# Unsupervised Identification of Study Descriptors in Toxicology Research: An Experimental Study

**Drahomira Herrmannova, Steven R. Young, Robert M. Patton, Christopher G. Stahl**
Oak Ridge National Laboratory, TN, USA
{herrmannovad, youngsr, pattonrm, stahlcg}@ornl.gov

**Nicole C. Kleinstreuer**
NICEATM, NTP, NIEHS, NIH
Research Triangle Park, NC, USA
nicole.kleinstreuer@nih.gov

**Mary S. Wolfe**
NTP, NIEHS, NIH
Research Triangle Park, NC, USA
wolfe@niehs.nih.gov

## Abstract

Identifying and extracting data elements such as study descriptors in publication full texts is a critical yet manual and labor-intensive step required in a number of tasks. In this paper we address the question of identifying data elements in an unsupervised manner. Specifically, provided a set of criteria describing specific study parameters, such as species, route of administration, and dosing regimen, we develop an unsupervised approach to identify text segments (sentences) relevant to the criteria. A binary classifier trained to identify publications that met the criteria performs better when trained on the candidate sentences than when trained on sentences randomly picked from the text, supporting the intuition that our method is able to accurately identify study descriptors.

## Acknowledgments

Support for this research was provided by a grant from the National Institute of Environmental Health Sciences (AES 16002-001), National Institutes of Health to Oak Ridge National Laboratory.

This research was supported in part by an appointment to the Oak Ridge National Laboratory ASTRO Program, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan[1].

## 1 Introduction

Extracting data elements such as study descriptors from publication full texts is an essential step in a number of tasks including systematic review preparation (Jonnalagadda et al., 2015), construction of reference databases (Kleinstreuer et al., 2016), and knowledge discovery (Smalheiser, 2012). These tasks typically involve domain experts identifying relevant literature pertaining to a specific research question or a topic being investigated, identifying passages in the retrieved articles that discuss the sought after information, and extracting structured data from these passages. The extracted data is then analyzed, for example to assess adherence to existing guidelines (Kleinstreuer et al., 2016). Figure 1 shows an example text excerpt with information relevant to a specific task (assessment of adherence to existing guidelines (Kleinstreuer et al., 2016)) highlighted.

[1] http://energy.gov/downloads/doe-public-access-plan



*Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018)*, pages 71–82
Brussels, Belgium, October 31, 2018. 

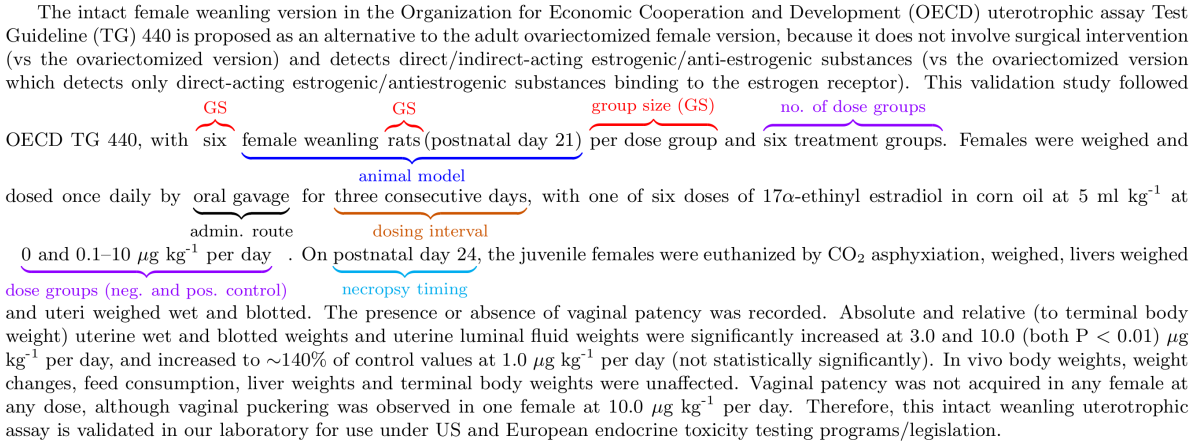


Figure 1: Text excerpt from a reference database of rodent uterotrophic bioassay publications (Kleinstreuer et al., 2016). The text in this example was manually annotated by one of the authors to highlight information relevant to guidelines for performing uterotrophic bioassays set forth by (OECD, 2007).

Extracting the data elements needed in these tasks is a time-consuming and at present a largely manual process which requires domain expertise. For example, in systematic review preparation, information extraction generally constitutes the most time consuming task (Tsafnat et al., 2014). This situation is made worse by the rapidly expanding body of potentially relevant literature with more than one million papers added into PubMed each year (Landhuis, 2016). Therefore, data annotation and extraction presents an important challenge for automation.

