A preliminary approach to recognize generic drug names by combining UMLS resources and USAN naming conventions

Isabel Segura-Bedmar
Computer Sciences Department
Carlos III University of Madrid
Avd. Universidad, 30, Leganés,
28911, Madrid, Spain
isegura@inf.uc3m.es

Paloma Martínez
Computer Sciences Department
Carlos III University of Madrid
Avd. Universidad, 30, Leganés,
28911, Madrid, Spain
pmf@inf.uc3m.es

Doaa Samy
Linguistic Department
Cairo University
Egypt

dsamy@cu.edu.eg

Abstract

This paper presents a system[1] for drug name identification and classification in biomedical texts.

1 Introduction

Numerous studies have tackled gene and protein names recognition (Collier et al, 2002), (Tanabe and Wilbur, 2002). Nevertheless, drug names have not been widely addressed (Rindflesch et al., 2000).

Automating the process of new drugs recognition and classification is a challenging task. With the rapidly changing vocabulary, new drugs are introduced while old ones are made obsolete. Though the terminological resources are frequently updated, they can not follow the accelerated pace of the changing terminology.

Drug receives three distinct names: the chemical name, the generic (or nonproprietary) name, and the brand (or trademark) name. The U.S. Adopted Name (USAN) Council establishes specific nomenclature rules for naming generic drugs. These rules rely on the use of affixes that classify drugs according to their chemical structure, indication or mechanism of action. For example, analgesics substances can receive affixes such as *-adol-, -butazone, -fenine, -eridine* and *–fentanil*. In the present work, we focus, particulary, on the implementation of a set of 531 affixes approved by the USAN Council and published in 2007[2]. The affixes allow a specific classification of drugs on pharmacological families, which ULMS Semantic NetWork is unable to provide.

2 The System

The system consists of four main modules: a basic text processing module, WordNet look-up module, UMLS look-up module and the USAN rules module, as shown in Figure 1.

A corpus of 90 medical abstracts was compiled for the experiment. For the basic processing of the abstracts, GATE[3] architecture is used. This text processing provides sentence segmentation, tokenization and POS tagging. Tokens which receive a noun or proper noun POS tag are extracted.

The nouns found on WordNet are discarded and those which are not found in WordNet are looked up in the UMLS Metathesaurus. If a noun is found in UMLS, it is tagged with its corresponding semantic types as assigned by UMLS. A subset of these nouns is tagged as "drug" if their semantic types are "Pharmacological Substance" or "Antibiotic". Finally, nouns which have not been found in UMLS are tagged as "unknown".

The list of nouns tagged as "drug" is passed to the rule module to detect their pharmacological families according to the affixes. In addition, the rule module processes the list of "unknown" nouns which are not found in UMLS to check the presence of affixes, and thereby, of possible drugs.

3 Preliminary results

[1] This work has been partially supported by the projects: FIT-350300-2007-75 (Semantic Interoperability in Electronic Health Care) and TIN2007-67407-C03-01 (BRAVO: Advanced Multimodal and Multilingual Question Answering).

[2] http://www.ama-assn.org/ama1/pub/upload/mm/365/usan_stem_list.pdf Accessed January 2008

[3] http://www.gate.ac.uk/

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 100–101,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

A manual evaluation by a domain[4] expert was carried out. The list of nouns not found in WordNet contained 1885 initial candidates. This initial list is looked up in UMLS and 93.4% of them (1761) is linked with some concepts of UMLS. The UMLS module recognized 1400 nouns as pharmacological substances or antibiotics. The rest of nouns, 361, are detected by UMLS but neither as pharmacological substance nor as antibiotics.

The expert manually evaluated the set of nouns detected by UMLS as pharmacological substances or antibiotics (1400). Evaluation showed that only 1100 were valid drugs.

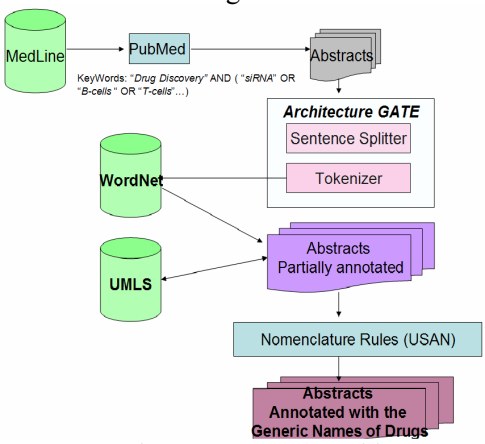

Figure 1 System Architecture

The list of nouns (124) which have not been found in UMLS are processed by the rule module to detect new candidate drugs not included in UMLS. This module only detects 17 candidate drugs. The manual evaluation showed that 7 of them were valid drugs and the rest of nouns are biomedical concepts not included in UMLS. Some of these drugs are *Mideplanin*, *Tomopenem*, *Elvitegravir*, and so on. The rest of nouns neither detected by the UMLS module nor by the rules module, 106, were also validated by the expert in order to estimate the overall coverage of our approach. The evaluation of these nouns shows that only 7 of them are valid drugs, however, the rest of the nouns are named entities of the general domain (organization, person names or cities) or biomedical concepts. Introducing a module of generic NER should decrease the noise caused by such entities.

Finally, precision and recall of the overall system combining UMLS and rules were calculated. The system achieved 78% of precision and 99.3% of recall

3.1 The classification in pharmacological families

Once processed by the rule module, 73.8% of the candidate drugs recognised by UMLS were also classified in pharmacological families by the USAN naming rules. Expert's evaluation of the rule-based classification showed that rules achieved 89% precision. Short affixes such as –ol, –pin and -ox are responsible of the wrong classifications. Thus, additional clues are necessary to detect these drug families.

4 Some Conclusions

As a preliminary approach, it is a first step towards a useful Information Extraction System in the field of Pharmacology. Though evaluation reveals that rules alone are not feasible enough in detecting drugs, but they help to improve the coverage. In addition, rules provide a drug classification in pharmacological families. Such classification is an added value in the development of NLP applications within the pharmacological domain.

For future work, the approach will be extended to address additional information about pharmacologic classes included in many biomedical terminologies integrated in the UMLS such as MeSH or SNOMED.

Future work will also target a wider coverage and a bigger set of drug types through including more affixes, detecting complex entities (multi-words), detecting synonyms, resolving acronyms and ambiguities as well as using contextual information to disambiguate the correct semantic type of each term occurring in the texts.

References

Collier N, Takeuchi K. 2004. Comparison of characterlevel and part of speech features for name recognition in bio-medical texts:423− 35.

Rindflesch, T.C., Tanabe,L., Weinstein,J.N. and Hunter,L. 2000. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac. Symp. Biocomput. 5, 517–528

Tanabe, L. y Wilbur, W.J. 2002. Tagging gene and protein names in biomedical text. Bioinformatics 18, 1124–1132

[4] The authors are grateful to Maria Bedmar Segura, Manager of the Drug Information Center, Mostoles University Hospital, for her valuable assistance in the evaluation of the system.