BioRo: The Biomedical Corpus for the Romanian Language

Maria Mitrofan and Dan Tufiş
Research Institute for AI ”Mihai Drăgănescu”, Romanian Academy
”Calea 13 Septembrie”, Bucharest 050711, Romania
{maria, tufis}@racai.ro

Abstract

The biomedical domain provides a large amount of linguistic resources usable for biomedical text mining. While most of the resources used in biomedical Natural Language Processing are available for English, for other languages including Romanian the access to language resources is not straight-forward. In this paper, we present the biomedical corpus of the Romanian language, which is a valuable linguistic asset for biomedical text mining. This corpus was collected in the contexts of CoRoLa project, the reference corpus for the contemporary Romanian language. We also provide informative statistics about the corpus, a description of the data-composition. The annotation process of the corpus is also presented. Furthermore, we present the fraction of the corpus which will be made publicly available to the community without copyright restrictions.

Keywords: BioNLP; Romanian; Corpora; Specialized Domain;

1. Introduction

Natural language processing (NLP) technologies have been used to extract useful information from different types of biomedical texts, such as: biomedical literature, research papers, medical school lecture notes, clinical notes, discharge summaries, clinical practice guidelines, etc.

In order to be able to obtain relevant knowledge form textual data, high-quality resources (i.e. corpora, lexica, terminologies, tesauri etc.) are needed. At international level various biomedical textual resources have been developed, but most of them are available in English (e.g.GENIA corpus (Kim et al., 2003), AnEM corpus (Ohta et al., 2012), CellFinder corpus (Neves et al., 2012), etc.).

In many countries where the official language is not English, there is a technical barrier in using Natural Language Processing in the biomedical domain (bioNLP) due to the fact that textual resources are more scarce. Nevertheless, a significant progress has been made over the past decades thanks to the contribution of various active NLP communities.

For example, for French the language Zweigenbaum et al. (2005) created the ”Unified Medical Lexicon for French” (UMLF), a reference resource for bioNLP. Another important linguistic resource is Corpus Medical du Centre de Recherche en Terminologie et Traduction (CMCRTT), which is a monolingual French corpus for bioNLP and it is publicly available both as plain or POS tagged text (Neveol et al., 2014).

For the Bulgarian language important efforts have been made in developing resources usable for various NLP tasks. Boytcheva et al. (2009) collected a biomedical corpus containing 6400 words, 2000 of them are part of the Bulgarian medical terminology.

Swedish is another language which has important resources for bioNLP. In 2012, Velupillai (2012) created an annotated gold standard of medical records, used for terminology management and linguistic explorations. Also, a negation and clinical uncertainty taxonomy schema was proposed for Swedish language and was mapped to an English annotation schema (Mowery et al., 2012).

The Romanian language is an under-resourced language, regarding resources available for bioNLP. Therefore, considerable efforts are carried on in order to improve the availability of the Romanian biomedical resources usable for bioNLP. At this moment the most important project is the CoRoLa project (Mititelu et al., 2018), started in 2012 by the Romanian Academy Research Institute for Artificial Intelligence ”Mihai Drăgănescu” (RACAI) and the Institute for Computer Science in Iaşi. This is an on-going project that aims to create a reference corpus of contemporary Romanian (from 1945 onwards). In the context of the COROLA project we created an important biomedical corpus for the Romanian language (BioRo) that can be used for different bioNLP tasks.

Developing a corpus for a specialized domain, in this case for the biomedical domain, is not an easy task and there are several steps that need to be taken: defining the structure and the linguistic coverage of the corpus, collecting the texts following the established structure, addressing copyright problems, processing the corpus with the NLP technologies available (segmentation, tokenization, lemmatization, tagging etc.) and establishing the availability of the data.

In what follows, the structure of the corpus is presented, describing the process of acquiring it and the pre-processing steps (Section 2.). Section 3. describes the statistics of the corpus. Section 4. presents the processing tool used for annotation and the results obtained after the annotation step. The availability of the data is presented in section 5.. In section 6. we draw the conclusions and present the future work.

2. BioRo Corpus Structure

BioRo corpus contains, excluding the punctuation, 9,864,707 tokens distributed in different medical subdomains such as: diabetes, endocrinology, cardiology, oncology, neurology etc. (Table 1). All the texts are tokenized, lemmatized and morpho-syntactically tagged.