A typical approach to automated identification of relevant information in biomedical texts is to infer a prediction model from labeled training data – such a model can then be used to assign predicted labels to new data instances. However, obtaining training data for creating such prediction models can be very costly as it involves the step which these models are trying to automate – manual data extraction. Furthermore, depending on the task at hand, the types of information being extracted may vary significantly. For example, in systematic reviews of randomized controlled trials this information generally includes the *patient* group, the *intervention* being tested, the *comparison*, and the *outcomes* of the study (PICO elements) (Tsafnat et al., 2014). In toxicology research the extraction may focus on routes of exposure, dose, and necropsy timing (Kleinstreuer et al., 2016). Previous work has largely focused on identifying specific pieces of information such as biomedical events (Gonzalez et al., 2015) or PICO elements (Jonnalagadda et al., 2015). However, depending on the domain and the end goal of the extraction, these may be insufficient to comprehensively describe a given study.

Therefore, in this paper we focus on *unsupervised methods* for identifying text segments (such as sentences or fixed length sequences of words) relevant to the information being extracted. We develop a model that can be used to identify text segments from text documents without labeled data and that only requires the current document itself, rather than an entire training corpus linked to the target document. More specifically, we utilize representation learning methods (Mikolov et al., 2013a), where words or phrases are embedded into the same vector space. This allows us to compute semantic relatedness among text fragments, in particular sentences or text segments in a given document and a short description of the type of information being extracted from the document, by using similarity measures in the feature space. The model has the potential to speed up identification of relevant segments in text and therefore to expedite annotation of domain specific information without reliance on costly labeled data.

We have developed and tested our approach on a reference database of rodent uterotropic bioassays[2] (Kleinstreuer et al., 2016) which are labeled according to their adherence to test guidelines set forth in (OECD, 2007). Each study in the database is assigned a label determining whether or not it met each of six main criteria defined by the

[2] https://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/endocrine-disruptors/ref-data/edhts.html

guidelines; however, the database does not contain sentence-level annotations or any information about where the criteria was mentioned in each publication. Due to the lack of fine-grained annotations, supervised learning methods cannot be easily applied to aid annotating new publications or to annotate related but distinct types of studies. This database therefore presents an ideal use-case for unsupervised approaches.

While our approach doesn't require any labeled data to work, we use the labels available in the dataset to evaluate the approach. We train a binary classification model for identifying publications which satisfied given criteria and show the model performs better when trained on relevant sentences identified by our method than when trained on sentences randomly picked from the text. Furthermore, for three out of the six criteria, a model trained solely on the relevant sentences outperforms a model which utilizes full text. The results of our evaluation support the intuition that semantic relatedness to criteria descriptions can help in identifying text sequences discussing sought after information.

There are two main contributions of this work. We present an unsupervised method that employs representation learning to identify text segments from publication full text which are relevant to/contain specific sought after information (such as number of dose groups). In addition, we explore a new dataset which hasn't been previously used in the field of information extraction.

The remainder of this paper is organized as follows. In the following section we provide more details of the task and the dataset used in this study. In Section 3 we describe our approach. In Section 4 we evaluate our model and discuss our results. In Section 5 we compare our work to existing approaches. Finally, in Section 6 we provide ideas for further study.

## 2 The Task and the Data

This section provides more details about the specific task and the dataset used in our study which motivated the development of our model.

### 2.1 Task Description

Significant efforts in toxicology research are being devoted towards developing new *in vitro* methods for testing chemicals due to the large number of untested chemicals in use (>75,000-80,000 (Judson et al., 2009; Kleinstreuer et al., 2016)) and the cost and time required by existing *in vivo* methods (2-3 years and millions of dollars per chemical (Judson et al., 2009)). To facilitate the development of novel *in vitro* methods and assess the adherence to existing study guidelines, a curated database of high-quality *in vivo* rodent uterotrophic bioassay data extracted from research publications has recently been developed and published (Kleinstreuer et al., 2016).

The creation of the database followed the study protocol design set forth in (OECD, 2007), which is composed of six minimum criteria (MC, Table 1). An example of information pertaining to the criteria is shown in Figure 1. Only studies which met all six minimum criteria were considered guideline-like (GL) and were included in a follow-up detailed study and the final database (Kleinstreuer et al., 2016). However, of the 670 publications initially considered for inclusion, only 93 ($\sim$14%) were found to contain studies which met all six MC and could therefore be included in the final database; the remaining 577 publications could not be used in the final reference set. Therefore, significant time and resources could be saved by automating the identification and extraction of the MC.

While each study present in the database is assigned a label for each MC determining whether a given MC was met and the pertinent protocol information was manually extracted, there exist no fine-grained text annotations showing the exact location within each publication's full text where a given criteria was met. Therefore, our goal was to develop a model not requiring detailed text annotations that could be used to expedite the annotation of new publications being added into the database and potentially support the development of new reference databases focusing on different domains and sets of guidelines. Due to the lack of detailed annotations, our focus was on identification of potentially relevant text segments.

