biomedicines text mining,
named entity recognition,
synonyms and abbreviation extraction

Katarzyna WĘGRZYN-WOLSKA*

LINKING WITH BIAM: SEARCHING FOR DRUGS AND PHARMACEUTICAL SUBSTANCES

The quantity of biomedical publications is growing at an exponential rate. With such explosive growth of the content, it is more and more difficult to locate, retrieve and manage the resulting information. This is why text mining has become a necessity. The main goal of biomedical research is to put knowledge to practical use in the form of diagnoses, prevention, and treatment. It is important to pool the resources between the different individuals researching results. The objective of this paper is to discuss the variety of issues and challenges surrounding the perspectives regarding the use of Information Retrieval and Text Mining methods in biomedicine. The article will first look at the directions in biomedical TM and then describe the work done for the BIAM project, the French on-line Medical Data Base.

1. INTRODUCTION

The volume of biomedical publication is growing at an exponential rate. With such explosive growth of the content, it is more and more difficult to locate, retrieve and manage the resulting information. That is why text mining has become a necessity. The main objective of biomedical TM is to enable scientists to identify the necessary information as efficiently as possible, finding the relationships between available information by applying algorithmic, statistical, and data management methods to the biomedical knowledge, thus knowledge can be pooled.

This article surveys the main directions in biomedical TM and presents the work done for the BIAM project, the French on-line Medical Data Base. The first section introduces current active areas of research. The next section presents both the general and specific problems of linking information in biomedicine. The conclusion summarises our work and introduces future challenges of biomedical TM.

2. BIOMEDICINE TEXT MINING

Biomedicine is an inter-disciplinary science connecting medicine, chemistry and biology. Thus, the same concept can be analysed according to several approaches.

* Ecole Supérieure d'Ingenieurs en Informatique et Génie de Télécommunication, 77-215 Avon Fontainebleau, France.

2.1. CURRENT AREAS OF RESEARCH

We present the most important areas of research in separate categories of TM task [1][6][9]:

- Named Entity Recognition (NER) - the recognition of terms denoting specific classes of biomedical entities (ex. gene and protein names),
- Text Classification (TC) – the automatic determination if an entire document or only part of it has particular characteristics of interest. Typically the information of interest is provided as a set of relevant (the positive training set) information, or not relevant information (the negative training set). Text classification systems must automatically extract the features that help determine the interest of text.
- Synonym and Abbreviation Extraction - a collection of synonyms and abbreviations which help users automatically find the information
- Relationship Extraction- occurrence detection of a pre-specified type of relationship between a pair of entities of given types; which is usually very specific (ex: genes, proteins, or drugs), the type of relationship very general (ex. any biochemical association) or very specific (ex. a regulatory relationship).
- Hypothesis Generation - uncovering relationships not presented directly in the text, but instead inferred by the presence of other more explicit relationships. The goal is to uncover previously unrecognised relationships worthy of further investigation.
- Integration Frameworks – integration of TM (ex. The “MedScan” [7] system combines lexicons with syntactic and semantic templates in a general-purpose TM system to extract relationships between biomedical entities).

Our work focused on two branches of biomedicine TM, Named Entity Recognition (NER) and Synonym and Abbreviation Extraction. That is why only these two areas are described in more detail.

2.1.1. NAMED ENTITY RECOGNITION

Named Entity Recognition is the most important step in Information Retrieval and Extraction (IR). NER goal is to identify, within a collection of texts, all of the different instances of a name, for example, all of the drug names within a collection of journals. This task is challenging for several reasons. Firstly, no complete dictionary exists for most types of biological entities. The simple text-matching algorithms are not sufficient. The main problems arise from the fact that there is often no one-to-one correspondence between concepts and terms. In addition, the same word or phrase can refer to a different thing depending upon context. Moreover, many biological entities have several names (e.g., PTEN and MMAC1 refer to the same gene). Biological entities may also have multi-word names (e.g., carotid artery).

The NER approach generally uses three categories of recognition: lexicon-based, rule-based, and statistic based.

2.1.2. SYNONYM AND ABBREVIATION EXTRACTION

Parallel to the growth of biomedical documents, is the growth in biomedical terminology. Many biomedical entities have multiple names and abbreviations. It would be very useful to have a collection of all existing synonyms and abbreviations. Furthermore, other TM tasks could be done more efficiently if all of the synonyms and abbreviations for one entity could be mapped by a single term.

3. FRENCH MEDICAL DATA BASE: BIAM

BIAM[14] (*Banque des Données Automatisée sur les Medications*) is one of the French data bases specialised in the cataloguing of drugs and substances used by pharmaceutical laboratories. It was created by the associated initiatives of French universities and the pharmaceutical industry. This free-access DB is used 60% by doctors, 30% by pharmaceutics employers and 10% by other health professionals. In order to provide information about drugs and active substances it would be useful to supply links to additional information already existing on the Web. It is evident that these links must be as reliable as possible.

