Overview of the TAC 2018 Systematic Review Information Extraction Track

Charles Schmitt[1], Vickie Walker[1], Ashley Williams[2], Arun Varghese[2], Yousuf Ahmad[2], Andy Rooney[1], Mary Wolfe[1]

1. National Toxicology Program, National Institute of Environmental Health Sciences, Durham, NC
2. ICF, Durham, NC

Abstract

This paper describes the Systematic Review Information Extraction (SRIE) track that was conducted as part of the 2018 National Institute for Standards and Technology (NIST) Text Analysis Conference (TAC). Participating teams were provided with annotated text passages of methods sections from research articles in PubMed. The annotations focused on details of experimental methods as well as grouping of the details into concepts. Participants were challenged in two tasks to develop computer models that could extract the mentions (Task 1) and group the mentions into concepts (Task 2). Seven teams submitted at least one run with 18 submissions in total across both tasks.

Background

The National Toxicology Program (NTP), the National Institute of Environmental Health Sciences (NIEHS, part of the National Institutes of Health), and the Environmental Protection Agency (EPA) routinely conduct systematic reviews of environmental agents to identify potential human health hazards. These reviews collect toxicity and health effects information on different chemicals from the published scientific literature including study details such as outcomes assessed, test methods, animal models, and results. Because this information can vary widely from study to study, systematic reviews serve a critical purpose by providing a transparent, standardized, multistep approach to identify, select, critically assess, and synthesize information for developing objective, evidence-based conclusions about potential chemical hazards [7]. Furthermore, because research practices and reporting procedures change over time, the systematic review approach serves to promote transparency and facilitate reproducibility of literature-based evaluations on environmental agents [12].

Some elements of information extraction (IE) in systematic reviews are straightforward, such as identifying the species or sex of the experimental models. Others are more complex such as the results as publications may report multiple experiments with various exposures and doses and evaluate several endpoints. Authors may report experimental details using different measurement units, different names for the same chemical, and other variations in terminology. In addition, this information may be located in the text of the publication or in a table, figure caption, or the figure itself. Currently, the information extracted in a systematic review is collected through a labor-intensive, manual, and well-structured process [11] that is slow and often costly. NTP and EPA are interested in developing and adopting automated or semi-automated processes for IE in systematic reviews of environmental chemicals in order to reduce time and labor-costs while maintaining quality and reproducibility.

The purpose of the SRIE track was to develop and evaluate IE methods that would increase the use of automation in systematic reviews of potential health effects from exposure to environmental agents. This track focused on IE of experimental design factors found in the Material and Methods section ("methods section") of published studies of experimental animals exposed to environmental chemicals. The first goal of the track was to identify and annotate the experimental design factors. The second goal of the track was to identify relations between different experimental design factors and assign the factors into logical groups.

Related Work

The use of structured systematic review has been adopted in multiple application areas, including hazard identification, clinical and public health interventions, adverse effects assessment, and economic evaluations. Although there is variation in the procedures across these disciplines, the systematic review methodology has a common multistep process including problem formulation, identification and selection of relevant documents/articles, de-duplication of articles, data extraction, risk of bias assessment of individual studies, data analysis or meta-analysis, and evidence integration [7]. While efforts exist in all of these steps to increase the use of automation, there exists a deficit in automated and semi-automated tools to aid in the data extraction stage [9], in part due to the lack of training sets and comparative metrics for developing data extraction algorithms. As recently noted in Jonnalagadda's review of data extraction automation efforts [4], "Biomedical natural language processing techniques have not been fully utilized to fully or even partially automate the data extraction step of systematic reviews."

The NIST TAC challenge and associated SRIE training and test sets are meant to help address this gap by beginning to develop gold standard corpora and performance benchmarks for future methods. A limited number of related data sets exist, including the Cochrane Database of Systematic Reviews (CDSR) [3]. The CDSR contains a large set of manually annotated systematic reviews that have been used in model development, e.g., to develop risk of bias models [6], to aid in data extraction in clinical reviews [2], and to allow the use of distant supervision as a model development technique to overcome the lack of training data [10]. Jonnalagadda et al. includes an assessment of data sets that have been used to support model or tool development. The list of data sets includes 17 generated from abstracts only and 9 generated from portions of full text articles. Only 11 of the data sets focus on the extraction of concepts versus identification of relevant sentences and only 5 focus on the extraction of concepts from full text. Most data sets are specific to clinical interventions, e.g., PICO (Patient, Intervention, Condition, Outcome), PECO (Patient, Exposure, Condition, Outcome), or PIBOSO (Patient, Intervention, Background, Outcome, Study Design, and Other). Jonnalagadda et al. also includes an assessment of 26 published data extraction systems. Importantly, only 3 used a common corpus, the PIBOSO corpus, which hinders direct comparison of the systems. The PIBOSO corpus, which was developed from 1000 medical abstracts, is targeted at classification of sentences [5] and was released under a Kaggle competition [1].

In the area of hazard and exposure related systematic reviews, we are not aware of other openly available corpus that can support development of data extraction tools and serve to provide benchmarks for methods development.

SRIE Tasks

Task 1: Task 1 focused on accuracy (F1 score) in extracting mentions of experimental design factors, such as species of animal, substances that animals were exposed to, and dose of exposure. This is similar to natural language processing (NLP) named entity recognition (NER) tasks.

Task 2: Building on Task 1, Task 2 focused on accuracy (F1 score) in grouping related mentions extracted as part of Task 1. This is similar to NLP slot filling tasks.