### 2.2 The Dataset

The version of the database which contains both GL and non-GL studies consists of 670 publications (spanning the years 1938 through 2014) with results from 2,615 uterotrophic bioassays. Specifically, each entry in the database describes one study, and studies are linked to publications using PubMed reference numbers (PMIDs). Each study

| Criteria name | Description |
|---|---|
| MC 1: Animal model | Immature rats, ovariectomized (OVX) adult rats, or OVX adult mice are acceptable (immature mice are not acceptable). OVX animals: OVX should be performed between six and eight weeks of age (allowing at least 14 days post-surgery before dosing for rats and seven days post-surgery for mice). Immature rats: dosing should begin between postnatal day (PND) 18 and PND 21, and be completed by PND 25. |
| MC 2: Group size | Each control group should have a minimum of three animals and each test group should have a minimum of five animals. |
| MC 3: Route of administration | Acceptable routes of administration: oral gavage (p.o.), subcutaneous (s.c.) injection, or intraperitoneal (i.p.) injection. |
| MC 4: Number of dose groups | Minimum of two dose level groups. Must have positive control and negative control. |
| MC 5: Dosing interval | Dosing for a minimum of three consecutive days. Complete by PND 25 in immature animals. |
| MC 6: Necropsy timing | Should be carried out 18-36 hours after the last dose. |

Table 1: Minimum criteria for guideline-like studies. The descriptions are reprinted here from (Kleinstreuer et al., 2016).

is assigned seven 0/1 labels – one for each of the minimum criteria and one for the overall GL/non-GL label. The database also contains more detailed subcategories for each label (for example "species" label for MC 1) which were not used in this study. The publication PDFs were provided to us by the database creators. We have used the Grobid[3] library to convert the PDF files into structured text. After removing documents with missing PDF files and documents which were not converted successfully, we were left with 624 full text documents.

Each publication contains on average 3.7 studies (separate bioassays), 194 publications contain a single study, while the rest contain two or more studies (with 82 being the most bioassays per publication). The following excerpt shows an example sentence mentioning multiple bioassays (with different study protocols):

> *With the exception of the first study (experiment 1), which had group sizes of 12, all other studies had group sizes of 8.*

For this experiment we did not distinguish between publications describing a single or multiple studies. Instead, our focus was on retrieving all text segments (which may be related to multiple studies) relevant to each of the criteria. For each MC, if a document contained multiple studies with different labels, we discarded that document from our analysis of that criteria; if a document contained multiple studies with the same label, we simply combine all those labels into a single label. Table 2 shows the final size of the dataset.

[3] https://github.com/kermitt2/grobid

| Criteria | 0 | 1 | Total | % of 1 |
|---|---|---|---|---|
| MC 1 | 414 | 175 | 589 | 29.71 |
| MC 2 | 35 | 577 | 612 | 94.28 |
| MC 3 | 70 | 536 | 606 | 88.45 |
| MC 4 | 309 | 206 | 515 | 40.00 |
| MC 5 | 96 | 490 | 586 | 83.62 |
| MC 6 | 228 | 340 | 568 | 59.86 |
| GL | 522 | 72 | 594 | 12.12 |

Table 2: Label statistics. Column *0* shows number of publications per MC which did not meet the criteria and column *1* shows number of publications which met the criteria. The last column in the table shows proportion of positive (i.e. criteria met) labels.

## 3 Approach

In this section we describe the method we have used for retrieving text segments related to the criteria described in the previous section. The intuition is based off question answering systems. We treat the criteria descriptions (Table 1) as the question and the text segments within the publication that discusses the criteria as the answer. Given a

full text publication, the goal is to find the text segments most likely to contain the answer.

We represent the criteria descriptions and text segments extracted from the documents as vectors of features, and utilize relatedness measures to retrieve text segments most similar to the descriptions. A similar step is typically performed by most question answering (QA) systems – in QA systems both the input documents and the question are represented as a sequence of embedding vectors and a retrieval system then compares the document and question representations to retrieve text segments most likely containing the answer (Mishra and Jain, 2016).

To account for the variations in language that can be used to describe the criteria, we represent words as vectors generated using Word2Vec (Mikolov et al., 2013a). The following two excerpts show two different ways MC 6 was described in text:

> *Animals were killed 24 h after being injected and their uteri were removed and weighed.*
>
> *All animals were euthanized by exposure to ethyl ether 24 h after the final treatment.*

We hypothesize that the use of word embedding features will allow us to detect relevant words which are not present in the criteria descriptions. (Mikolov et al., 2013b) have shown that an important feature of Word2Vec embeddings is that similar words will have similar vectors because they appear in similar contexts. We utilize this feature to calculate similarity between the criteria descriptions and text segments (such as sentences) extracted from each document. A high-level overview of our approach is shown in Figure 2.

We use the following method to retrieve the most relevant text segments:

**Segment extraction:** First, we break each document down into shorter sequences such as sentences or word sequences of fixed length. While the first option (sentences) results in text which is easier to process, it has the disadvantage of resulting in sequences of varying length which may affect the resulting similarity value. However, for simplicity, in this study we utilize the sentence version.

