MC-UC3M Participation at TAC 2017 Adverse Drug Reaction Extraction from Drug Labels

José L. Martínez[1], **Paloma Martínez**[2], **Isabel Segura-Bedmar**[2], **Adrián Carruana**[2], **Ali Naderi**[1], **Conchi Polo**[1]
[1]MeaningCloud LLC, USA
[2]Computer Science Department, Universidad Carlos III de Madrid, Spain
[1]jmartinez,anaderi,polo@meaningcloud.com
[2]pmf,isegura,acarruan@inf.uc3m.es

Abstract

The Adverse Drug Reaction Extraction from Drug Labels ADR track at NIST TAC 2017 aims to extract information on adverse drug reactions (ADR) from drug labels. In this paper, we describe the MC-UC3M'system that participated in this track. The system developed includes text analytics processes provided by MeaningCloud APIs, which are flexible enough to implement dictionaries with ADRs lexicons, and machine learning components based on SVM. MedDRA[1] and SIDER[2] resources have been integrated into the system to identify ADR mentions. Moreover, we studied several corpora containing negation expressions and proposed a set of rules to identify negation particles.

1 Introduction

The Adverse Drug Reaction Extraction from Drug Labels ADR track (Roberts et al., 2017) at NIST TAC 2017 aims to extract information on adverse drug reactions (ADR) from drug labels. The track consists of four subtasks: 1) the identification of ADR and related mentions (Severity, Factor, DrugClass, Negation, Animal); 2) the detection of relations between ADR and related mentions (Negated, Hypothetical and Effect); 3) the identification of positive ADR, that is, those that have not been negated or do not have a Hypothetical relation to a DrugClass or Animal, and 4) the linking of the positive ADR with its corresponding MedDRA Preferred Terms (PT) and Lower Level Terms (LLT).

Text mining and Natural Language Processing (NLP) methods can be used to gather significant information on ADRs and Drug-Drug interactions (DDIs) from different and heterogeneous textual sources, supporting researchers and clinicians with the challenging task of improving patient safety as well as transforming unstructured into structured data. Dealing with health related data requires dealing with complex vocabularies and specific syntax. This means that linguistically motivated natural language processes demand dictionaries and syntactic rules developed specifically for the health domain.

There are several resources developed by different institutions that can be applied for this purpose, such as SNOMED[3], ICD[4], ATC[5] and other vocabularies, most of them included under the umbrella of UMLS[6]. Of course, if adverse reactions are taken into account lexicons such as MedDRA must be integrated in those linguistic based processed. If patient oriented vocabulary has to be analyzed, then such technical terminologies are not adequate. Initiatives such as Consumer Health Vocabulary[7] help to understand what people is saying in social media (blogs, twitter, etc.) as well as is a useful resource to translate medical knowledge to lay people.

Concerning recognizing ADR, three types of research can be distinguished depending of the sources

[1]https://www.meddra.org
[2]http://sideeffects.embl.de
[3]http://www.snomed.org
[4]http://www.who.int/classifications/icd/en/
[5]https://www.whocc.no/atc_ddd_index/
[6]https://www.nlm.nih.gov/research/umls/
[7]http://www.consumerhealthvocab.org/

analyzed (drug labels, MedLine articles and social media). Li et al. (2013) described a hybrid NLP system that combines Conditional Random Fields (CRF) algorithms with hand crafted rules and UMLS to identify medical conditions (disease and signs) from FDA drug labels as part of a framework to build a database with information extracted. The system was evaluated using 52 drug labels manually annotated obtaining an F-measure of 0.90.

Gurulingappa et al. (2012a) introduced a system not only to recognize ADR but also relations between ADR and drugs. The pipeline integrates a Support Vector Machines (SVM) method for relation extraction and it was evaluated using the ADE corpus, Gurulingappa et al. (2012b), which contains 2972 MedLine case reports. The system obtained an F-measure of 0.87.

Xu and Wang (2013) developed pattern-learning approach to extract drug-disease pairs (indications) from 20 million MedLine biomedical abstracts in order to complement information in existing drug-disease treatment databases. The system integrates a drug lexicon from DrugBank and a disease lexicon form UMLS and Human Disease Ontology[8]. These resources were used to annotate drug-disease pairs in clinical trials XML files extracted from ClinicalTrials.gov and used as seeds to look form textual patterns that contain them from MedLine abstracts. These patterns were used to find new drug-disease pairs from MedLine. The algorithm achieved a precision of 0.904 and a recall of 0.131 in extracting all drug-disease pairs.