Data

The SRIE track targeted IE of experimental design factors found in the Materials and Methods section of published experimental animal exposure studies. Extracted data included the mention of specific types of entities (mention annotations) as well as grouping of those entities into related concepts (grouping annotations). Table 1 lists the mention and group types that our team selected for the 2018 SRIE TAC challenge as relevant for animal exposure studies. This list is a subset of types that we considered of interest for animal exposure studies, for instance it does not include mentions such as 'Endpoint Method' or 'Test Article Source'. Those additional types may be pursued in future efforts.

Table 1: Annotation Types for Animal Exposure Study Methods

Category	Annotation Tag	Description
Exposure	TestArticle	Test article or exposure evaluated
	Vehicle	The solution the test article is in
	TestArticlePurity	Purity of test article
	TestArticleVerification	Text indicating that the test article was confirmed, if present, typically just a statement saying the purity was confirmed by a third party
Animal Group	GroupName	If reported, a name given to animal treatment groups (e.g., 'DES-10', 'treated') or control groups ('negative control', 'positive control').
	GroupSize	The number of animals in a group where a group is a set of animals given the same dosing regimen or used for an endpoint measurement.
	SampleSize	The number of animals used in an experiment
	Species	The species names
	Strain	The strain names
	Sex	Sex of the animal group(s)
	CellLine	The cell line name used in the experiment
Dose Group	Dose	Dose
	DoseUnits	Units of dose
	DoseFrequency	Frequency at which doses are given
	DoseDuration	Duration of treatment (dose)
	DoseDurationUnits	Units of dose duration
	DoseRoute	Route of administration
	TimeAtDose	Time when dose is given (typically the age)
	TimeUnits	Units used for time (typically days)
	TimeAtFirstDose	Time at which first dose is given
	TimeAtLastDose	Time at which last dose is given
Endpoint	Endpoint	Endpoint evaluated
	EndpointUnitOfMeasure	Units of measured endpoint
	TimeEndpointAssessed	Time at which the endpoint was accessed (typically number of days after some event)

Group	Annotation Tag	Description
TestArticleGroup	TestArticle	Test article or exposure evaluated
	Vehicle	The solution the test article is in
	TestArticlePurity	Purity of test article
	TestArticleVerification	Verification of purity of test article
AnimalGroup	Species	The species names
	Strain	The strain names
	Sex	Sex of the animal group(s)
	Group name (possibly)	Animal group name if multi-generational

Selection of research articles: Studies from toxicological, open-access (CC0, CC-BY), peer-reviewed journals in PubMed were selected for the challenge. Studies were randomly assigned to a training set (n=100) and a test set (n=100). See Table 2 for article counts by journal. A majority of studies came from four journals (Environmental Health Perspectives, PLoS One, International Journal of Environmental Research and Public Health, Particle and Fibre Toxicology), which is a recognized limitation of the data set that reflects the challenge of developing training sets from open access articles.

Table 2: Article Count by Journal

Journal	Count
Appl Environ Microbiol	1
Basic Clin Pharmacol Toxicol	2
Birth Defects Res B Dev Reprod Toxicol	1
BMC Pharmacol Toxicol	13
Elife	2
Environ Health	14
Environ Health Perspect	246
Environ Health Prev Med	1
Environ Mol Mutagen	7
Food Chem Toxicol	1
Inhal Toxicol	4
Int J Environ Res Public Health	74
J Immunotoxicol	1
J Toxicol Environ Health A	2
Microbes Environ	1
Nanotoxicology	2
Part Fibre Toxicol	84
PLoS Biol	2
PLoS Genet	1
PLoS Med	1
PLoS Negl Trop Dis	4
PLoS One	102
PLoS Pathog	4
Radiat Environ Biophys	2
Toxicol Lett	1
Toxicol Mech Methods	5
Toxicol Sci	**1**

Training/Test Sets: The training data set was released in two parts. The first release, Task 1, consisted of a set of 100 studies with the experimental design factors annotated (the mentions). The second release, Task 2, consisted of the same data from the first release with the inclusion of group annotations.

A test data set was generated and used as the gold standard for evaluation of the submitted IE models. For the test data set, another set of 100 studies was annotated for mentions and groups. Text files of these 100 studies were released to study participants with an additional 344 studies to make up the full test set. The 344 studies were not annotated and were pulled from the same set of journals.

The training and test sets were annotated with the BRAT annotation tool [8]. A customized version of BRAT was developed to support the generation of groups of mentions (see example annotations below). Participants were provided with the training set (100 studies) as text files (*.txt), BRAT-annotation files (*.ann), and XML formatted annotation files (*.xml) (see [13] for details). Participants were given 444 studies of the test set as text files. Challenge participants trained their IE models using the training data set, ran their IE models on the test set, and then submitted their results for evaluation (in .xml format).

Annotation Process: To guide the annotation process and ensure consistency, a set of Annotation Guidelines was developed [13]. The guidelines included a listing of mention and groups types to annotate as well as details on how to use the BRAT software, how to handle fragmented annotations, how to generate the grouping of mentions, and how to use keyboard shortcuts within the BRAT tool to make annotating the studies easier.

For both Tasks 1 and 2, an initial set of pilot annotations was produced on three to five studies by two or more of the annotators and other members of the SRIE team. The guidelines were then updated based on the review and discussion of identified issues.