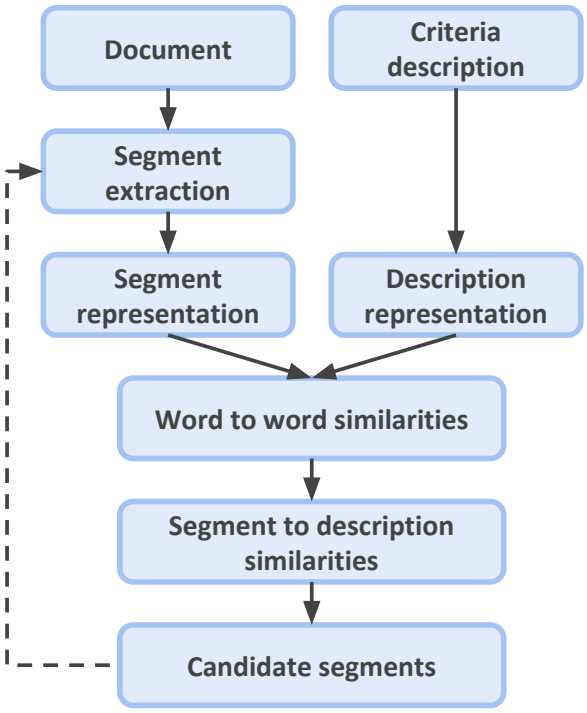


Figure 2: High level overview of our approach. The dotted line represents an optional step of finding smaller sub-segments within the candidate segments. For example, in our case, we first retrieve the most similar sentences and in the second step find most similar continuous 5-grams found withing those sentences.

**Segment/description representation:** We represent each sequence and the input description as a sequence of vector representations. For this study we have utilized Word2Vec embeddings (Mikolov et al., 2013a) trained using the Gensim library on our corpus of 624 full text publications.

**Word to word similarities:** Next we calculate similarity between each word vector from each sequence $s_i$ and each word vector from the input description $d$ using *cosine similarity*. The output of this step is a similarity matrix $\mathbf{S}_i \in \mathbb{R}^{N_i \times M_d}$ for each sequence $s_i$, where $N_i$ is the number of unique words in the sequence and $M_d$ is the number of unique words in the description $d$.

**Segment to description similarities:** To obtain a similarity value representing the relatedness of each sequence to the input description we first convert each input matrix $\mathbf{S}_i$ into a vector $v_i \in \mathbb{R}^{N_i}$ by choosing the maximum similarity value for each word in the sequence, that is $v_i = \max_{rows}(\mathbf{S}_i)$. Each sequence is then assigned a similarity value $r_i \in \mathbb{R}$ which is calculated as $r_i = \text{avg}(v_i)$. In the future we are planning to experiment with different ways of calcu-

lating relatedness of the sequences to the descriptions, such as with computing similarity of embeddings created from the text fragments using approaches like Doc2Vec (Le and Mikolov, 2014). In this study, after finding the top sentences, we further break each sentence down into continuous n-grams to find the specific part of the sentence discussing the MC. We repeat the same process described above to calculate the relatedness of each n-gram to the description.

**Candidate segments:** For each document we select the top $k$ text segments (sentences in the first step and 5-grams in the second step) most similar to the description.

### 3.1 Example Results

Figures 3, 4, and 5 show example annotations generated using our method for the first three criteria. For this example we ran our method on the abstract of the target document rather than the full text and highlighted only the single most similar sentence. The abstract used to produce these figures is the same as the abstract shown in Figure 1. In all three figures, the lighter yellow color highlights the sentence which was found to be the most similar to a given MC description, the darker red color shows the top 5-gram found within the top sentence, and the bold underlined text is the text we are looking for (the correct answer). Annotations generated for the remaining three criteria are shown in Appendix A.

Due to space limitations, Figures 3, 4, and 5 show results generated on abstracts rather than on full text; however, we have observed similarly accurate results when we applied our method to full text. The only difference between the abstracts and the full text version is how many top sentences we retrieved. When working with abstracts only, we observed that if the criteria was discussed in the abstract, it was generally sufficient to retrieve the single most similar sentence. However, as the criteria may be mentioned in multiple places within the document, when working with full text documents we have retrieved and analyzed the top $k$ sentences instead of just a single sentence. In this case we have typically found the correct sentence/sentences among the top 5 sentences. We have also observed that the similar sentences which don't discuss the criteria directly (i.e. the "incorrect" sentences) typically discuss related topics. For example, consider the following three sentences:

> *After weaning on pnd 21, the dams were euthanized by CO2 asphyxiation and the juvenile females were individually housed.*
>
> *Six CD(SD) rat dams, each with reconstituted litters of six female pups, were received from Charles River Laboratories (Raleigh, NC, USA) on offspring postnatal day (pnd) 16.*
>
> *This validation study followed OECD TG 440, with six female weanling rats (postnatal day 21) per dose group and six treatment groups.*

These three sentences were extracted from the abstract and the full text of a single document (document 20981862, the abstract of which is shown in Figures 1 and 3-8). These three sentences were retrieved as the most similar to MC 1, with similarity scores of 70.61, 65.31, and 63.69, respectively. The third sentence contains the "answer" to MC 1 (underlined). However, it can be seen the top two sentences also discuss the animals used in the study (more specifically, the sentences discuss the animals' housing and their origin).

## 4 Evaluation

The goal of this experiment was to explore empirically whether our approach truly identifies mentions of the minimum criteria in text. As we did not have any fine-grained annotations that could be used to directly evaluate whether our model identifies the correct sequences, we have used a different methodology. We have utilized the existing 0/1 labels which were available in the database (these were discussed in Section 2) to train one binary classifier for each MC. The task of each of the classifiers is to determine whether a publication met the given criteria or not. We have then compared a baseline classifier trained on all full text with three other models:

- A model which, instead of all full text, utilized only the top $k$ sentences most similar to the given MC. The top $k$ sentences were identified using our model introduced in the previous section.
- A model which utilized only the $k$ least similar sentences.