4. AUTOMATIC LINKING

4.1. THE OBJECTIVES

The objectives of our work consisted of searching for and locating the content corresponding to the BIAM interrogated pages and thus automatically generating the corresponding links. Searching such specialised information as we find in biomedicine using general-purpose search engines such as Google is neither reliable nor efficient. That is why when querying biomedical publications, it is better to develop a specialised IR tool.

4.2. BIAM CONTENT

BIAM contains the descriptions of drugs and substances used by pharmaceutical laboratories:

Pharmaceutic products: over 4200 products,

Equivalences – French terms and their corresponding foreign product names

- Over 3000 substances with the appropriate information such as: active ingredient identification (DCI and other denominations), chemical form and chemical class.

Active substances

- desired effects
- pharmaceutical properties, therapeutic indications
- undesirable and unpleasant side effects, possibility of addiction
- precautions and disqualification for use

- list of medical tests, which can be affected by the use of certain drugs
- overdose signs and treatment
- pharmacological-addiction possibilities
- dosage, mode of administration
- general bibliographical references

Drugs interactions – over 10 000 pairs of interaction between the active substances.
The BIAM data base is updated every week.

4.2.1. INTERROGATED MEDICAL DATA BASES

To perform our test we interrogated the specialised, databases accessible on-line on the Web:

- Clinical Pharmacology[15] and Alchemy[16] of the "Gold Standard Multimedia" GSM which are the guides of utilisation to the most popular and often used drugs.
- RxList[17]; the database of the top of 200 of drugs.
- MedicineNet[18] provides a lot of medical services like on-line doctor consultation, reliable produced health and medical information and has the hug medication's data base which contents over 2500 common drugs.
- Internet Mental Health[19]; since 1995, Internet Mental Health has provided information on mental health free-of-charge.
- The data published by the Cancer Imaging Program[20] of National Cancer Institute.
- National Toxicology Program[21] by National Institute of Environmental Health Science; with the tree specialised agencies; National Institute of Environmental Health Sciences of the National Institutes of Health (NIEHS/NIH), National Institute for Occupational Safety and Health of the Centres for Disease Control and Prevention (NIOSH/CDC), National Centre for Toxicological Research of the Food and Drug Administration (NCTR/FDA).
- DRUG Infonet[22] provides drug and disease information about the healthcare, we can find here the answers to common health questions, the links to pharmaceutical company pages, etc…
- US Food and Drug Administration[23]; which provides with the special Center for Drugs Evaluation and Research same very interesting databases like Adverse Event Reporting System for all approved drug and therapeutic biologic products, Approved Drug Products with Therapeutic Equivalence Evaluations, Drugs@FDA with information about FDA-approved brand name and generic prescription and over-the-counter human drugs and biological therapeutic products. Drugs@FDA includes most of the drug products approved since 1939. The majority of patient information, labels, approval letters, reviews, and other information are available for drug products approved since 1998; Drug Firm Annual Registration Status database which allows to search for information submitted by drug firms, etc…
- Enviro-Net Environmental Professionals[24]; drugs database done by the University of Utah.

- EUSHC Labs[25]; "Electronic Laboratory Manual" from "The Emory University System of Health Care".

4.3. DATA SEARCHING AND EXTRACTION

The main task is to determine not the similar information but the information which is precisely corresponding to the "drug and substance" from BIAM. This information is used to linking the data. Each substance in BIAM Data Base is identifying by catalogue number, its principal name (generic name), synonyms and the CAS® Registry Number[26].

Finding the correct nomenclature for a particular drug is very important part of searching for pharmaceutical information. Generally, the drug data bases identify the product by its generic names, trade names, lab codes, CAS® Registry Numbers, synonyms for drugs and sometimes also other molecular entities.

4.3.1. SEARCHING BY CAS® NUMBER

CAS registry numbers are unique numerical identifiers for chemical compounds, polymers, biological sequences and mixtures. Chemical Abstracts Service (CAS), a division of the American Chemical Society, assigns these identifiers to every chemical that has been described in the literature. While CAS® Registry number is a unique number for each chemical substance, it could be very effectively used to search the information. Unfortunately, it is not possible to limit the searching only by this method. Some DB doesn't use the CAS identification and prefer to use their own denomination. Moreover, some substances in BIAM data bases haven't the CAS identifier. To connect all of the data it is necessary to provide also the searching by correspondent name.

The main task in linking textual biomedical information with "searching by name" method is to determine that the two names are the denomination of the same substance. One of the major problem to do it is:

- the imprecise and ambiguity terminology,
- the variety of the same substances,
- the variation of denomination in the different languages

It is necessary to determine if the substances, which have similar or "almost identical" name (with the weak semantic difference) can be consider as the same, identical substance. While this task can be very simple for an expert, sometimes it can be much more complicated to proceeding it automatically without human interaction.