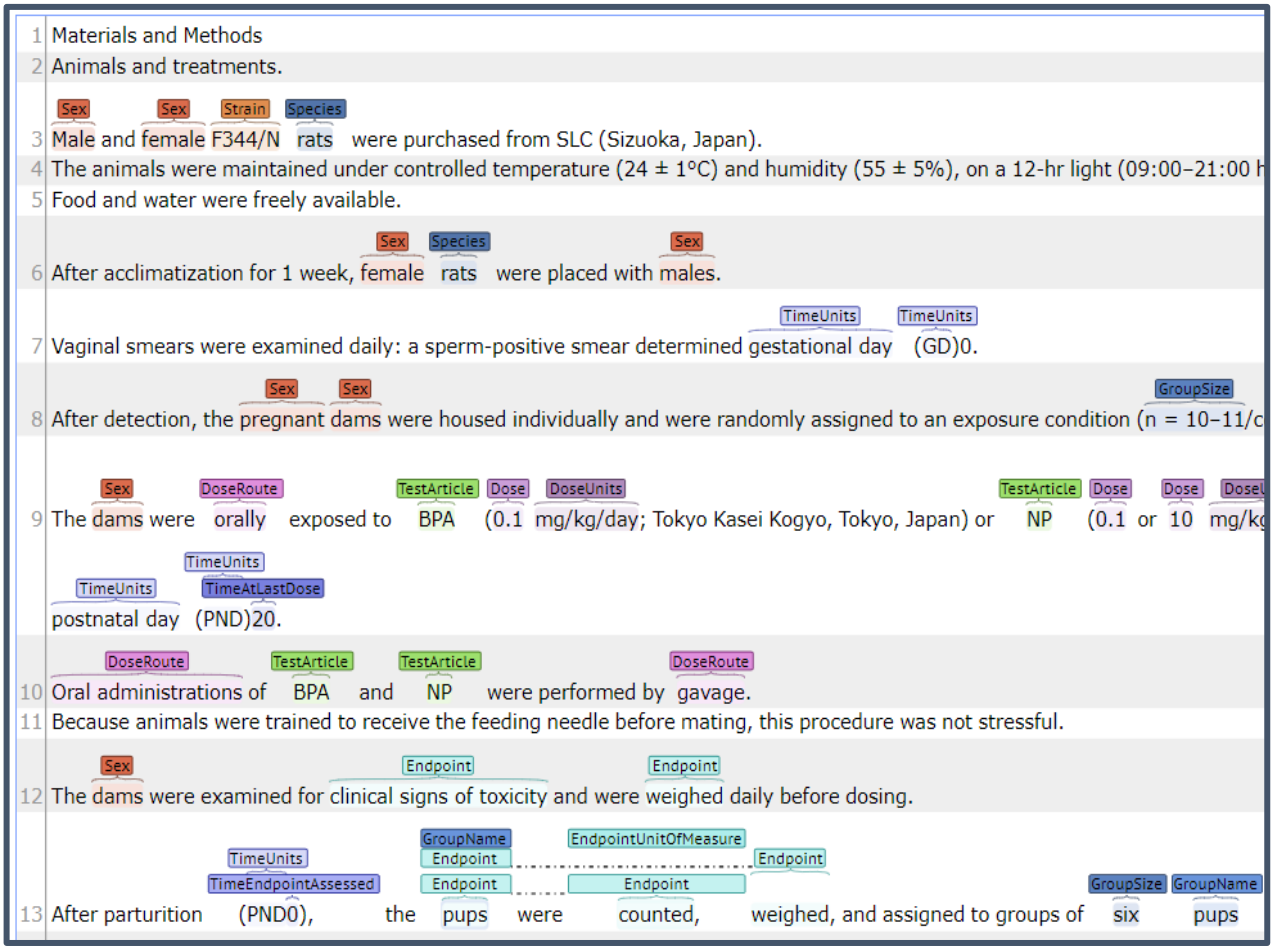

Figure 1: An example of mention annotations

The process for generating final annotations for both the training and test sets was as follows. Three directories were created within BRAT, labeled Annotator1, Annotator2, and QA. Two annotators independently annotated each study, and the second annotator copied their version to the QA directory. A

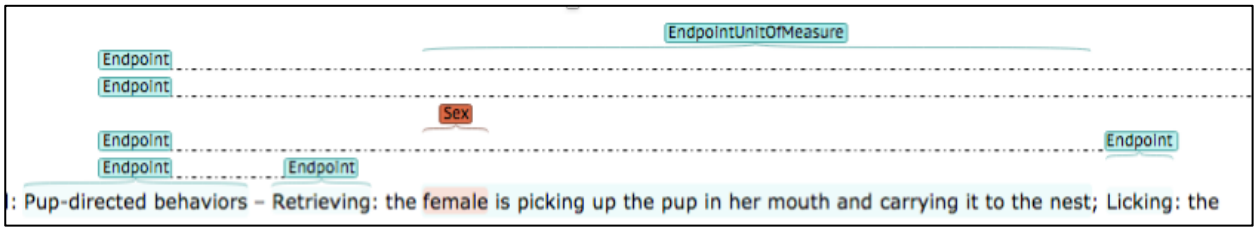

Figure 2: An example of Endpoint and EndpointUnitOfMeasure mentions

third reviewer compared the annotations from Annotator 1 to Annotator 2, resolved any discrepancies, and produced a final QA version by modifying the QA copy as needed. Articles were replaced by an alternative study if the methods section was unclear or focused on in vitro methods. These annotation guidelines were updated periodically when there was a change in guidance or more specific guidance was provided based on questions that arose during the annotation process.