2.1. Collecting the Data

The process of collecting the texts was a difficult task, because the laws of intellectual property are very restrictive and also because, in general, most of the biomedical literature is not published in the Romanian language. The main providers of medical texts contained in BioRo corpus are: the Romanian Academy Publishing House, Polirom publishing house, PIM publishing house, Timpul publishing house, the Romanian Medical Journal, medical blogs and medical school lecture notes.

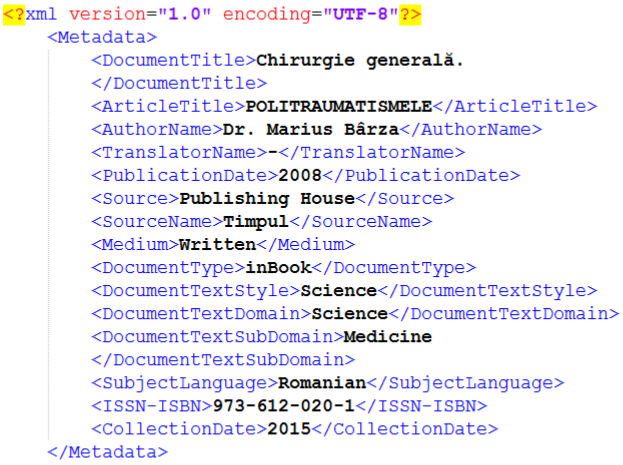
2.2. Cleaning the Data

Initially, the textual resources contained in BioRo corpus were available in various formats such as unprotected .pdf and .doc. All the texts have been converted into a raw text format which fits to our processing tools (Tufis et al., 2008). The conversion of the medical files included a boilerplate removal phase in which all the figures, tables, headers, footers, etc. were removed. The non-standard codes for diacritics have been replaced with the proper ones while the missing diacritics have been inserted automatically.

Moreover in order to prepare the corpus for the processing step we decided to correct various types of misspellings such as: missing letters ”diabe” instead of ”diabet” (en. ”diabetes”), extra spaces ”cardio logie” instead of ”cardiologie” (en. ”cardiology”), etc. Nevertheless, after this cleaning phase (given that automatic procedures are not error-free), roughly 3-4% of the words still need to be corrected.

2.3. Metadata Creation

All the files contained in the corpus have an associated XML file representing the corresponding metadata scheme. Each XML file contains specific information at the document level about source, author, genre and type of the text, etc. Most of the information contained in the metadata scheme is relevant for the indexing of the corpus. All 823 files have manually created metadata descriptors, each being created following the metadata scheme used in CoRoLa (Figure 1).



```
<?xml version="1.0" encoding="UTF-8"?>
    <Metadata>
        <DocumentTitle>Chirurgie generală.
        </DocumentTitle>
        <ArticleTitle>POLITRAUMATISMELE</ArticleTitle>
        <AuthorName>Dr. Marius Bârza</AuthorName>
        <TranslatorName>-</TranslatorName>
        <PublicationDate>2008</PublicationDate>
        <Source>Publishing House</Source>
        <SourceName>Timpul</SourceName>
        <Medium>Written</Medium>
        <DocumentType>inBook</DocumentType>
        <DocumentTextStyle>Science</DocumentTextStyle>
        <DocumentTextDomain>Science</DocumentTextDomain>
        <DocumentTextSubDomain>Medicine
        </DocumentTextSubDomain>
        <SubjectLanguage>Romanian</SubjectLanguage>
        <ISSN-ISBN>973-612-020-1</ISSN-ISBN>
        <CollectionDate>2015</CollectionDate>
    </Metadata>
```

Figure 1: A example of metadata scheme.

3. BioRo Corpus Statistics

In order to present general statistics about the biomedical corpus, we counted all the sentences, unique lemmas, tokens and content words (nouns, main verbs, adjective and adverbs) (Table 1). The punctuation is obtained by subtracting the words count from the tokens count. Figure 2 indicates the percentage of tokens for each biomedical subdomain contained in the corpus. In Figure 2 it is also shown that the biomedical corpus has an unbalanced distribution regarding the number of tokens for each biomedical subdomain, but this comes from the fact that in the process of obtaining the data there are many difficulties, due to the copyright restrictions.