In recent years, there has been an increasing interest in the extraction of ADRs from social media as a mean to help in pharmacoviligance tasks mainly motivated by the limited use that patients do of spontaneous report systems to notify ADRs. A relevant survey that describes the approaches that use social media for pharmacovigilance is given in (Sarker et al., 2015). Besides, Sarker et al. (2015) introduced different recent research works about mining the pharmacovigilance literature.

Moreover, these NLP medical concepts and relationships recognizers analyze other languages apart from English; see for instance the system to monitor Spanish health social media streams presented in (Martínez et al., 2016; Segura-Bedmar et al., 2015).

Other works have been devoted to annotate corpora that are required to train and test machine learning NLP methods. Neves (2014) collected the most significant text corpora in the biological domain. Some examples are the corpus explained in (Ginn et al., 2014), which contains 10,822 tweets annotated with 74 drugs and their variants and ADRs, the span of the ADR mention and its UMLS identifier. Other corpora have been focused on other languages, such as the corpus described in (Segura-Bedmar et al., 2014), a Spanish corpus of user comments extracted from a health forum that is annotated with drugs and ADRs.

Ramesh et al. (2014) introduced a corpus of 122 Food and Drug Adsministrations Adverse Event Reporting System (FAERS) narratives with annotated medication information and adverse event entities.

Regarding research projects that have working in ADR, EU-ADR Project (Coloma et al., 2011), focused on combining spontaneous reports with electronic healthcare records (EHR) to investigate adverse drug events in Europe. WEB-RADR project[9] was funded by the Innovative Medicines Initiative (IMI) to address the potential of the reporting of ADRs through mobile applications and the identification of drug safety events in user posts.

Apart from these frameworks, APIs developed by companies and institutions that can be integrated in systems are very valuable development resources. For instance, The OpenFDA project[10] provides open APIs, to access medical device reports, enforcement reports and drug adverse event reports since 2004 annotated with named entities. PatientOpinion[11] gathers patients opinions about health issues and treatments and offers an API.

Finally, a great effort in standardization of medicinal products is being done during last years. In this regard, ISO 11616:2017[12] is aimed to characterize and identify regulated medicinal products during their entire life-cycle by establishing concepts and definitions and describing data elements and their structural relationships.

In this paper, we describe our participation at

[8] http://bioportal.bioontology.org/ontologies/1009

[9] https://web-radr.eu/

[10] https://open.fda.gov/drug/event/

[11] https://www.patientopinion.org.uk/

[12] https://www.iso.org/obp/ui/#iso:std:iso:11616:ed-2:v1:en

TAC ADR track. The proposed system adopts a hybrid approach combining dictionary-based, rules-based and machine learning techniques. This work supposes an important step towards representation, structuration and posterior inference of knowledge about drugs and related information.

2 System architecture

The organizers provided 101 documents annotated with entity mentions (drugsclass, ADR, factor, serverity, negation and animal) and relations (negated, hipothetical, effect). Moreover, they also provided a collection of 2,208 unannotated drug labels (which are publicly accessible). The test dataset was also included into this collection. The data is provided in an XML format. In this section, we describe the approaches that we use to solve the different tasks.

2.1 Subtask 1: Identification of mentions

In our approach, mentions identification is a resource-based task. The terms, lexical entries and linguistic structures used for referencing adverse effects form a very specific set of resources that must be under the review of professionals in the area. Fortunately, there exist a wide range of taxonomies in electronic format containing part of the information needed to identify adverse effects, although they have not been specifically created for that task. Among these resources it is possible to find:

- MedDRA: The Medical Dictionary for Regulatory Activities is a multilingual standard defined to make it easier to share information about medical products consumed by humans. The standard can be used for regulatory communication and monitoring of health products.

- SIDER: The Side Effect Resource is a compilation of drugs available and their identified ADRs. The information is gathered from public documents

- UMLS: The Unified Medical Language System provides an umbrella integrating different terminologies, classifications and coding standards under a unique network, including semantic information.

- ATC: The Anatomical Therapeutic Chemical is a standard defined by the World Health Organization to categorize pharmaceutical products.