Example Annotations: Figure 1 shows a set of annotations extracted from one of the research articles and illustrates some of the extraction challenges. For instance, the extraction of TestArticle is similar to the task of extracting chemical names (the majority of TestArticles are chemicals); however, a TestArticle must also be given to the test animals for the purpose of evaluating the impact (as opposed to being given to the animal as an anesthetic, for example). For many of the mention types, extraction was highly dependent on the content within the sentence and paragraph. Figure 2 illustrates a set of Endpoint and EndpointUnitOfMeasure mentions, which presented additional challenges as these mention types often included multiple spans that could cross multiple words and included word combinations that were not trivial to recognize from look-up tables, dictionaries, or ontologies.

Figure 3 shows examples of Animal Group and Equivalence Group annotations. In this case, the annotator generated four different animal groups (Animal-0, Animal-1, Animal-2, and Animal-3) and two equivalence groups (Equiv-1, Equiv-2). Mentions labeled with the same group type (e.g., Animal) and same

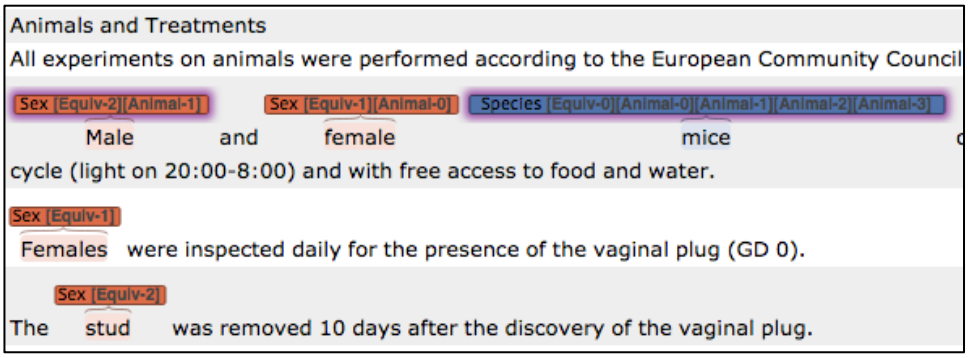

Figure 3: Examples of group annotations

group number (e.g., 0) were assigned to the same group. For instance, in this case, Animal Group 0 consists of the Male and Mice mentions, Animal Group 1 consists of the Female and Mice mentions, and the mention Male is in an equivalence relationship with the mention Stud. To support this type of grouping, the BRAT tool was modified to allow annotators to create, edit, and delete group instances and easily assign a set of mentions to a group.

Table 3 provides the count of annotations types produced for Task 1 and Task 2.

Table 3: Counts of Annotation Types

Table	Training	Test
Type	**Count**	**Count**
CellLine	39	91
Dose	659	611
DoseDuration	216	188
DoseDurationUnits	204	176
DoseFrequency	96	106
DoseRoute	572	524
DoseUnits	493	441
Endpoint	4411	3756
EndpointUnitOfMeasure	706	698
GroupName	963	1058
GroupSize	387	496
SampleSize	45	74
Sex	612	608
Species	1624	1639
Strain	375	338
TestArticle	1922	2207
TestArticlePurity	28	19
TestArticleVerification	6	2
TimeAtDose	117	56
TimeAtFirstDose	47	66
TimeAtLastDose	23	44
TimeEndpointAssessed	672	830
TimeUnits	608	733
Vehicle	440	358
Total	**15265**	**15119**
Group Animal		602
Group Equivalence		1375
Group TestArticle		445
Total		**2422**

Evaluation

Participants submitted results for all of the 444 documents within the test set. Only 100 of the test set documents were annotated; however, participants were not aware of which documents were annotated. The evaluation software (see https://github.com/niehs/systematic_review_eval_nist2018) was available to all participants to use in the development of their models.

For the evaluation of mentions, the evaluation script computed the number of true positives (TP), number of false positives (FP), and number of false negatives (FN) for each paper and each mention type and then computed precision, recall, and F1 measures for each mention type across the 100 test articles and computed the overall scores. To compute per article TPs, FPs, and FNs, the evaluation script computed a distance between each model-generated annotation and each human-provided annotation, where the distance between two annotations was equal to the sum of the overlap between the annotations divided by the total length of annotations. Distances below a threshold, Tmd, were set to 0. The distance matrix created a bipartite graph in which one model annotation may be assigned to more than one human-provided annotation (and vice versa). To create unique assignments, the python linear_sum_assignment algorithm (an implementation of the Hungarian algorithm for unique assignments on bipartite graphs [11]) from the

scipy.optimize library was used. After unique assignments were made, the values for TP, FP, and FN were computed.

The evaluation of groups used the same approach as for the evaluation of mentions. In this case, however, the overlap between two groups was computed as the number of matching mentions divided by the total number of mentions in the two groups. Mentions were considered to be matching if they were determined to be matching during the mention evaluation (so group scores were dependent upon the value for Tmd). As with mentions, distances below a threshold, Tgd, were set to 0. The code for evaluation of groups did not account for equivalence relations, which is a shortcoming to be addressed in future work.

For the challenge, teams were evaluated using Tmd=0.5 and Tgd=0.5. However, to understand the impact of thresholds on results, we present scores for Tmd values ranging from 0.1 to 0.99 (Tgd was also set equal to Tmd).