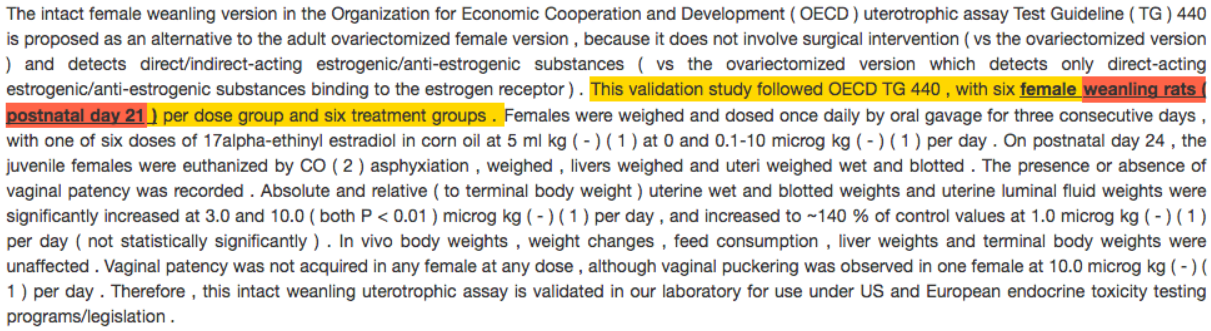


Figure 3: Annotations generated using our method for the abstract from Figure 1. The sentence which was found to be the most similar to the description for "MC 1: Animal model" is highlighted in yellow and the most similar sequence of words within that sentence is highlighted in red. The text we are looking for is highlighted with bold underlined text. For this example we ran our method on the abstract of the target document rather than the full text and highlighted only the single most similar sentence.

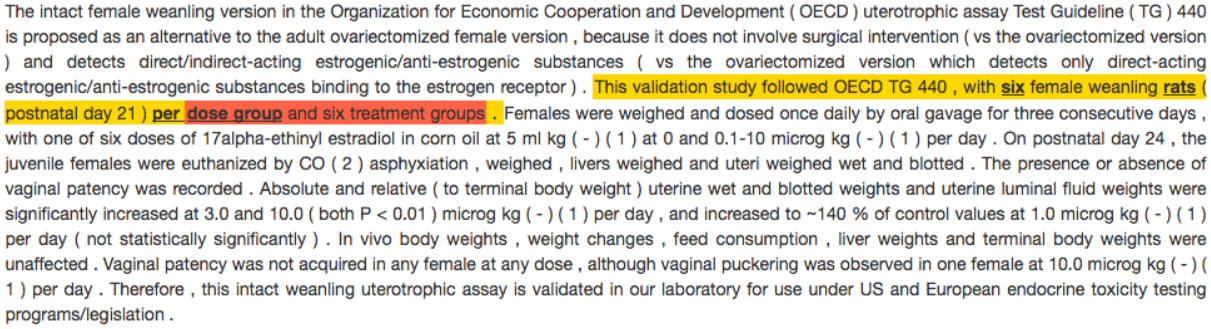


Figure 4: Annotations generated using our method for "MC 2: Group size". The highlighting used is the same as in Figure 3.

- A model which utilized only $k$ random sentences (but none of the top or bottom $k$ sentences – the sentences were chosen at random from the interval $(k, n-k)$ where $n$ is the number of sentences in the document and where sentences are sorted from the most similar to the least similar).

The only difference between the four models is which sentences from each document are passed to the classifier for training and testing. The intuition is that a classifier utilizing the correct sentences should outperform both other models.

To avoid selecting the same sentences across the three models we removed documents which contained less than $3*k$ sentences (Table 3, row *Number of documents* shows how many documents satisfied this condition). In all of the experiments presented in this section, the publication full text was tokenized, lower-cased, stemmed, and stop words were removed. All models used a Bernoulli Naïve Bayes classifier (scikit-learn implementation which used a uniform class prior) trained on binary occurrence matrices created using 1-3-grams extracted from the publications, with n-grams appearing in only one document removed. The complete results obtained from leave-one-out cross validation are shown in Table 3. In all cases we report classification accuracy. In the case of the *random-k sentences* model the accuracy was averaged over 10 runs of the model.

We compare the results to two baselines: (1) a baseline obtained by classifying all documents as belonging to the majority class (*baseline 1* in Table 3) and (2) a baseline obtained using the same setup (features and classification algorithm) as in the case of the *top-/random-/bottom-k sentences* models but which utilized all full text instead of selected sentences extracted from the text only (*baseline 2* in Table 3).