In practice, our application is faced with the problems of term variation and term ambiguity, which make the integration of information available in text difficult. Term variation originates from the ability of a natural language to express a single concept in a number of ways. For example, in biomedicine there are many synonyms for proteins, enzymes, genes, etc. Having several synonyms for a single substance is very often in this domain. The probability that two experts use the same term to denominate the same entity is less than 20% [12]. In addition, biomedicine includes pharmacology, where numerous trademark names refer to the same compound (ex. Advil, Brufen, Motrin, Nuprin and Nurofen all refer to ibuprofen). Term ambiguity occurs when the same term is used to refer to multiple concepts. Ambiguity is an inherent feature of natural language. Words typically have multiple dictionary entries and the meaning of a word can be altered by its context.

How "delicate" is this determination we can observe in the following examples:

Example 1.

a) *ra-**n**-itidine*
*ra-**m**-itidine*

Can we consider these two substances as the identical? They seems to be the same, that have the name and the pharmaceutic proprieties similar, but they **are not identical**. - FDA (Federal Drug Approvals – US) data bases provides for these two names two different numbers App_N020095 et App_N020251.

b) *ranitidin-**e***
*Ranitidin-**a***
ranitidin

These two substances **are identical**, the morphological difference is caused by their origin denomination's inscription (different natural languages).

Can we suppose (or determine), that if the names are different only on the last position (like in the last example **-a -e** -) the drugs or the substances are identical? If the difference is caused only by the different natural language grammar (declination, lemmatisation during indexing, etc...).

The next two examples show that our last hypothesis was wrong, and that it is impossible to determine the two identical substances without the full form of the name.

Example 2.

a) *vitamine**A***
*vitamine**E***
for two **different** substances.

b) ***vitamina***
vitamineA
vitaminaA
for identical substances.

Next example illustrates the Synonym and Abbreviation Extraction problems (Section0).

Example 3.

The „Cimitidine" is identify in BIAM Data Bases by the follows denomination:

- CIMETIDINE
- RANITIDINE
- NIZATIDINE
- FAMOTIDINE
- Numéro CAS 51481-61-9

It is evident, how advantageous should be the lexicon of synonyms and abbreviations.

5. CONCLUSION AND FUTURE CHALLENGES

The objectives of our work consisted in searching and locating the content corresponding to the BIAM interrogated pages and than to generate automatically the corresponding links. The most important and at the same time difficult part of developed linking system was determination if the corresponding drugs and substance are identical.

The results of the developed system are satisfied both in qualities and reliabilities of the found links.

The most important direction for future progress is interdisciplinary coordination and cooperation. TM researchers; publishers, and biomedical researchers have to work together to design the systems that produce consistent, measurable, and verifiable results.

BIBLIOGRAPHY

[1] ANANIADU S., Text Mining for Biology and Biomedicine, Boston and London: Artech House, 2006, 286 pp; hardbound, ISBN 1-58053-984-X.
[2] BRUIJN B, MARTIN J. Getting to the (c)ore of knowledge: mining biomedical literature. *Int J Med Inf* 2002;67(1-3), pp.7-18.
[3] BRILL E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 1995;21(4):543-565.
[4] Center for Drug Evaluation and Research, Drug Approvals, 1994, http://www.fda.gov/cder/da94.htm
[5] Chemical Abstracts Service http://www.th-darmstadt.de:81/ze/online/stn-info/about.html
[6] COHEN M., A Survey of Current Work in Biomedical Text Mining, Briefings in Bioinformatics, Vol. 6, No. 1.,pp.57-71, 2005.
[7] DZICZKOWSKI G, WEGRZYN-WOLSKA K., Graph Based System purpose-built for automatic retrieval and extraction of the electronics, In Proceeding of "Internet and Multimedia Systems and Applications" IASTED, 2007.
[8] HIRSHMAN L, MORGAN AA, YEH AS. Rutabaga by any other name: extracting biological names. *J Biomed Inform,* 2002, 35(4), pp. 247-59.
[9] IBEKWE-SANJUAN F, Fouille de texte, methods, outils et applications, Hermes Science, LAVOISIR, 2007.
[10] NOVICKOVA S, EGROV S, DARASELIA N., MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* 2003;19(13):1699-706.
[11] SPASIC I., ANANIADU S, McNAUGHTt J, KUMAR A, Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform, Vol. 6, No. 3., pp. 239-251, September 2005.
[12] SWANSON DR. Medical literature as a potential source of new knowledge. *Bull Med Libr Assoc* 1990;78(1), pp. 29-37.
[13] WEGRZYN-WOLSKA K, Etude et realisation d'un robot pour la recherché medical d'information sur le Web, Rapport DEA d'Informatique, Université d'Evry-Val-d'Essonne, Doc. E/193/CRI, 1996.
[14] www.biam2.org/accueil.html
[15] www.clinicalpharmacology.com
[16] www.alchemyrx.com
[17] www.rxlist.com
[18] www.medicinenet.com
[19] www.mentalhealth.com
[20] imaging.cancer.gov
[21] ntp.niehs.nih.gov
[22] www.druginfonet.com
[23] www.enviro-net.com
[24] www.whsc.emory.edu
[25] A CAS® Registry number is a unique number for each chemical substance assigned by Chemical Abstracts Service.