PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts

Franck Dernoncourt*
Adobe Research
dernonco@adobe.com

Ji Young Lee*
MIT
jjylee@mit.edu

Abstract

We present PubMed 200k RCT[1], a new dataset based on PubMed for sequential sentence classification. The dataset consists of approximately 200,000 abstracts of randomized controlled trials, totaling 2.3 million sentences. Each sentence of each abstract is labeled with their role in the abstract using one of the following classes: background, objective, method, result, or conclusion. The purpose of releasing this dataset is twofold. First, the majority of datasets for sequential short-text classification (i.e., classification of short texts that appear in sequences) are small: we hope that releasing a new large dataset will help develop more accurate algorithms for this task. Second, from an application perspective, researchers need better tools to efficiently skim through the literature. Automatically classifying each sentence in an abstract would help researchers read abstracts more efficiently, especially in fields where abstracts may be long, such as the medical field.

1 Introduction

Short-text classification is an important task in many areas of natural language processing, such as sentiment analysis, question answering, or dialog management. For example, in a dialog management system, one might want to classify each utterance into dialog acts (Stolcke et al., 2000).

In the dataset we present in this paper, PubMed 200k RCT, each short text we consider is one sentence. We focus on classifying sentences in medical abstracts, and particularly in randomized controlled trials (RCTs), as they are commonly considered to be the best source of medical evidence (Tianjing Li, 2015). Since sentences in an abstract appear in a sequence, we call this task the *sequential sentence classification* task, in order to distinguish it from general text or sentence classification that does not have any context.

The number of RCTs published every year is steadily increasing, as Figure 1 illustrates. Over 1 million RCTs have been published so far and around half of them are in PubMed (Mavergames, 2013), which makes it challenging for medical investigators to pinpoint the information they are looking for. When researchers search for previous literature, e.g., to write systematic reviews, they often skim through abstracts in order to quickly check whether the papers match the criteria of interest. This process is easier when abstracts are *structured*, i.e., the text in an abstract is divided into semantic headings such as objective, method, result, and conclusion. However, over half of published RCT abstracts are *unstructured*, as shown in Figure 2, which makes it more difficult to quickly access the information of interest.

Consequently, classifying each sentence of an abstract to an appropriate heading can significantly reduce time to locate the desired information, as Figure 3 illustrates. Besides assisting humans, this task may also be useful for a variety of downstream applications such as automatic text summarization, information extraction, and information retrieval. In addition to the medical applications, we hope that the release of this dataset will help the development of algorithms for sequential sentence classification.

* These authors contributed equally to this work.
[1] The dataset is freely available at https://github.com/Franck-Dernoncourt/pubmed-rct

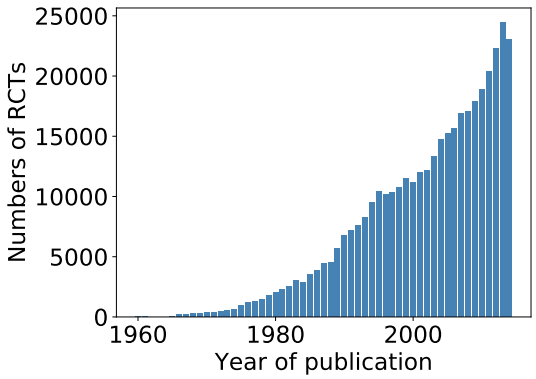

Figure 1: Number of RCTs present in PubMed published yearly between 1960 and 2014 (inclusive). The first documented controlled trial dates back 1747 (Dunn, 1997), but the scientific value of RCTs became widely recognized only by the late 20th century as the standard method for medical evidence (Meldrum, 2000).

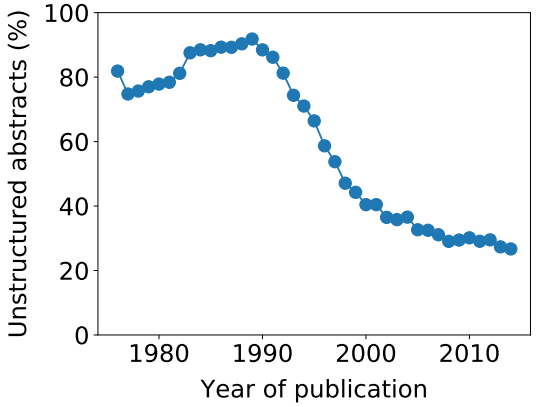

Figure 2: Evolution of the percentage of RCT abstracts present in PubMed that are unstructured between 1975 and 2014 (inclusive). The years before 1975 were omitted due to the low number of RCTs. Overall, approximately half of the RCT abstracts are unstructured. An RCT abstract is considered as unstructured if and only if at least one of its section is labeled as "None".