### 4.1 Results analysis

Table 3 shows that for four out of the six criteria (MC 1, MC 4, MC 5, and MC 6) the *top-k sentences* model outperforms *baseline 1* as well the *bottom-k* and the *random-k sentences* models by a significant margin. Furthermore, for three of the

The intact female weanling version in the Organization for Economic Cooperation and Development ( OECD ) uterotrophic assay Test Guideline ( TG ) 440 is proposed as an alternative to the adult ovariectomized female version , because it does not involve surgical intervention ( vs the ovariectomized version ) and detects direct/indirect-acting estrogenic/anti-estrogenic substances ( vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor ) . This validation study followed OECD TG 440 , with six female weanling rats ( postnatal day 21 ) per dose group and six treatment groups . Females were weighed and dosed once daily by **oral gavage** for three consecutive days , with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg ( - ) ( 1 ) at 0 and 0.1-10 microg kg ( - ) ( 1 ) per day . On postnatal day 24 , the juvenile females were euthanized by CO ( 2 ) asphyxiation , weighed , livers weighed and uteri weighed wet and blotted . The presence or absence of vaginal patency was recorded . Absolute and relative ( to terminal body weight ) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 ( both P < 0.01 ) microg kg ( - ) ( 1 ) per day , and increased to ~140 % of control values at 1.0 microg kg ( - ) ( 1 ) per day ( not statistically significantly ) . In vivo body weights , weight changes , feed consumption , liver weights and terminal body weights were unaffected . Vaginal patency was not acquired in any female at any dose , although vaginal puckering was observed in one female at 10.0 microg kg ( - ) ( 1 ) per day . Therefore , this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation .

Figure 5: Annotations generated using our method for "MC 3: Route of administration". The highlighting used is the same as in Figure 3.

| **Approach** | **MC1** | **MC2** | **MC3** | **MC4** | **MC5** | **MC6** |
|---|---|---|---|---|---|---|
| Baseline 1: Most frequent label | 70.35 | 94.43 | 88.74 | 59.48 | 84.30 | 60.44 |
| Baseline 2: All full text | 78.25 | 92.06 | 89.59 | 67.94 | 84.83 | 74.05 |
| Top-k sentence | 76.84 | 91.55 | 87.71 | 68.35 | 88.54 | 74.23 |
| Bottom-k sentences | 70.00 | 91.39 | 88.23 | 63.10 | 80.60 | 63.70 |
| Random-k sentences | 73.26 | 93.72 | 88.43 | 65.65 | 85.29 | 68.28 |
| Number of documents | 570 | 592 | 586 | 496 | 567 | 551 |
| Number of pos. labels | 169 | 559 | 520 | 201 | 478 | 333 |

Table 3: Evaluation results.

six criteria (MC 4, MC 5, and MC 6) the *top-k sentences* model also outperforms the *baseline 2* model (model which utilized all full text). This seems to confirm our hypothesis that semantic relatedness of sentences to the criteria descriptions helps in identifying sentences discussing the criteria. These seems to be the case especially given that for three of the six criteria the *top-k sentences* model outperforms the model which utilizes all full text (*baseline 2*) despite being given less information to learn from (selected sentences only in the case of the *top-k sentences* model vs. all full text in the case of the *baseline 2* model).

For two of the criteria (MC 2 and MC 3) this is not the case and the *top-k sentences* model performs worse than both other models in the case of MC 3 and worse than the *random-k* model in the case of MC 2. One possible explanation for this is class imbalance. In the case of MC 2, only 33 out of 592 publications (5.57%) represent negative examples (Table 3). As the *top-k sentences* model picks only sentences closely related to MC 2, it is possible that due to the class imbalance the top sentences don't contain enough negative examples to learn from. On the other hand, the *bottom-k* and *random-k sentences* models may select text not necessarily related to the criteria but potentially containing linguistic patterns which the model learns to associate with the criteria; for example, certain chemicals may require the use of a certain study protocol which may not be aligned with the MC and the model may key in on the appearance of these chemicals in text rather than the appearance of MC indicators. The situation is similar in the case of MC 3. We would like to emphasize that the goal of this experiment was not to achieve state-of-the-art results but to investigate empirically the viability of utilizing semantic relatedness of text segments to criteria descriptions for identifying relevant segments.

## 5 Related Work

In this section we present studies most similar to our work. We focus on unsupervised methods for information extraction from biomedical texts.

Many methods for biomedical data annotation and extraction exist which utilize labeled data and supervised learning approaches ((Liu et al., 2016) and (Gonzalez et al., 2015) provided a good overview of a number of these methods); however, unsupervised approaches in this area are much scarcer. One such approach has been introduced by (Zhang and Elhadad, 2013), who have proposed a model for unsupervised Named En-

tity Recognition. Similar to our approach, their model is based on calculating the similarity between vector representations of candidate phrases and existing entities. However, their vector representations are created using a combination of TF-IDF weights and word context information, and their method relies on a terminology. More recently, (Chen and Sokolova, 2018) have utilized Word2Vec and Doc2Vec embeddings for unsupervised sentiment classification in medical discharge summaries.

A number of previous studies have focused on unsupervised extraction of relations such as protein-protein interactions (PPI) from biomedical texts. For example, (Quan et al., 2014) have utilized several techniques, namely kernel-based pattern clustering and dependency parsing, to extract PPI from biomedical texts. (Alicante et al., 2016) have introduced a system for unsupervised extraction of entities and relations between these entities from clinical texts written in Italian, which utilized a thesaurus for extraction of entities and clustering methods for relation extraction. (Rink and Harabagiu, 2011) also used clinical texts and proposed a generative model for unsupervised relation extraction. Another approach focusing on relation extraction has been proposed by (Madkour et al., 2007). Their approach is based on constructing a graph which is used to construct domain-independent patterns for extracting protein-protein interactions.