# Tokens	9,864,707
# Unique lemmas	237,620
# Punctuation	1,498,218
# Sentences	561,978
Tokens per sentence	17.55
Punctuation per sentence	2.66

Table 1: General statistics over the biomedical corpus

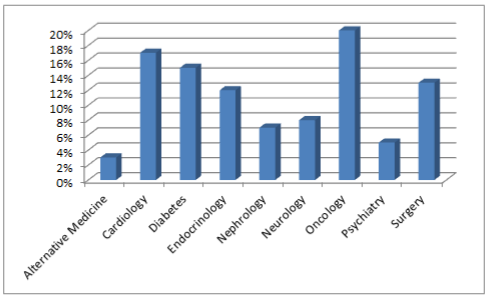


Figure 2: The distribution of the medical sub-domains in the corpus.

A very important component of the BioRo corpus is represented by the texts obtained from online sources (322,005 tokens and 18,226 sentences), free of copyright restriction and which will be freely available to the community.

In order to facilitate cross-linguistic comparison, each file included in this part of the BioRo corpus have another metadata scheme, enhanced with standard categories used in the biomedical corpora (Figure 3). The medical categories are extracted from the Medical Subject Headings(MeSH) [1] thesaurus. Table 2 describes the sub-corpus differentiating by biomedical sub-domains.

4. Corpus Annotation

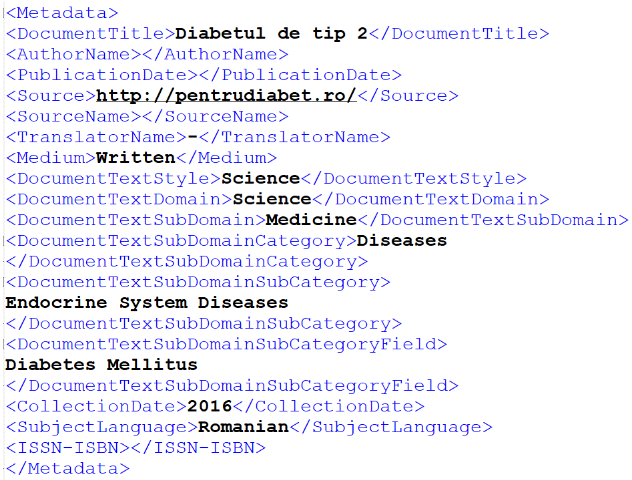
The entire corpus was subjected to an annotation phase using the Tokenizing, Tagging and Lemmatizing (TTL) text processing platform developed at RACAI (Ion, 2007).

TTL is a Perl platform supporting different languages such as Romanian, English and French, and performs the following functionalities: named entity recognition (NER), sentence splitting, tokenization, POS tagging and chunking.

[1] https://meshb.nlm.nih.gov/treeView

	Sentences	Tokens	Content words	Unique lemmas	Punctuation
Neurology	7,632	78,901	44,739	8,940	15,973
Diabetes	5,736	124,035	66,401	9,102	17,916
Endocrinology	1,840	40,038	21,514	4,928	5,915
Cardiology	1,235	33,674	18,782	4,100	4,355
Oncology	1,127	28,286	15,902	3,963	3,587
Nephrology	536	13,753	7,590	2,449	1,977
Alternative medicine	120	3,318	1,774	872	336
Total	**18,226**	**322,005**	**176,702**	**34,354**	**50,059**

Table 2: Statistics over the biomedical sub-corpus extracted from online sources.



```
<Metadata>
<DocumentTitle>Diabetul de tip 2</DocumentTitle>
<AuthorName></AuthorName>
<PublicationDate></PublicationDate>
<Source>http://pentrudiabet.ro/</Source>
<SourceName></SourceName>
<TranslatorName>-</TranslatorName>
<Medium>Written</Medium>
<DocumentTextStyle>Science</DocumentTextStyle>
<DocumentTextDomain>Science</DocumentTextDomain>
<DocumentTextSubDomain>Medicine</DocumentTextSubDomain>
<DocumentTextSubDomainCategory>Diseases
</DocumentTextSubDomainCategory>
<DocumentTextSubDomainSubCategory>
Endocrine System Diseases
</DocumentTextSubDomainSubCategory>
<DocumentTextSubDomainSubCategoryField>
Diabetes Mellitus
</DocumentTextSubDomainSubCategoryField>
<CollectionDate>2016</CollectionDate>
<SubjectLanguage>Romanian</SubjectLanguage>
<ISSN-ISBN></ISSN-ISBN>
</Metadata>
```

Figure 3: A example of metadata scheme for biomedical domain.