Participants

Seven participants submitted results for Task 1 and two teams submitted results for Task 2. A brief description of the approaches of teams that submitted reports follows:

- DASCIM: Ecole Polytechnique. For Task 1, the team utilized a multi-level entity detection approach in which qualitative and unit-based mentions were first identified to aid in identification of context specific mentions. Identification rules included string matching based on training set data combined with context-based rules.
- EP: Evidence Prime, LLC. The team developed a deep learning architecture based on a bidirectional Long Short-Term Memory (LSTM) units with highway connections coupled with a Conditional Random Field (CRF). The team made use of novel approaches to connect layers, pre-trained word embeddings using GloVe [15] and ELMo [16], and used various regularization techniques as ways to address the large model parameter space of their neural network-based model.
- ICF: ICF Inc. The ICF team worked with the organizers to provide a baseline model in advance of the evaluation process. This model was used to validate the evaluation software, uncover annotation issues, explore precision versus recall tradeoffs, and develop an expectation of model performance. ICF was instructed to not use external data and to limit model development time and complexity. The team employed a CRF model based on preceding tokens and parts of speech with minimal tuning. For Task 2, the ICF team developed simple, easy-to-implement heuristics to assign nearby mentions into groups. The developed models were submitted to the challenge.
- Sciome: Sciome, LLC. For Task 1, the team employed separate models for different mention types, including linguistic rules and regular expressions for simpler mentions (e.g., sex, species), CRF models for mention types with less than 1000 data points, and LSTM-CRF neural network-based models for remaining mention types. For Task 2, the team used heuristics based on distance between mentions to develop groups.
- VCU: Virginia Commonwealth University. The VCU team used a multi-class classification system for identifying mentions. The system is based upon a Python package, MedaCy, which includes four components, text tokenization, rule-based token grouping, feature extraction, and a CRF model. Of note, the feature extraction included terms from medical terminologies, including UMLS concept mappings along with morphological, orthographic, lexical, synatactic, and semantic features.

Results

Mention Results: Table 4 shows the precision, recall, and F1 scores across all mention types from the seven teams. Each team was allowed to submit up to three models for evaluation. Several teams steered away from deep learning approaches, or limited their use, out of concern over the ability to train such models with the available data. The success of the EP and Sciome teams in using deep learning approaches

demonstrates that transfer learning and regularization techniques can provide competitive models despite modest training set size.

Table 4: Precision, Recall, and F1 Scores for Mentions. FDUKW is a team from Fudan University University in Shanghai, China. AIMRL is a team from the Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan

Team	Model	Precision	Recall	F1
EP	ep_2	58.79	63.17	60.90
EP	ep_3	58.30	63.11	60.61
EP	ep_1	58.65	62.02	60.29
Sciome	sciome_2	53.55	46.46	49.76
Sciome	sciome_1	53.87	44.83	48.94
Sciome	sciome_3	47.87	47.57	47.72
FDUKW	fdukw_1	57.07	40.28	47.23
VCU	vcu_1	48.61	28.27	35.75
ICF	icf_1	20.26	44.68	27.88
DASCIM	dascim_2	28.14	22.57	25.05
DASCIM	dascim_1	23.35	26.09	24.64
ICF	icf_2	13.63	48.46	21.28
ICF	icf_3	10.68	49.79	17.59
AIMRL	aimrl_2	5.36	2.01	2.92
AIMRL	aimrl_1	5.60	1.92	2.86

To provide a reference for evaluating the computational models, we compared the final annotations produced by the third annotator against those produced by the two initial annotators using the evaluation script to generate an inter-annotator human F1 score. Figure 4 shows the top F1 score by mention type for the human annotators (blue) versus the top F1 score (orange) from the submitted models.

Generally, the F1 scores in Figure 4 for the human annotators are in the 80+ range with a few noticeable exceptions. The agreement is lowest for CellLine. Cell Line was originally not included in the mention types as the focus of the data set is on animal studies; however, we noted during the course of annotating studies that many articles contained a combination of animal and molecular studies on cell lines. As such, we made a late decision to capture cell lines and the low F1 score reflects that the guidance was only implemented by the QA annotator for the studies that had already been annotated by Annotators 1 and 2 at the time of the revised guidance. The lower score for TestArticlePurity also likely reflects refinement issues with the annotation guidelines as there were very few instances of TestArticlePurity across studies. Finally, we believe the lower score for EndpointUnitOfMeasure, TimeAtFirstDose, TimeAtLastDose, TimeEndpointAssessed, and TimeUnit (all slightly below 80) reflects the high variability by authors in describing these types.

In looking at the scores from the top models in Figure 4, we see similar trends to those from the inter-annotator scores, namely that time-based scores and EndpointUnitOfMeasure were especially challenging. In addition, the algorithms did especially poorly on SampleSize compared to human annotators. However, the algorithms came close to matching annotator agreement on Sex, Species, CellLine, and TestArticlePurity. The better performance on these mentions likely reflects a combination of fewer unique instances for these mentions as well as less dependence on context.

