation (grant n. P103/12/G084) and the EU FP7 project Khresmoi (contract no. 257528).

## References

1. Amati, G.: Probability models for information retrieval based on divergence from randomness. Ph.D. thesis, University of Glasgow (2003)
2. Fox, S.: Health Topics: 80% of internet users look for health information online. Tech. rep., Pew Research Center (2011)
3. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., Mueller, H.: Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In: Proceedings of CLEF 2014 (2014)
4. Hiemstra, D., Kraaij, W.: Twenty-one at TREC-7: ad-hoc and cross-language track. In: Proceedings of the seventh Text Retrieval Conference TREC-7. pp. 227–238. US National Institute of Standards and Technology (1999)
5. Kelly, L., Goeuriot, L., Suominen, H., Schrek, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the share/clef ehealth evaluation lab 2014. In: Proceedings of CLEF 2014. Lecture Notes in Computer Science (LNCS), Springer (2014)
6. Kohlschütter, C., Fankhauser, P., Nejdil, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. pp. 441–450. ACM, New York, NY, USA (2010)
7. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006) (2006)
8. Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlaváčová, J., Jones, G., Kelly, L., Leveling, J., Mareček, D., Novák, M., Popel, M., Rosa, R., Tamchyna, A., Urešová, Z.: Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial Intelligence in Medicine* (2014)
9. Pomikalek, J.: Removing Boilerplate and Duplicate Content from Web Corpora. Ph.D. thesis, Ph.D. thesis, Masaryk University (2001)
10. Urešová, Z., Hajič, J., Pecina, P., Dušek, O.: Multilingual test sets for machine translation of search queries for cross-lingual information retrieval in the medical domain. In: Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association, Reykjavik, Iceland (2014)
11. U.S. National Library of Medicine: UMLS reference manual (2009), metathesaurus. Bethesda, MD, USA