TAC SRIE 2018: Extracting Systematic Review Information with MedaCy

Andriy Mulyar, Darshini Mahendran, Luke Maffey, Amy Olex, Grant Matteo, Neha Dill, Nastassja Lewinski and Bridget McInnes

Department of Computer Science
Virginia Commonwealth University
Richmond, VA 23284 USA

Abstract

This draft paper describes our participation in the TAC SRIE 2018 Challenge for extracting experimental design factors for the systematic review process from the method section of biomedical journal articles. We discuss our system, medaCy, a python-based supervised multi-class classification system that uses Conditional Random Fields. Our system place fourth for the challenge. The results show that we are better able to identify Animal and Dose Group design factors than Exposure and Endpoint. This paper includes an analysis of our results and future directions to improve the extraction.

1 Introduction

Systematic reviews are necessary to create an exhaustive summary of current evidence relevant to a research question. They are a fundamental for experimental design and analysis. The process is essentially an information organization task consisting of information retrieval, extraction and analysis task [LB16]. However the time required to conduct a review is not trivial. Petrosino, et al. [Pet99] estimate that it takes approximately six people more than 1,000 hours to complete a systematic reviews.

To address this challenge, we present medaCy, a supervised multi-class classification system to automatically identify experimental design factors for the categories of exposure, animal group, dose group, and endpoint from journal articles describing experiments related to toxicity and health effects of environmental agents.

2 Methods

This section describes the underlying methodology of our NER system.

2.1 MedaCy

The python based NER framework, medaCy, is utilized to facilitate the construction and training of supervised machine learning models for the extraction of experimental design factors from the literature. MedaCy decomposes the learning task of named entity recognition into four sequential components: text tokenization, token-grouping by rule-sets, feature extraction, and training/prediction. Each of these components are tuned and optimized for the particular entities in need of extraction. This framework comes pre-equipped with feature extraction utilities that directly incorporate various tools (e.g. MetaMap [Aro01]) and lexicons (e.g. []). MedaCy's flexibility allows for the easy creation of pipelines and includes several pre-defined pipelines as adjustable starting-off points for the building of highly predictive bio-medical text NER systems.

2.2 MedaCy on Systematic Reviews

The extraction of systematic reviews manifests as an extension of the pre-built medaCy pipeline. Namely, the following steps occur in sequence:

1. **Preprocessing**
    - Remove non-ascii characters from the text documents
2. **Tokenization**
    - Character-level tokenization is applied to the original text.
    - Tokens (at this stage individual characters) are merged together if certain conditions are met. Example merge-rules

include membership in medical lexicons such as pre-defined units of mass or volume or the existence of a mapping to a UMLS concept. These rules are not used to classify entities but rather to aggregate similar characters together prior to feature extraction and supervised model training.

3. **Feature Extraction**.

   - For each merged and un-merged token, the standard set of text processing features are extracted alongside UMLS concept mappings and unit pattern matching in a window size of five. These feature include: morphological, orthographic, lexical, syntactic and semantic.

4. **Machine Learning**

   - The Conditional Random Fields (CRF) supervised classification algorithm is trained and then utilized for prediction. The benefit of CRFs is they are capable of capturing label inter-dependencies.

The incorporation of character-level splitting and later selective merging aid in the decomposition of text into meaningful instances for classification and speculatively, however intuitively, it assists the classification algorithm in overcoming inevitable annotation noise present in training data.

2.3 Replicability and Reproducibility

The utilization of a standardized framework such as medaCy ensures the replicability of the system described above. MedaCy can be found and installed here: *https://github.com/NanoNLP/medaCy*. The system described above has been wrapped under the SystematicReviewPipeline and can be utilized as described in the documentation. To reproduce results found below, a user must run the SystematicReviewPipeline over the TAC 2018 SRIE dataset.

