Extracting Factual Min/Max Age Information from Clinical Trial Studies

Yufang Hou[1], **Debasis Ganguly**[1], **Léa A. Deleris**[2], **Francesca Bonin**[1]
[1]IBM Research, Ireland
{yhou,debasga1,fbonin}@ie.ibm.com
[2]BNP Paribas
lea.deleris@bnpparibas.com

Abstract

Population age information is an essential characteristic of clinical trials. In this paper, we focus on extracting minimum and maximum (min/max) age values for the study samples from clinical research articles. Specifically, we investigate the use of a neural network model for question answering to address this information extraction task. The min/max age QA model is trained on the massive structured clinical study records from *ClinicalTrials.gov*. For each article, based on multiple min and max age values extracted from the QA model, we predict both actual min/max age values for the study samples and filter out non-factual age expressions. Our system improves the results over (i) a passage retrieval based IE system and (ii) a CRF-based system by a large margin when evaluated on an annotated dataset consisting of 50 research papers on smoking cessation.

1 Introduction

Clinical trials are an important source of scientific evidence for guiding the practice of evidence-based medicine. However, many characteristics of clinical trials are only reported in the published research articles. The health service community could benefit from knowledge bases populated with detailed information from clinical trials reported in research articles. With this in mind, clinical information extraction aims to extract such information from journal articles that report randomized controlled trials (Kiritchenko et al., 2010; Wallace et al., 2016).

Relevant information about clinical trials can be categorised along: *(i)* trial's population characteristics (e.g. minimum and maximum age of the participants, education level, marital status, health status), *(ii)* intervention methods, both what is being done (e.g. specific drug and dosage, planning sessions, use of an app for daily reporting) and how it is being administered (e.g., where, how often and by whom), and *(iii)* outcome of the study (e.g., *30% of the population stopped smoking after 6 months*).

In this paper, we focus on extracting population characteristics and in particular minimum and maximum (min/max) age values associated with the study samples from clinical trials research articles.

Unlike (Summerscales, 2013), our aim is to extract information from the full article, rather than only from the abstract, as we have observed that age information is not always described in the abstracts. In our testing dataset consisting of 50 research papers, only nine papers describe the min/max age information in their abstracts.

Naturally, analysing the entire article presents many challenges. Our goal is to identify the factual min/max age value information for the persons who actually participated in the clinical trial (see Example 1 and Example 2 below). This should be distinguished from non-factual min/max age information (Example 3 and Example 4) and also from min/max age information which is not related to the participants in the study (Example 5 and Example 6).

(1) Participants were 83 smokers, who were **18-23** years old and undergraduate students ...

(2) participants aged **18-24** years were randomized to a brief office intervention (n=99) or to an expressive writing plus brief office intervention (n=97).

(3) To be included in the study, smokers had to be between the ages of **18** and **60** years ...

(4) The subjects were eligible for inclusion if they were at least **18** years of age, reported smoking 10 or more cigarettes per day, ...

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 107–116
Minneapolis, Minnesota, June 7, 2019.

(5) An estimated 23.6% of young adults aged **18-24** years are current smokers.

(6) Smoking Dutch youths had in many cases tried their first cigarette at the age of **11-12** years.

Our proposed system extracts factual min/max age values of the study samples directly from research articles in PDF format. We leverage the massive structured clinical study records from *ClinicalTrials.gov* to provide distant supervision for min/max age value extraction. Furthermore, inspired by the work on hedging detection on Bioscience domain (Light et al., 2004; Kilicoglu and Bergler, 2008; Farkas et al., 2010), we explore a list of "speculation cues" to filter out non-factual min/max age expressions. Our system improves the results over *(i)* a passage retrieval based IE system and *(ii)* a CRF-based system by a large margin when evaluated on an annotated dataset consisting of 50 research papers on smoking cessation.

2 Related Work

2.1 Clinical Information Extraction

In general, research on information extraction from medical literature is still in its infancy involving a number of limitations, such as lack of common benchmarking datasets, and a lack of general consensus on the class of approaches that are reported to work well on such benchmarks.

Some work has been conducted on supervised approaches for medical information extraction. Multiple studies have concentrated their efforts on medical abstract. In (Kim et al., 2011), the authors propose a conditional random filed (CRF) classification method for labelling medical abstract sentences according to medical categories, such as outcome, intervention, population. Hansen et al., 2008 (Hansen et al., 2008) developed a Support Vector Machine algorithm for extracting the number of trials participants from medical abstracts, while in (Hassanzadeh et al., 2014), the authors use a machine learning approach for classifying abstract sentences according to the PICO (Population, Intervention, Comparison, Outcome) scheme.