TTL's tokenizer is language independent, recognizes multiword expressions (MWEs), clitics and contractions, assuming that language-dependent resources are available. The POS tagger uses tiered tagging methods (Tufis, 1999) and it is a reimplementation of the Hidden Markov Models (HMM) tagger described in Brants (2000). For the Romanian language the MSD tagset has 614 (Boros et al., 2013) labels compatible with the MULTEXT-East morpho-lexical specifications [2]. External annotations compatible with the Universal Dependencies format can also be generated using an existing multilingual platform from text processing which is available online[3] (Dumitrescu et al., 2017).

After the completion of the POS tagging step the lemmatization begins and a human-validated Romanian word-form lexicon with almost 1,200,000 entries is used by the TTL lemmatizer. In the case of out-of-dictionary words, the lemmatizer uses a five-gram letter Markov Model-based guesser to select the most probable lemma.

Another functionality performed by the TTL platform is chunking. This process is guides by a set of rules based on regular expressions applied on MSDs. The TTL chunker deals with recognizing nominal, verbal, adjectival, adverbial and prepositional phrases.

The biomedical corpus was annotated with the baseline TTL model, which is trained over texts corrected by trained linguists at word-level. The POS tagging accuracy for the general purpose Romanian language is over 98% (Tufis, 1999), and for the biomedical domain the accuracy is 97.83% (Mitrofan and Ion, 2017).

Following the automatic annotation step all the tokens included in the corpus the lemmas and POS tagging were submitted to a partial manual revision. The main clear cases of errors produces by the tagger were wrong lemmas and wrong POS-tags. Table 3 describes the distribution of content words according to the POS tags types found in the sub-corpus extracted from online sources. A important feature of the biomedical domain, also present in table 3, is a higher frequency of nouns and adjectives.

5. The Availability of the Data

The entire BioRo corpus, part of CoRoLa corpus is available for query via KorAP interface (Diewald et al., 2016; Banski et al., 2014; Banski et al., 2013). The search results will be downloadable (this facility, taking into account the IPR restrictions, is under development in KorAP). The KorAP interface allows (among other things) building virtual corpora by observing IPR restrictions - if any, multiple types of linguistic interrogations, various levels of annotation etc. The sub-corpus obtained from online sources (322,005 tokens and 18,226 sentences) will be freely available for download [4] and non-commercial use. The sub-corpus will be accessible in both raw text and annotated formats.

6. Conclusion and Future Work

We presented the BioRo corpus which contains morpho-syntactically annotation. We described the corpus and the annotation process. To our knowledge this is the first biomedical corpus for the Romanian language compiled for biomedical text mining. Although CoRoLa corpus, thus BioRo, is not downloadable, an important part of BioRo is free for non-commercial use.

We plan to annotate it with biomedical named entities and to parse it with the Romanian Universal Dependencies parser developed in the SSPR project (Mititelu et al., 2016). Also we are in the process of creating a tool which is able to automatically label text with BioNER entities, which is trained on the described corpora. We have experimented with multiple strategies (Boros, 2013; Boroş et al., 2017;

[2] http://nl.ijs.si/ME/V4/msd/html/

[3] http://slp.racai.ro/index.php/mlpla-new/

[4] http://slp.racai.ro/index.php/resources/

	Nouns	Verbs	Adjectives	Adverbs	Total
Neurology	26,514	6,489	10,291	1,445	44,739
Diabetes	38,463	13,459	11,667	2,812	66,401
Endocrinology	11,946	4,545	4,240	783	21,514
Cardiology	10,706	3,444	4,001	631	18,782
Oncology	8,538	3,261	3,629	474	15,902
Nephrology	4,357	1,332	1,604	297	7,590
Alternative medicine	885	475	331	83	1,774

Table 3: Statistics over the biomedical sub-corpus extracted from online sources.

Boros and Dumitrescu, 2018) and we are currently turning toward graph-based decoding of named entities.