3 Data

The 2018 TAC SRIE dataset provided by the TAC organizers contain experimental design factors for the categories of exposure, animal group, dose group, and endpoint from journal articles describing experiments related to toxicity and health effects of environmental agents. The data was divided into two subsets: training set and test set. The training set includes 97 articles and the test set contains 444 articles of which 100 were annotated as the gold standard. Table 1 shows each the entities categories and each entity type with the number of instances in the training data (Train) and test data (Test).

Table 1: Entities from the TAC SRIE dataset

Category	Entities	Train	Test
Exposure	TestArticle	1831	2207
	Vehicle	417	358
	TestArticlePurity	26	19
	TestArticleVerification	5	2
Animal Group	GroupName	939	1058
	GroupSize	367	496
	SampleSize	43	74
	Species	1585	1639
	Strain	372	338
	Sex	574	608
	CellLine	39	91
Dose Group	Dose	637	611
	DoseUnits	472	441
	DoseFrequency	92	106
	DoseDuration	210	188
	DoseDurationUnits	198	176
	DoseRoute	558	524
	TimeAtDose	115	56
	TimeUnits	594	733
	TimeAtFirstDose	45	66
	TimeAtLastDose	21	44
Endpoint	Endpoint	4316	3756
	EndpointUnitOfMeasure	682	698
	TimeEndpointAssessed	659	830

4 Results

Table 2 shows our reported results and the results of the top performing system. Results show that our system obtains an overall higher precision at extracting the entities than recall.

Table 2: Reported performances of our system and the best system

Team run	TP	FP	FN	Precision	Recall	F_1
Best system	9401	6846	5720	0.579	0.622	0.6
Our system	4183	4582	10893	0.477	0.278	0.351

Table 3 shows the results over each of the different design factor categories. The results show that our system is better able to identify Animal and Dose Group entities than Endpoint and Exposure.

Table 4 shows the Precision, Recall and F-1 score on the training corpus for each of the different entities using 10-fold cross validation. The ta-

Table 3: Performance of each category

Category	Precision	Recall	F_1
Exposure	0.398	0.162	0.215
Animal Group	0.658	0.485	0.668
Dose Group	0.417	0.399	0.414
Endpoint	0.269	0.454	0.194

ble also contains the minimum and maximum F-1 score over the folds. These results are based on an exact match of the system prediction with the human annotations. The results show a high volatility of scores between the folds indicating a high variability of the data within the folds.

Table 4: 10-fold Cross Validation Results on Training Data

Category	Entity	Precision	Recall	F_1	F_1Min	F_1Max
Exposure	TestArticle (1831)	0.559	0.339	0.372	0.265	0.515
	Vehicle (417)	0.578	0.33	0.371	0.201	0.495
	TestArticlePurity (26)	0.398	0.126	0.172	0	0.4
Animal Group	GroupName (939)	0.585	0.334	0.379	0.13	0.491
	GroupSize (367)	0.682	0.547	0.587	0.222	0.724
	SampleSize (43)	0.292	0.2	0.197	0	0.791
	Sex (574)	0.818	0.622	0.69	0.411	0.829
	Species (1585)	0.792	0.632	0.697	0.493	0.777
	Strain (372)	0.816	0.522	0.626	0.397	0.756
	CellLine (39)	0.853	0.502	0.587	0.273	0.778
Dose Group	Dose (637)	0.621	0.481	0.514	0.274	0.605
	DoseDuration (210)	0.429	0.3	0.296	0.1	0.488
	DoseDurationUnits (198)	0.422	0.276	0.304	0	0.634
	DoseFrequency (92)	0.325	0.11	0.148	0	0.444
	DoseRoute (558)	0.547	0.196	0.221	0.03	0.466
	DoseUnits (472)	0.703	0.585	0.623	0.485	0.733
	TimeAtDose (115)	0.252	0.077	0.109	0	0.31
	TimeAtFirstDose (45)	0.02	0.033	0.025	0	0.25
	TimeAtLastDose (21)	0	0	0	0	0
	TimeUnits (594)	0.549	0.305	0.378	0.224	0.55
Endpoint	Endpoint (4316)	0.528	0.164	0.242	0.157	0.337
	EndpointUnit0fMeasure (682)	0.39	0.11	0.126	0.022	0.259
	TimeEndpointAssessed (659)	0.405	0.19	0.242	0.131	0.37
System	Train	0.474	0.356	0.311	0.114	0.397

Table 5 shows the Precision, Recall and F-1 score on the test corpus of our submitted results using a mention threshold of 0.5. Given the threshold these results are lower than what would be expected from the training data. We believe that this is due to the removal of the non-ascii characters from the data set which throws off our spacing.