2 Related Work

Existing datasets for classifying sentences in medical abstracts are either small, not publicly available, or do not focus on RCTs. Table 1 presents an overview of existing datasets.

The most studied dataset to our knowledge is the NICTA-PIBOSO corpus published by Kim et al. (2011). This dataset was the basis of the ALTA 2012 Shared Task (Amini et al., 2012), in which 8 competing research teams participated.

Achilles tendinopathy (AT) is a common and difficult to treat musculoskeletal disorder. The purpose of this study is to examine whether 1 injection of platelet-rich plasma (PRP) would improve outcomes more effectively than placebo (saline) after 3 months when used to treat AT. A total of 24 male patients with chronic AT (median disease duration, 33 months) were randomized (1:1) to receive either a blinded injection of PRP (n = 12) or saline (n = 12). Patients were informed that they could drop out after 3 months if they were dissatisfied with the treatment. After 3 months, all patients were reassessed (no dropouts). No difference between the PRP and the saline group could be observed with regard to the primary outcome (VISA-A score: mean difference [MD], -1.3; 95% CI, -17.8 to 15.2; P = .868). Secondary outcomes were pain at rest (MD, 1.6; 95% CI, -0.5 to 3.7; P = .137), pain while walking (MD, 0.8; 95% CI, -1.8 to 3.3; P = .544), pain when tendon was squeezed (MD, 0.3; 95% CI, -0.2 to 0.9; P = .208). PRP injection did not result in an improved VISA-A score over a 3-month period compared with placebo. The conclusions are limited to the 3 months after treatment owing to the large dropout rate.

Figure 3: Example of abstract with the method section highlighted. Abstracts in the medical field can be long. This abstract was taken from (Krogh et al., 2016) and several sentences have been removed for the sake of conciseness. Providing clinical researchers and practitioners a tool that would allow them to highlight the section(s) that they are interested in would help them explore the literature more efficiently.

Only the dataset published in (Davis-Desmond and Mollá, 2012) is publicly available: two datasets can only be obtained via email inquiries, and the other datasets are not accessible (unanswered email requests or negative replies). The only public dataset is also the smallest one.

3 Dataset Construction

3.1 Abstract Selection

Our dataset is constructed upon the MEDLINE/PubMed Baseline Database published in 2016 , which we will refer to as PubMed in this paper. PubMed can be accessed online by anyone, free of charge and without having to go through any registration. It contains 24,358,442 records. A record typically consists of metadata on one article, as well as the article's title and in many cases its abstract.

We use the following information from each PubMed record of an article to build our dataset: the PubMed ID (PMID), the abstract and its structure if available, and the Medical Subject Head-

Dataset	Size	Manual	RCT	Available
Hara et al. (2007)	200	y	y	email
Hirohata et al. (2008)	104k	n	n	no
Chung (2009)	327	y	y	no
Boudin et al. (2010)	29k	n	n	no
Kim et al. (2011)	1k	y	n	email
Huang et al. (2011)	23k	n	n	no
Robinson (2012)	1k	n	y	no
Zhao et al. (2012)	20k	y	n	no
Davis et al. (2012)	194	n	y	public
Huang et al. (2013)	20k	n	y	no
PubMed 200k RCT	196k	n	y	no

Table 1: Overview of existing datasets for sentence classification in medical abstracts. The size is expressed in terms of number of abstracts.

ings (MeSH) terms. MeSH is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed.

We select abstracts from PubMed based on the two following criteria:

- the abstract must belong to an RCT. We rely on the article's MeSH terms only to select RCTs. Specifically, only the articles with the MeSH term D016449, which corresponds to an RCT, are included in our dataset. 399,254 abstracts fit this criterion.

- the abstract must be structured. In order to qualify as structured, it has to contain between 3 and 9 sections (inclusive), and it should not contain any section labeled as "None", "Unassigned", or "" (empty string). Only 0.5% of abstracts have fewer than 3 sections or more than 9 sections: we chose to discard these outliers. The label of each section was originally given by the authors of the articles, typically following the guidelines given by journals: as many labels exist, PubMed maps them into a smaller set of standardized labels: background, objective, methods, results, conclusions, "None", "Unassigned", or "" (empty string).