A similar but distinct approach to unsupervised extraction is distant supervision. Similarly as unsupervised extraction methods, distant supervision methods don't require any labeled data, but make use of weakly labeled data, such as data extracted from a knowledge base. Distant supervision has been applied to relation extraction (Liu et al., 2014), extraction of gene interactions (Mallory et al., 2015), PPI extraction (Thomas et al., 2012; Bobić et al., 2012), and identification of PICO elements (Wallace et al., 2016). The advantage of our approach compared to the distantly supervised methods is that it does not require any underlying knowledge base or a similar source of data.

## 6 Conclusions and Future Work

In this paper we presented a method for unsupervised identification of text segments relevant to specific sought after information being extracted from scientific documents. Our method is entirely unsupervised and only requires the current document itself and the input descriptions instead of corpus linked to this document. The method utilizes short descriptions of the information being extracted from the documents and the ability of word embeddings to capture word context. Consequently, it is domain independent and can potentially be applied to another set of documents and criteria with minimal effort. We have used the method on a corpus of toxicology documents and a set of guideline protocol criteria needed to be extracted from the documents. We have shown the identified text segments are very accurate. Furthermore, a binary classifier trained to identify publications that met the criteria performed better when trained on the candidate sentences than when trained on sentences randomly picked from the text, supporting our intuition that our method is able to accurately identify relevant text segments from full text documents.

There are a number of things we plan on investigating next. In our initial experiment we have utilized criteria descriptions which were not designed to be used by our model. One possible improvement of our method could be replacing the current descriptions with example sentences taken from the documents containing the sought after information. We also plan on testing our method on an annotated dataset, for example using existing annotated PICO element datasets (Boudin et al., 2010).

## References

Anita Alicante, Anna Corazza, Francesco Isgrò, and Stefano Silvestri. 2016. Unsupervised entity and relation extraction from clinical records in italian. *Computers in biology and medicine*, 72:263–275.

Tamara Bobić, Roman Klinger, Philippe Thomas, and Martin Hofmann-Apitius. 2012. Improving distantly supervised extraction of drug-drug and protein-protein interactions. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 35–43. Association for Computational Linguistics.

Florian Boudin, Jian-Yun Nie, Joan C Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. 2010. Combining classifiers for robust pico element detection. *BMC medical informatics and decision making*, 10(1):29.

Qufei Chen and Marina Sokolova. 2018. Word2vec and doc2vec in unsupervised sentiment analysis

of clinical discharge summaries. *arXiv preprint arXiv:1805.00352*.

Graciela H Gonzalez, Tasnia Tahsin, Britton C Goodale, Anna C Greene, and Casey S Greene. 2015. Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in bioinformatics*, 17(1):33–42.

Siddhartha R. Jonnalagadda, Pawan Goyal, and Mark D. Huffman. 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews*, 4(1).

Richard Judson, Ann Richard, David J Dix, Keith Houck, Matthew Martin, Robert Kavlock, Vicki Dellarco, Tala Henry, Todd Holderman, Philip Sayre, et al. 2009. The toxicity data landscape for environmental chemicals. *Environmental health perspectives*, 117(5):685.

Nicole C. Kleinstreuer, Patricia C. Ceger, David G. Allen, Judy Strickland, Xiaoqing Chang, Jonathan T. Hamm, and Warren M. Casey. 2016. A curated database of rodent uterotrophic bioactivity. *Environmental Health Perspectives*, 124(5).

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Feifan Liu, Jinying Chen, Abhyuday Jagannatha, and Hong Yu. 2016. Learning for biomedical information extraction: Methodological review of recent advances. *arXiv preprint arXiv:1606.07993*.

Mengwen Liu, Yuan Ling, Yuan An, and Xiaohua Hu. 2014. Relation extraction from biomedical literature with minimal supervision and grouping strategy. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 444–449. IEEE.

Amgad Madkour, Kareem Darwish, Hany Hassan, Ahmed Hassan, and Ossama Emam. 2007. Bionoculars: extracting protein-protein interactions from biomedical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 89–96. Association for Computational Linguistics.

Emily K Mallory, Ce Zhang, Christopher Ré, and Russ B Altman. 2015. Large-scale extraction of gene interactions from full-text literature using deepdive. *Bioinformatics*, 32(1):106–113.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR Workshop*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Amit Mishra and Sanjay Kumar Jain. 2016. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361.

OECD. 2007. Test No. 440: Uterotrophic Bioassay in Rodents. In *OECD Guidelines for the Testing of Chemicals, Section 4*. OECD Publishing, Paris.

Changqin Quan, Meng Wang, and Fuji Ren. 2014. An unsupervised text mining method for relation extraction from biomedical literature. *PloS one*, 9(7):e102039.

Bryan Rink and Sanda Harabagiu. 2011. A generative model for unsupervised discovery of relations and argument classes from clinical texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 519–528. Association for Computational Linguistics.

Neil R. Smalheiser. 2012. Literature-based discovery: Beyond the ABCs. *Journal of the American Society for Information Science and Technology*, 63(2):218–224.