Overall the test results show that our system obtains a higher precision than recall for a majority of the entities. This indicates the variability in the entities is quite high.

5 Error Analysis

One limitation of current system is that it does not take non-ascii characters as input. Our pre-preprocessing step removes contiguous non-ascii characters replacing them with a space. This allows for utf-16 characters to be replaced with a single space but threw off our character count when there were two utf-8 characters were in a row. For example, *-TCR-PerCP* has two utf-8 characters and should be replaced with two spaces; and *mean SEM* contains a utf-16 character and should be replaced with one character. Our preprocessing did not take this into consideration.

Table 5: Reported results for each entity types on test data

Category	Entity	Precision	Recall	F_1
Exposure	TestArticle (2207)	0.606	0.117	0.196
	Vehicle (358)	0.463	0.264	0.336
	TestArticlePurity (19)	0.125	0.105	0.114
Animal Group	GroupName (1058)	0.537	0.241	0.333
	GroupSize (496)	0.713	0.447	0.549
	SampleSize (74)	0.444	0.541	0.964
	Sex (608)	0.889	0.697	0.781
	Species (1639)	0.855	0.607	0.71
	Strain (338)	0.766	0.421	0.544
	CellLine(91)	0.4	0.440	0.792
Dose Group	Dose (611)	0.458	0.251	0.325
	DoseDuration (188)	0.145	0.642	0.889
	DoseDurationUnits (176)	0.354	0.971	0.153
	DoseFrequency (106)	0.548	0.219	0.313
	DoseRoute (524)	0.584	0.508	0.543
	DoseUnits (441)	0.473	0.283	0.354
	TimeAtDose (56)	0.682	0.536	0.6
	TimeAtFirstDose (66)	0.588	0.152	0.241
	TimeAtLastDose (44)	0.5	0.227	0.435
	TimeUnits (733)	0.522	0.198	0.287
Endpoint	Endpoint (3756)	0.291	0.254	0.271
	EndpointUnit0fMeasure (698)	0.252	0.981	0.141
	TimeEndpointAssessed (830)	0.263	0.126	0.17
System	Test	0.486	0.283	0.358

Our current system combines *Dose* and *Dose Units*; and *DoseDuration* and *DoseDurationUnits*. In the future, we need to incoporate a post-processing step to separate the entity into their respective subsets.

For the Endpoint entity, the recall results are quite low. One difficulty with Endpoint is that the mentions of the endpoint are not contiguous. For example, in the segement: *uteri were dissected free of adhering fat and connective tissue, drained of intraluminal fluid, and weighed.* The endpoint label is *uteri weighed.* Currently, our system does not support this type of annotation; and trains the entire segment as an Endpoint. This increases the variability of endpoints in the data making it difficult to learn.

6 Conclusion

This paper described our participation in the TAC SRIE 2018 Challenge for extracting experimental design factors for the systematic review process from the method section of biomedical journal articles. We discuss our system, medaCy, a python-based supervised multi-class classification system that uses Conditional Random Fields. Our system place fourth for the challenge. The results show that we are better able to identify Animal and Dose Group design factors than Exposure and Endpoint.

References

[Aro01] Alan R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. 2001.

[LB16] Ana Lucic and Catherine L Blake. Improving endpoint detection to support automated systematic reviews. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1900. American Medical Informatics Association, 2016.

[Pet99] Anthony Petrosino. Lead authors of cochrane reviews: Survey results. *Report to the Campbell Collaboration. Cambridge, MA: University of Pennsylvania*, 1999.