195,654 abstracts fit these two criteria, i.e., belong to RCTs and are structured.

3.2 Dataset Split

The dataset contains 195,654 abstracts and is randomly split into three sets: a validation set containing 2500 abstracts, a test set containing 2500 abstracts, and a training set containing the remaining 190,654 abstracts. Since 200k abstracts may be too many for some applications, we also provide a smaller dataset, PubMed 20k RCT, which contains 15000 abstracts for the training set, 2500 abstracts for the validation set, and 2500 abstracts for the test set. The 20k abstracts were chosen from the 200k abstracts by taking the most recently published ones. Table 2 presents the number of abstracts and sentences for both PubMed 20k RCT and PubMed 200k RCT, for each split of the data set.

Dataset	$\|V\|$	Train	Validation	Test
PubMed 20k	68k	15k (180k)	2.5k (30k)	2.5k (30k)
PubMed 200k	331k	190k (2.2M)	2.5k (29k)	200 (29k)

Table 2: Dataset overview. $|V|$ denotes the vocabulary size. For the train, validation and test sets, we indicate the number of abstracts followed by the number of sentences in parentheses.

3.3 Dataset Format

The dataset is provided as three text files: one for the training set, one for the validation set, and one for the test set. Each file has the same format: each line corresponds to either a PMID or a sentence with its capitalized label at the beginning. Each token is separated by a space. Listing 1 shows an excerpt from these files.

For each abstract, sentence and token boundaries are detected using the Stanford CoreNLP toolkit (Manning et al., 2014). We provide two versions of the dataset: one with the original text, and one where digits are replaced by the character @ (at sign).

```
###9813759
OBJECTIVE This study evaluated an  [...]
OBJECTIVE It was hypothesized that [...]
METHODS Participants were @ men    [...]
METHODS Psychological functioning  [...]
RESULTS Intervention group subject [...]
RESULTS Compared to the control    [...]
CONCLUSIONS This study has shown   [...]
```

Listing 1: Example of one abstract as formatted in the PubMed 200k RCT dataset set. The PMID of the corresponding article is 9813759; the article can be found that https://www.ncbi.nlm.nih.gov/pubmed/9813759.

4 Dataset Analysis

Figure 4 counts the number of sentences per label: the least common label (objective) is approximately four times less frequent than the most common label (results), which indicates that the dataset is not excessively unbalanced. Figure 5 shows the distribution of the number of tokens the sentence. Figure 6 shows the distribution of the number of sentences per abstract. Figures 4, 5 and 6 are based on PubMed 200k RCT.

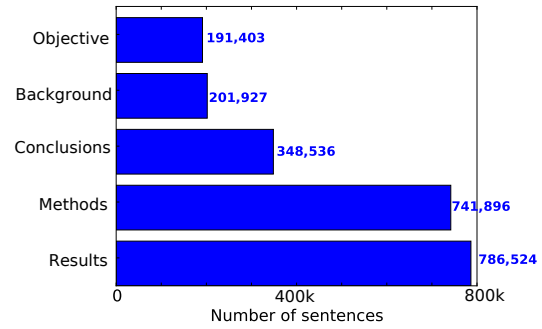

Figure 4: Number of sentences per label

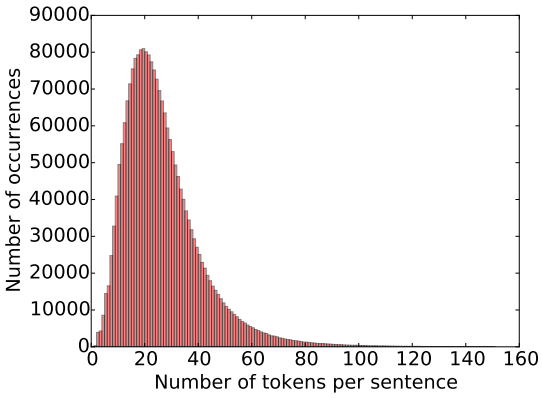

Figure 5: Distribution of the number of tokens the sentence. Minimum: 1; mean: 26.2; maximum: 338; variance: 227.6; skewness: 2.0; kurtosis: 8.7.