- ICD: The International Classification of Diseases is a standard classification defined by the World Health Organization to standardize diagnostics and diseases.

These are only some of the resources available for health data standardization. Someone could think on just including these terms on dictionaries to be used by text analytics processes but some issues arise:

- MedDRA: MedDRA is structured around different axis according to the body part or system affected. It includes terms related with adverse effects, but they are not classified according to the role they play in defining the adverse effect. So, including MedDRA is not enough.

- SIDER: It includes terms referring specifically to adverse effects with the corresponding MedDRA code, allowing the identification of terms of MedDRA referring to adverse reactions.

- Partial matching: adverse effects mentions does not appear exactly as they are referenced in mentioned dictionaries.

- Additional mentions: the goal is not only to identify adverse effects but also severity markers, drug classes, animals, factors, ...

- Negation: Dealing with negation requires dedicated approaches, which are explained in subsection 2.2.1.

The way that MC_UC3M team has approached these problems is based on MeaningCloud Text Analytic Platform[13]. MeaningCloud provides a set of text analytics related APIs such as topic extraction, text classification, sentiment analysis, lemmatization and Part-of-speech (POS) tagging, Deep Categorization among others. The system built to take part in TAC uses GATE[14] as a way to combine the different linguistic capabilities offered by MeaningCloud using the plugin provided to integrate with

[13] http://www.meaningcloud.com

[14] http://gate.ac.uk

GATE. MeaningCloud platform provides an easy way to integrate text analytics capabilities in any application, including domain customization. Domain customization is an important feature for any text analytics platform because the ability to deal with domain vocabulary and linguistic structures is the key to obtain accurate results. This is the case in the health domain; drug, diseases and signs names are not part of the common vocabulary in any language. So, identifying mentions of adverse reactions, severity markers, factor markers, and so on, has been implemented through dedicated MeaningCloud dictionaries. Table 1 shows the size and structure of those dictionaries:

Dictionary	**#entries**
Adverse Reactions	21,826
Factor	41
Severity	158
Animal	27
DrugClass	101

Table 1: MeaningCloud dictionary sizes.

The ADR dictionary has been built from SIDER and from the training collection. Dictionaries for the rest of elements: Factor, Severity, Animal and DrugClass have been built based only on the training collection.

2.2 Subtask 2: Identification of relations

We propose a supervised machine learning approach for extracting semantic relations between adverse reactions and some of the above described mentions (Severity, DrugClass, Negation, Animal and Factor). More specifically, the task is formulated as a classification task of AdverseReation-Other pairs. We apply the well-known Support Vector Machine (SVM) (Cortes and Vapnik, 1995) classifier and a set of lexical features that capture the contextual information around the candidate mentions. To obtain these features, the drug labels were tokenized and lemmatized using the nltk library[15]. The features are described bellow:

- M1TXT: the text of the first mention.
- C1BOW: bag-of-words of the first mention.
- C1POS: part of speech of the first mention.
- M2TXT: the text of the second mention.
- C2BOW: bag-of-words of the second mention.
- C2POS: part of speech of the second mention.
- BWTXT: the text between the two mentions.
- LWLEM: the lemmas between the two mentions.
- PWPOS: the PoS tags between the two mentions.
- WB1TXT: the two tokens before the first mention.
- LB1LEM: the lemmas of the two tokens before the first mention.
- PB1POS: the PoS tags of the two tokens before the first mention.
- WA1TXT: the two tokens after the first mention.
- LA1LEM: the lemmas of the two tokens after the first mention.
- PA1POS: the PoS tags of the two tokens after the first mention.
- WB2TXT: the two tokens before the second mention
- LB2LEM: the lemmas of the two tokens before the second mention
- PB2POS: the PoS tags of the two tokens before the second mention.
- WA2TXT: the two tokens after the second mention.
- LA2LEM: the lemmas of the two tokens after the second mention.
- PA2POS: the PoS tags of the two tokens after the second mention.
- NTOKB: the number of tokens between the two mentions.