Figure 5 details the impact of the number of annotations on performance. In this figure, the F1 score of the top algorithm (orange) and the inter-annotator F1 score (blue) are plotted for each mention type against the

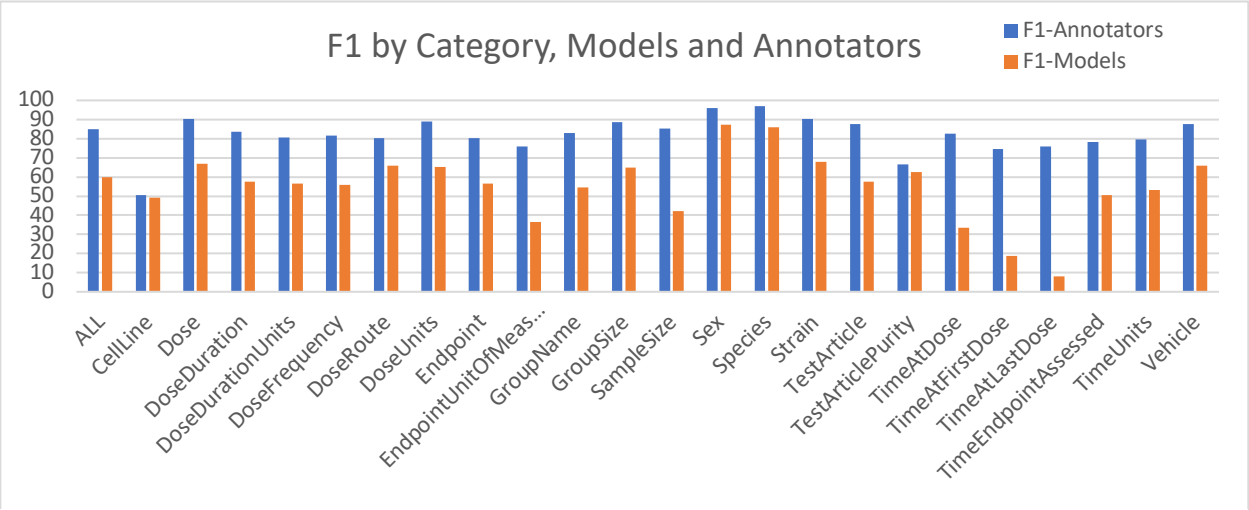

Figure 4: F1 scores by mention type

number of annotations for the mention type. A few trends are worth noting. First, the performance of the models drops dramatically below the inter-annotator score around 50-60 annotation instances. Second, the performance of the models' results clusters around 60 while the inter-annotator scores cluster around 80, despite the number of annotation instances. The two highest data points for the models are Species and Sex, which are generally easier annotations to extract. This result hints at the idea of generating limited annotations and using those to understand the gaps between human and machine results before investing in

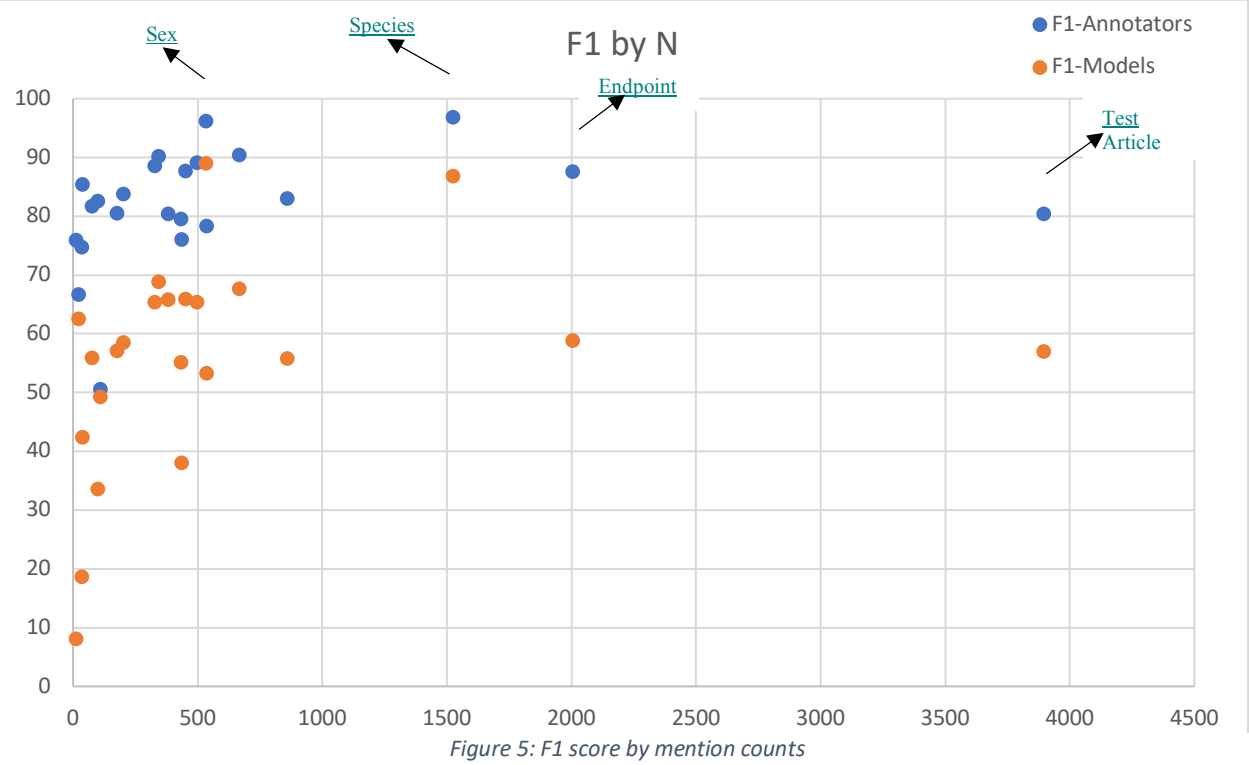

Figure 5: F1 score by mention counts

large scale annotations, especially given that the mentions annotated in this data set were specifically

limited to experimental animal studies and their methods and covered only a portion of the domain where systematic review is employed for research and where training data are needed. The two right-most points in the graph represent Endpoints and TestArticle. Further exploration of whether machine performance would increase with additional annotations is worth pursuing.