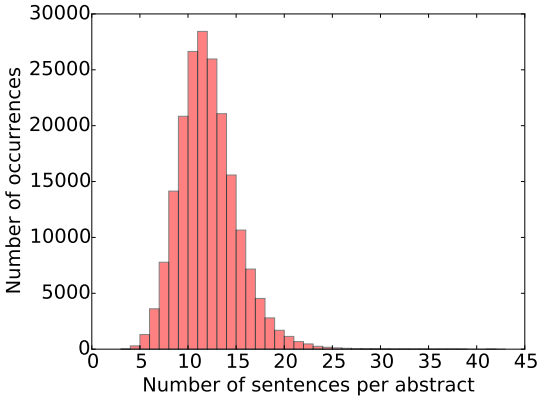

Figure 6: Distribution of the number of sentences per abstract. Minimum: 3; mean: 11.6; maximum: 51; variance: 9.5; skewness: 0.9; kurtosis: 2.6.

5 Performance Benchmarks

We report the performance of several systems to characterize our dataset. The first baseline is a classifier based on logistic regression (LR) using n-gram features extracted from the current sentence: it does not use any information from the surrounding sentences. This baseline was implemented with scikit-learn (Pedregosa et al., 2011).

The second baseline (Forward ANN) uses the artificial neural network (ANN) model presented in (Lee and Dernoncourt, 2016): it computes sentence embeddings for each sentence, then classifies the current sentence given a few preceding sentence embeddings as well as the current sentence embedding.

The third baseline is a conditional random field (CRF) that uses n-grams as features: each output variable of the CRF corresponds to a label for a sentence, and the sequence the CRF considers is the entire abstract. The CRF baseline therefore uses both preceding and succeeding sentences when classifying the current sentence. CRFs have been shown to give strong performances for sequential sentence classification (Amini et al., 2012). This baseline was implemented with CRFsuite (Okazaki, 2007).

The fourth baseline (bi-ANN) is an ANN consisting of three components: a token embedding layer (bi-LSTM), a sentence label prediction layer (bi-LSTM), and a label sequence optimization layer (CRF). The architecture is described in (Dernoncourt et al., 2016) and has been demonstrated to yield state-of-the-art results for sequential sentence classification.

Table 3 compares the four baselines. As expected, LR performs the worst, followed by the Forward ANN. The bi-ANN outperforms the CRF, but as the data set becomes larger the difference of performances diminishes.

Table 4 presents the precision, recall, F1-score and support for each class with the bi-ANN. Accurately classifying the background and objective classes is the most challenging. The confusion matrix in Table 5 shows that background sentences are often confused with objective sentences, and vice versa.

Table 6 gives more details on the LR baseline, and illustrates the impact of the choice of the n-gram size on the performance. By the same token, Table 7 shows the impact of the choice of the window size on the performance of the CRF.

Model	PubMed 20k	PubMed 200k
LR	83.1	85.9
Forward ANN	86.1	88.4
CRF	89.5	91.5
bi-ANN	**90.0**	**91.6**

Table 3: F1-scores on the test set of several baselines. The presented results for the ANN-based models are the F1-scores on the test set of the run with the highest F1-score on the validation set.

	Precision	Recall	F1-score	Support
Background	70.7	81.1	75.6	2663
Conclusions	94.6	93.7	94.2	4426
Methods	95.5	96.5	96.0	9751
Objective	77.1	65.3	70.7	2377
Results	95.6	94.8	95.2	10276
Total	91.7	91.6	91.6	29493

Table 4: Results for each class obtained by the bi-ANN model on the PubMed 200k RCT test set. The total support is 29493, i.e. the number of sentences in the test set.

	Backg.	Concl.	Methods	Obj.	Res.
Background	2760	12	62	424	5
Conclusions	41	4149	9	0	227
Methods	82	17	9409	31	212
Objective	757	0	69	1551	0
Results	14	208	303	5	9746

Table 5: Confusion matrix on the PubMed 200k RCT test set obtained with the bi-ANN model. Rows correspond to actual labels, and columns correspond to predicted labels. For example, 62 background sentences were predicted as method.

6 Conclusion

In this article we have presented PubMed 200k RCT, a dataset for sequential sentence classification. It is the largest such dataset that we are aware of. We have evaluated the performance of several baselines so that researchers may directly compare their algorithms against them without having to develop their own baselines. We hope that the release of this dataset will accelerate the development of algorithms for sequential sentence classification and increase the interest of the text mining community in the study of RCTs.