[15] http://www.nltk.org/

We use SVM implementation from sklearn library (python). Specifically, we use the SVC implementation of SVM, which is based on libsvm. We use the default values provided by the library for all the optional parameters. The kernel is linear. As the organizers did not provided a development set, we randomly took one-third of the training set as our development set. Then, we train our model using the rest of the training set. The results on the development set are shown in Table 2. It is important to emphasize that this table shows the results of a model that was trained using all gold-standard mentions annotated in the training dataset. In this way, we can evaluate the performance of our relation extraction module in an isolated manner. It is to be expected that when this module is evaluated in the pipeline, its performance will decrease notably because the mentions that form the candidate relation instances are not gold-standard mentions. Hypothetical type shows best results, followed by Effect type.

Type of Relation	P	R	F1
Negated	18.33	7	10.14
Hypothetical	68.63	42.25	52.3
Effect	51.8	34.79	41.62
Macro	53.17	47.17	48.61

Table 2: Task 2 results on the developmet set.

2.2.1 Negation relations identification

In order to identify mentions to negation particles in test documents, a specific model has been developed. The goal is to help the machine learning process developed for Task 2 through the identification of mentions to negation terms. The model developed is based on a set of patterns defined by linguistic professionals using a domain specific language provided by MeaningCloud Insights Engine API. An example of one of these patterns is shown in Figure 1. This pattern lets to identify negations such as "...without associated bleeding events. ...":

This negation model has been applied only to the identification of negative particles, which form the input to a machine learning process devoted to the identification of relations.

```
<<without|exclude|decrease|reduce>> :-
ENTITY{"type":"NegationLeft", "label":"$1"}
+ CONSUME{}
<<{AFFECTEDADR}>> :-
ENTITY{"type":"AffectedAdr", "label":"$1"}

"...without associated bleeding events. ..."
```

Figure 1: Negation pattern example.

2.3 Subtask 3: Identification of positive ADR

For this task only positive ADRs must be identified, that is, AdverseReactions mentions that are not negated or related to a DrugClass by a Hypothetical relation. Identifying ADRs matching these criteria have been done by filtering out ADRs according to the negations detected by the negative patterns defined in subsection 2.2.1. Moreover, ADRs that occur near some hypothetical mention were also rule out.

2.4 Subtask 4: Normalization of positive ADR

Normalization of positive ADRs has been implemented through the integration of lexical resources containing semantic information, including MedDRA codes. The ability of MeaningCloud platform to deal with synonyms and to perform fuzzy matching with words appearing in texts makes it simpler to obtain the corresponding ADR MedDRA code expression.

3 Results and discussion

Precision, recall and F-measure were metrics used to evaluate the different tasks.

Type	P	R	F1
Exact (+type)	54.79	66.33	60.01
Exact (-type)	55.78	66.34	60.60

Table 3: Task 1 results on the test set.

Table 3 shows the results for the task 1. The first row presents the results when types of mentions are taken into account in the evaluation and the second one when not. As can be observed from this table, the results are very close among them. This may indicate that the modules fails in the identification of

the mentions, but not in their classification. It is to be expected that if we integrate additional dictionaries, our recall will increase. Moreover, machine learning methods, such as a CRF algorithm, could help us to improve our performance.

We have made an analysis to know evaluation measures by type of mention (see Table 4). As expected, evaluation results for additional features of an adverse reaction such as severity or factor were poor. The linguistic resources based on training collection do not have a good coverage for animals.

Type	P	R	F1
AdverseReaction	63.82	70.77	67.12
Severity	37.13	49.52	42.44
Factor	4.05	7.65	5.3
Negation	10.59	53.76	17.7
DrugClass	19.23	39.63	25.9
Animal	76.56	56.98	65.33
Macro	54.79	66.33	60.01

Table 4: Task 1 results by type of mention on the test set.

The performance of our relation extraction module is very poor, as it is shown in Table 5. Features encoding richer syntactic information from a dependency parser or a syntactic parser should be integrated into the features set in order to improve our results. However, it should be taken into account that these results have been derived using noisy mentions (i.e., those produced by our previous module for mention recognition). To prove this, the same experiments on the test dataset were performed using the correct mentions (i.e, those manually annotated in the test dataset). The results are shown in Table 6. We can observe that the results are far better than those using noisy mentions.

Type of Relation	P	R	F1
Negated	8.43	4.86	6.17
Hypothetical	5.95	9.56	7.34
Effect	24.94	25.74	25.33
Macro	12.19	15.59	13.68

Table 5: Task 2 results on the test set.