7. Bibliographical References

Banski, P., Bingel, J., N.Diewald, Frick, E., Hanl, M., Kupietz, M., Pezik, P., Schnober, C., and Witt, A. (2013). The new corpus analysis platform at ids mannheim. In *Proceedings of the 6th Language and Technology Conference: LTC'13. 2013.*

Banski, P., Diewald, N., Hanl, M., Kupietz, M., and Witt, A. (2014). Access control by query rewriting. the case of korap. In *Proceedings of the Ninth Conference on International Language Resources and Evaluation.*, pages 3817–3822.

Boros, T. and Dumitrescu, S. D. (2018). Multilingual tokenization and part-of-speech tagging. lightweight versus heavyweight algorithms. *Lecture Notes in Artificial Intelligence*, Human Language Technology. Challenges for Computer Science and Linguistics.

Boros, T., Ion, R., and Tufis, D. (2013). Large tagset labeling using feed forward neural networks. case study on romanian language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 692–700.

Boroş, T., Pipa, S., Mititelu, V. B., and Tufiş, D. (2017). A data-driven approach to verbal multiword expression detection. parseme shared task system description paper. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 121–126.

Boros, T. (2013). A unified lexical processing framework based on the margin infused relaxed algorithm. a case study on the romanian language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 91–97.

Boytcheva, S., Nikolova, I., Paskaleva, E., Angelova, G., Tcharaktchiev, D., and Dimitrova, N. (2009). Extraction and exploration of correlations in patient status data. in proceedings of the workshop on biomedical information extraction. In *Proceedings of the Workshop on Biomedical Information Extraction*, pages 1–7.

Brants, T. (2000). Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231.

Diewald, N., Margaretha, M. H. E., Bingel, J., Kupietz, M., Banski, P., and Witt, A. (2016). Korap architecture diving in the deep sea of corpus data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3586–3591.

Dumitrescu, S. D., Boroş, T., and Tufiş, D. (2017). Racai's natural language processing pipeline for universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 174–181, Vancouver, Canada, August. Association for Computational Linguistics.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian (in Romanian).* Ph.D. thesis, Romanian Academy.

Kim, J., Ohta, T., Tateisi, Y., and Tsujii, J. I. (2003). Genia corpus - a semantically annotated corpus for biotextmining. *Bioinformatics.*

Mititelu, V. B., Ion, R., Simionescu, R., Scutelnicu, A., and Irimia, E. (2016). Improving parsing using morphosyntactic and semantic information.

Mititelu, V. B., Tufis, D., and Irimia, E. (2018). The reference corpus of the contemporary romanian language (corola). In *Proceedings of the 11th Language Resources and Evaluation Conference-LREC*, page In This Volume.

Mitrofan, M. and Ion, R. (2017). Adapting the ttl romanian pos tagger to the biomedical domain. In *Proceedings of the Biomedical NLP Workshop associated with RANLP 2017*, pages 8–14.

Mowery, D. L., Velupillai, S., and Chapman, W. W. (2012). Medical diagnosis lost in translation: analysis of uncertainty and negation expressions in english and swedish clinical texts. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics.*

Neveol, A., Grosjean, J., Darmoni, S. J., and Zweigenbaum, P. (2014). Language resources for french in the biomedical domain. In *LREC*, pages 2146–2151.

Neves, M., Damaschun, A., Kurtz, A., and Leser, U. (2012). Annotating and evaluating text for stem cell research. In *Proceedings of the Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012).*

Ohta, T., Pyysalo, S., Tsujii, J., and Ananiadou, S. (2012). Open-domain anatomical entity mention detection. In *Proceedings of ACL 2012 Workshop on Detecting Structure in Scholarly Discourse (DSSD)*, pages 27–36.

Tufis, D., Ion, R., Ceausu, A., and Stefanescu, D. (2008). Racai's linguistic web services. In *Proceedings of the 6th Language Resources and Evaluation Conference-LREC*, pages 28–30.

Tufis, D. (1999). *Tiered tagging and combined language models classifiers.* Springer.

Velupillai, S. (2012). *Shades of certainty: annotation and classification of swedish medical records.*

Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, Ã., Grabar, N., and Darmoni, S. (2005). Umlf: a unified medical lexicon for french. *International Journal of Medical Informatics*, pages 119–124.