N-gram size	Precision	Recall	F1-score	Runtime
1	82.3	82.7	82.4	4406
2	85.1	85.4	85.2	13237
3	85.5	85.8	85.6	20618
4	85.7	86.0	85.8	25553
5	**85.8**	**86.1**	**85.9**	35006

Table 6: Results obtained on the PubMed 200k RCT test set by the LR model with different size of n-grams as features. The n-gram size indicates the size of the largest n-grams: For example, if the n-gram size is 3, it means unigrams, bigrams and trigrams are extracted as features. The maximum n-gram size in our experiments is 5 due to RAM limitation. The runtime is expressed in seconds and comprises both training and testing times.

Window size	Precision	Recall	F1-score	Runtime
1	90.6	90.6	90.6	1565
2	91.0	91.0	91.0	2490
3	91.1	91.1	91.1	3908
4	**91.5**	**91.5**	**91.5**	4867
5	90.9	91.0	90.9	6424
6	91.4	91.4	91.4	7649
7	91.3	91.3	91.3	7929
8	90.9	90.9	90.9	7644
9	91.2	91.3	91.2	7891

Table 7: Results obtained on the PubMed 200k RCT test set by the CRF model with different window sizes. A window of size k means that for each token, features are extracted from the current token, the k preceding tokens as well as the k succeeding tokens. The runtime is expressed in seconds and comprises both training and testing times.

References

Iman Amini, David Martinez, and Diego Molla. 2012. Overview of the ALTA 2012 Shared Task. In *Australasian Language Technology Association Workshop 2012*. volume 7, page 124.

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.

Grace Yuet-Chee Chung. 2009. Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. *Journal of biomedical informatics* 42(5):790–800.

Patrick Davis-Desmond and Diego Mollá. 2012. Detection of evidence in clinical research papers. In

Proceedings of the Fifth Australasian Workshop on Health Informatics and Knowledge Management-Volume 129. Australian Computer Society, Inc., pages 13–20.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2016. Neural networks for joint sentence classification in medical paper abstracts. *European Chapter of the Association for Computational Linguistics (EACL) 2017* .

Peter M Dunn. 1997. James lind (1716-94) of edinburgh and the treatment of scurvy. *Archives of Disease in Childhood-Fetal and Neonatal Edition* 76(1):F64–F65.

Kazuo Hara and Yuji Matsumoto. 2007. Extracting clinical trial design information from medline abstracts. *New Generation Computing* 25(3):263–275.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Ke-Chun Huang, I-Jen Chiang, Furen Xiao, Chun-Chih Liao, Charles Chih-Ho Liu, and Jau-Min Wong. 2013. Pico element detection in medical text without metadata: Are first sentences enough? *Journal of biomedical informatics* 46(5):940–946.

Ke-Chun Huang, Charles Chih-Ho Liu, Shung-Shiang Yang, Furen Xiao, Jau-Min Wong, Chun-Chih Liao, and I-Jen Chiang. 2011. Classification of pico elements by text features systematically extracted from pubmed abstracts. In *Granular Computing (GrC), 2011 IEEE International Conference on*. IEEE, pages 279–283.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics* 12(2):S5.

Thøger P Krogh, Torkell Ellingsen, Robin Christensen, Pia Jensen, and Ulrich Fredberg. 2016. Ultrasound-guided injection therapy of achilles tendinopathy with platelet-rich plasma or saline a randomized, blinded, placebo-controlled trial. *The American journal of sports medicine* .

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Human Language Technologies 2016: The Conference of the North American Chapter of the Association for Computational Linguistics, NAACL HLT*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60.

Chris Mavergames. 2013. The future of knowledge: Cochranetech to 2020 (and beyond). 21st Cochrane Colloquium. http://mavergames.info/.

Marcia L Meldrum. 2000. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/oncology clinics of North America* 14(4):745–760.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

David Alexander Robinson. 2012. Finding patient-oriented evidence in pubmed abstracts. *Athens: University of Georgia* .

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26(3):339–373.

Kay Dickersin Tianjing Li. 2015. Introduction to systematic review and meta-analysis. *Coursera* .

Jin Zhao, Praveen Bysani, and Min-Yen Kan. 2012. Exploiting classification correlations for the extraction of evidence-based practice information. In *AMIA*.