Figure 6 highlights the impact of the evaluation threshold on performance. In this figure, the top F1 model scores are presented for each mention type and for each evaluation threshold (for thresholds ranging from 0.1 to 0.99). It is worth noting that the models were trained and evaluated using a threshold of 0.5. In examining the figure, a few mention types are only moderately impacted by threshold (e.g., Dose, Dose Unit), while a few are strongly impacted by the threshold (e.g., Endpoint, Strain, Vehicle). In examining the annotation results, one of the impacts on threshold is the presence of annotations that span multiple words, including words that are separated by a few and in some cases multiple, intervening, non-annotated words. In this case, performance increases when the threshold is low enough that getting only one of the sets of annotated words correct is sufficient to generate a hit call. The other factor is that lower thresholds allow for some amount of tolerance over word boundaries, especially for shorter unit and time-based mention types that may or may not include characters such as parenthesis or unit symbols.

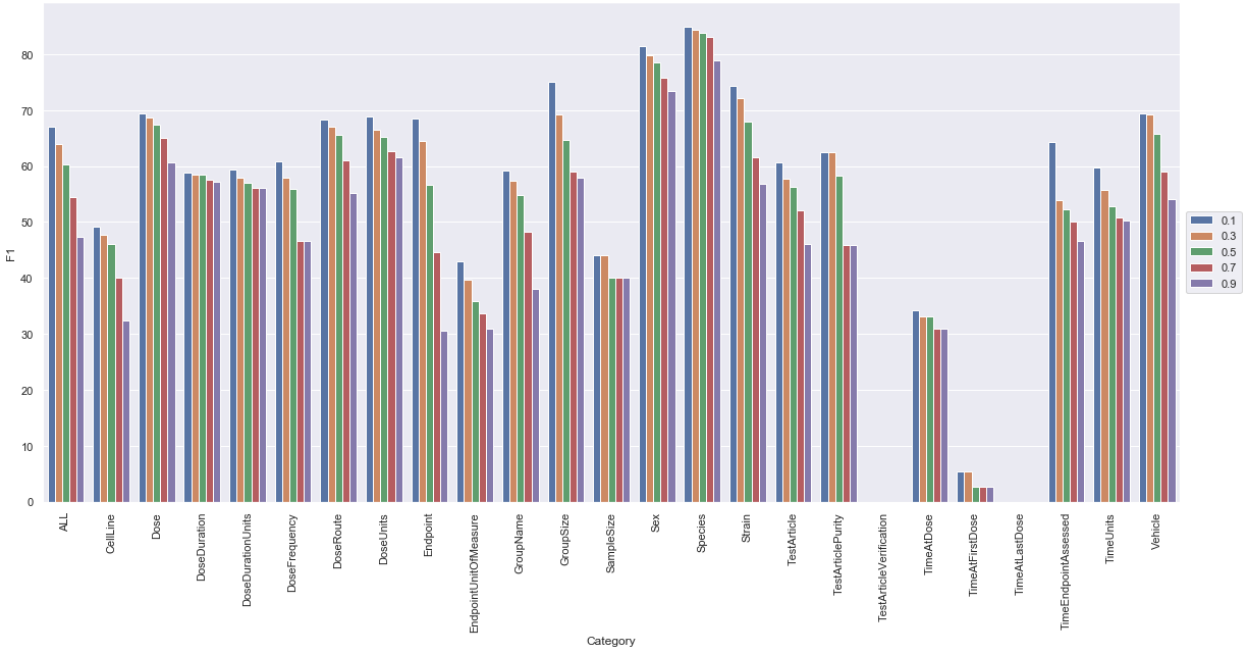

Figure 6: Top F1 score for each mention for varying threshold levels.

Grouping Results: Two teams competed in Task 2 and the results for their models are presented in Table 4. It is important to note that the ICF team was asked by the project team to develop a simple model for Task 2 in order to generate a baseline expectation of performance. In discussion with challenge teams, most teams did not compete on Task 2 due to the limited time provided within the challenge timeline and the dependency of Task 2 on good results from Task 1.

Table 4: Results of Models for Group Annotation

TeamId	Category	Precision	Recall	F1
sciome_1rel	Animal	43.48	46.15	44.78

sciome_1rel	Equiv	31.43	27.54	29.36
icf_2rel	Animal	12.95	31.28	18.32
icf_1rel	Animal	6.17	55.38	11.11
sciome_1rel	TestArticleGroup	9.88	8.56	9.17
icf_2rel	Equiv	3.31	7.57	4.60
icf_2rel	TestArticleGroup	1.48	23.29	2.78
icf_1rel	Equiv	0.76	35.28	1.49
icf_1rel	TestArticleGroup	0.46	39.38	0.90

Discussion

Overall, the results of the challenge were promising. While the gap between machine and human performance on mention extraction is clear, the performance of machine models was not unexpected given the limited time span participants were provided to generate models for an entirely new training set. The work by the various research teams showed that neural network-based deep learning models are performing at a high level. Despite the large number of parameters in such models, the SRIE track results demonstrated that the building of word embeddings and language models from articles outside the training set allows for the generation of data-driven models despite the limited number of annotated research articles. The SRIE results also identified several items for consideration in subsequent challenges and for subsequent/revised training sets, such as including other sections of research articles (e.g., title, abstract, results) and increases in targeted annotations (e.g., for cell lines, time units). Finally, the SRIE track results emphasized that use of computer models is beginning to be viable for inclusion in systematic review data extraction.