Table 7 shows the results for the identification of positive ADRs. These results show that the negative patterns are effective to rule out non-positive

Type of Relation	P	R	F1
Negated	1.12	27.59	2.15
Hypothetical	35.5	52.49	42.36
Effect	24.93	48.77	32.99
Macro	46.7	49.97	47.32

Table 6: Task 2 results using correct mentions on the test set.

ADRs. Filtering out those ADR near some hypothetical mention appears to be a good clue. On the other hand, most candidates are actually positive ADRs. Table 8 shows the results for the normalization of positive ADRs. The ability of MeaningCloud platform to deal with synonyms and to perform fuzzy matching with words appearing in texts makes it simpler to obtain the corresponding ADR MedDRA code expression.

	P	R	F1
Micro	70.03	71.42	70.71
Macro	69.23	72.93	70.13

Table 7: Task 3 results on the test set.

	P	R	F1
Micro	73.40	80.25	76.67
Macro	72.10	80.38	75.29

Table 8: Task 4 results on the test set.

Conclusions

In this paper, we describe our participation at TAC ADR track. The proposed system adopts a hybrid approach combining dictionary-based, rules-based and machine learning techniques.

The integration of additional terminological resources is required to improve the recall of our module for mention recognition. These dictionaries would also help us to improve the results of the normalization task. Moreover, training a CRF model could help us to improve this module. The improvement of the mention recognition task will have a very possitive effect on the relation extraction module.

Experiments show that the necessity of using a more sophisticated set of features for the relation ex-

traction task. In particular, we plan to use rich syntactic features such as dependencies. We also plan to explore deep learning techniques for classifying the relation instances.

Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain (projects DeepEMR (TIN2017-87548-C2-1-R) and eGovernAbility-Access project (TIN2014-52665-C2-2-R)).

References

Preciosa M Coloma, Martijn J Schuemie, Gianluca Trifirò, Rosa Gini, Ron Herings, Julia Hippisley-Cox, Giampiero Mazzaglia, Carlo Giaquinto, Giovanni Corrao, Lars Pedersen, et al. 2011. Combining electronic healthcare databases in europe to allow for large-scale drug safety monitoring: the eu-adr project. *Pharmacoepidemiology and drug safety*, 20(1):1–11.

Corinna Cortes and Vladimir Vapnik. 1995. Support vector networks. *Machine learning*, 20(3):273–297.

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen OConnor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. 2014. Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*.

Harsha Gurulingappa, Abdul Mateen-Rajpu, and Luca Toldo. 2012a. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):15.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012b. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Qi Li, Louise Deleger, Todd Lingren, Haijun Zhai, Megan Kaiser, Laura Stoutenborough, Anil G Jegga, Kevin Bretonnel Cohen, and Imre Solti. 2013. Mining fda drug labels for medical conditions. *BMC medical informatics and decision making*, 13(1):53.

Paloma Martínez, José L Martínez, Isabel Segura-Bedmar, Julián Moreno-Schneider, Adrián Luna, and Ricardo Revert. 2016. Turning user generated health-related content into actionable knowledge through text analytics services. *Computers in Industry*, 78:43–56.

Mariana Neves. 2014. An analysis on the entity annotations in biological corpora. *F1000Research*, 3.

Balaji Polepalli Ramesh, Steven M Belknap, Zuofeng Li, Nadya Frid, Dennis P West, and Hong Yu. 2014. Automatically recognizing medication and adverse event information from food and drug administrations adverse event reporting system narratives. *JMIR medical informatics*, 2(1).

Kirk Roberts, Dina Demner-Fushman, and Joseph M. Tonning. 2017. Overview of the tac 2017 adverse reaction extraction from drug labels track. In *Text Analysis Conference (TAC) 2017. Workshop notebook papers.*

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen OConnor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of biomedical informatics*, 54:202–212.

Isabel Segura-Bedmar, Ricardo Revert, and Paloma Martínez. 2014. Detecting drugs and adverse events from spanish health social media streams. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pages 106–115.

Isabel Segura-Bedmar, Paloma Martínez, Ricardo Revert, and Julián Moreno-Schneider. 2015. Exploring spanish health social media for detecting drug effects. *BMC medical informatics and decision making*, 15(2):S6.

Rong Xu and QuanQiu Wang. 2013. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. *BMC bioinformatics*, 14(1):181.