Philippe Thomas, Tamara Bobić, Ulf Leser, Martin Hofmann-Apitius, and Roman Klinger. 2012. Weakly labeled corpora as silver standard for drug-drug and protein-protein interaction. In *Proceedings of the Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM) on Language Resources and Evaluation Conference (LREC)*.

Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. 2014. Systematic review automation technologies. *Systematic reviews*, 3(1):74.

Byron C Wallace, Joel Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J Marshall. 2016. Extracting pico sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*, 17(1):4572–4596.

Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098.

## A Supplemental Material

This section provides additional details and results. Figures 6, 7, and 8 show example annotations generated for criteria MC 4, MC 5, and MC 6.

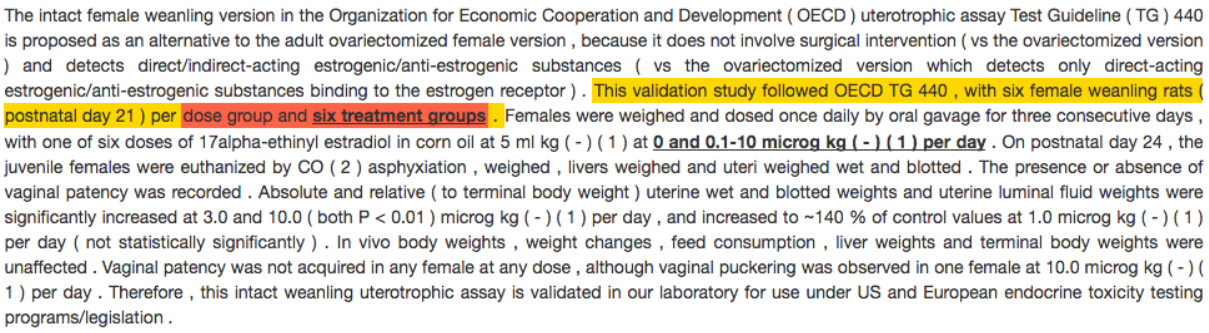


Figure 6: Annotations generated using our method for abstract from Figure 1. The sentence which was found to be the most similar to the description for "MC 4: Number of dose groups" is highlighted in yellow and the most similar sequence of words within that sentence is highlighted in red. The text we are looking for is highlighted with bold underlined text. For this example we ran our method on the abstract of the target document rather than the full text and highlighted only the single most similar sentence.

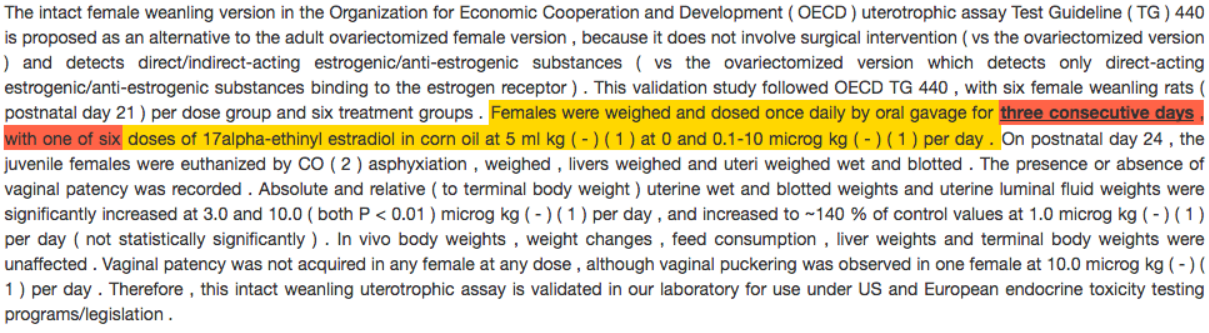


Figure 7: Annotations generated using our method for "MC 5: Dosing interval".

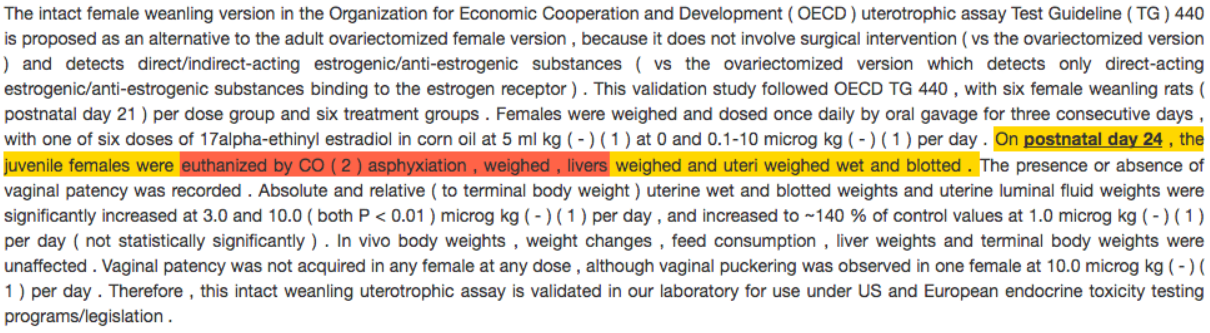


Figure 8: Annotations generated using our method for "MC 6: Necropsy timing".