Conclusion

The goal of the TAC SRIE track was to stimulate the development of machine-based approaches that can be employed in systematic reviews, particularly in machine-aided extraction of information from research articles by human reviewers. Seven teams submitted models. The top submission for Task 1, the annotation of design features from the methods section of experimental animal studies, achieved results, which although below human-level performance, suggested that computer-assisted IE is a viable option to assist researchers in the labor and resource intensive steps in the systematic review process.

Acknowledgements

Work for this project was provided by staff at the Division of the National Toxicology Program, National Institute for Environmental Health Sciences (NIEHS), National Institutes of Health; ICF International under contract GS00Q14OADU417; and through an interagency agreement of the NIEHS with the National Institute for Standards and Technology (NIST). The organizers would like to thank the ICF annotators and staff, including Marc Avey, Steven Black, Jon Davis, Pam Hartman, Cara Henning, Alexander Lindahl, Mustafa Ramadan, Delaney Reilly, Robert Shin, Kelly Shipkowski, Parnian Soleymani. The organizers would also like to thank Dina Demner Fushman of the National Library of Medicine; Byron Wallace and Ben Nye of Northeastern University for advice and work on the Brat software; NIEHS advisors including Nicole Kleinstreuer, John Bucher, and Andy Shapiro; advisors from the Environmental Protection Agency including Kristan Markey and Michelle Angrish; and technical assistance with setting up the 2018 TAC track provided by NIST staff including Hoa Dang and Ian Soboroff.

References

1. Amini, Imam, David Martinez, and Diego Molla. 2012. “Overview of the ALTA 2012 Shared Task.” *Australasian Language Technology Association Workshop* 7: 124.

2. Basu, Tanmay, Shraman Kumar, Abhishek Kalyan, Priyanka Jayaswal, Pawan Goyal, Stephen Pettifer, and Siddhartha R. Jonnalagadda. 2016. “A Novel Framework to Expedite Systematic Reviews by Automatically Building Information Extraction Training Corpora,” June. http://arxiv.org/abs/1606.06424.
3. Cochrane Collaboration. n.d. “Cochrane Register of Studies.” Accessed August 1, 2018. http://www.metaxis.com/CRSSoftwarePortal/Index.asp.
4. Jonnalagadda, Siddhartha R, Pawan Goyal, and Mark D Huffman. 2015. “Automating Data Extraction in Systematic Reviews: A Systematic Review.” *Systematic Reviews* 4 (1): 78. https://doi.org/10.1186/s13643-015-0066-7.
5. Kim, Su, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. “Automatic Classification of Sentences to Support Evidence Based Medicine.” *BMC Bioinformatics* 12 (Suppl 2): S5. https://doi.org/10.1186/1471-2105-12-S2-S5.
6. Marshall, Iain J, Joël Kuiper, and Byron C Wallace. 2016. “RobotReviewer: Evaluation of a System for Automatically Assessing Bias in Clinical Trials.” *Journal of the American Medical Informatics Association* 23 (1): 193–201. https://doi.org/10.1093/jamia/ocv044.
7. Rooney, Andrew A., Abee L. Boyles, Mary S. Wolfe, John R. Bucher, and Kristina A. Thayer. 2014. “Systematic Review and Evidence Integration for Literature-Based Environmental Health Science Assessments.” *Environmental Health Perspectives* 122 (7): 711–18. https://doi.org/10.1289/ehp.1307972.
8. Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. “BRAT: A Web-Based Tool for NLP-Assisted Text Annotation.” In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–7. EACL ’12. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2380921.2380942.
9. Tsafnat, Guy, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. 2014. “Systematic Review Automation Technologies.” *Systematic Reviews* 3 (1): 74. https://doi.org/10.1186/2046-4053-3-74.
10. Wallace, Byron C, Joël Kuiper, Aakash Sharma, Mingxi Brian Zhu, and Iain J Marshall. 2016. “Extracting PICO Sentences from Clinical Trial Reports Using Supervised Distant Supervision.” *Journal of Machine Learning Research : JMLR* 17. http://www.ncbi.nlm.nih.gov/pubmed/27746703.
11. NTP (National Toxicology Program), 2015c, Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence, Office of Health Assessment and Translation, RTP, NC (2015) (Available: http://ntp.niehs.nih.gov/go/38673 accessed 25 Jan 2019)
12. Thayer, Kristina A., Mary S. Wolfe, Andrew A. Rooney, Abee L. Boyles, John R. Bucher, and Linda S. Birnbaum. 2014. “Intersection of Systematic Review Methodology with the NIH Reproducibility Initiative.” *Environmental Health Perspectives* 122 (7): A176-7. https://doi.org/10.1289/ehp.1408671.
13. “TAC 2018 Systematic Review Information Extraction (SRIE)Track Guidelines.” 2018. https://tac.nist.gov/2018/SRIE/guidelines/SRIE_TrackGuidelines_20180614.pdf. (accessed 02/01/2019).
14. Kuhn, H. W. 1955. “The Hungarian Method for the Assignment Problem.” *Naval Research Logistics Quarterly* 2 (1–2): 83–97. https://doi.org/10.1002/nav.3800020109.
15. Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. “GloVe: Global Vectors for Word Representation.” In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43. http://www.aclweb.org/anthology/D14-1162.
16. Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. “Deep Contextualized Word Representations.” In *Proc. of NAACL*.