Other studies have exploited the entire article, for the extraction of papers' metadata as (Lin et al., 2010): the authors propose a preliminary system based on CRF for extracting formulaic text (authors names, email and institution) as well as some key study parameters as free text, from PubMedCentral articles. They reach promising results for the formulaic text, but only moderate success for the free text attributes. The study in (Luan et al., 2017) involves finding key-phrases from scientific articles and then classifying them. However, these categories are much broader (coarse-grained), e.g. 'process', 'task' etc., than the fine-grained categories in our task (min/max age).

A few studies have tackled the min/max age extraction problem. Most research work on extracting information from clinical trial literature considers "eligibility criteria" as a target element, which often contains min/max age information (de Bruijn et al., 2008; Kiritchenko et al., 2010).

However, min/max age information contained in the eligibility criteria refers to the planned min/max age and may be different from the actual min/max age values of the study samples (for example: the researchers could decide to test a population of women between 20-30 years, but realistically they could gather participants only between 22 and 28 years old). (Summerscales, 2013) carefully designed a number of heuristic rules to extract min/max age values of the study population from the abstracts. We differ from this latter work as we (a) extract such information from the full articles and (b) use a machine learning approach. In addition, we integrate the rules designed by (Summerscales, 2013) into our passage retrieval based IE system as a baseline.

Generally, in contrast to previous work, in this paper we a) concentrate on the extraction of population characteristics, b) use the entire article for detecting the min/max age and c) compare an unsupervised approach with a QA-based approach.

2.2 Question Answering

Most recently, *reading comprehension* or *question answering based on context* has gained popularity within the NLP community, in particular since (Rajpurkar et al., 2016) released a large-scale dataset (SQuAD) consisting of 100,000+ questions on a set of Wikipedia articles. In the medical domain, (Šuster and Daelemans, 2018) created a dataset of clinical case reports for machine reading comprehension (CliCR). The dataset contains around 100,000 gap-filling queries on 12,000 case reports. These queries are created by blanking out medical entities in the *learning points* sections using some heuristics.

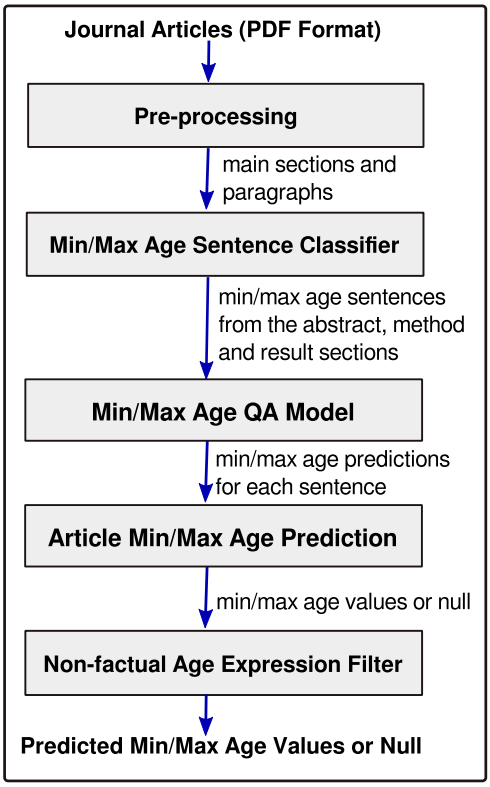

Figure 1: Proposed QA based factual min/max age value extraction framework.

We explore the QA framework for min/max age value extraction. Various neural network models have been proposed for question answering but these models trained on SQuAD or CliCR do not work well in our scenario because these datasets do not contain the queries targeting the specific min/max age values expressed in the text. Therefore, we leverage instead the massive structured clinical study records on *ClinicalTrials.gov* and create the training data for our min/max age value extraction component.

2.3 Non-factual Information Detection

There has been a significant amount of research in detecting speculative language in scientific research articles (Hyland, 1998; Light et al., 2004; Kilicoglu and Bergler, 2008; Medlock and Briscoe, 2007; Farkas et al., 2010; Morante and Sporleder, 2012). Our task requires to extract information from definite statements, therefore we use a list of speculation cues to filter out sentences where min/max age information are expressed speculatively.

3 Approach

We develop a pipeline to extract factual min/max age information from clinical trial studies. We divide the task in two steps: 1) finding sentences containing min/max age information; 2) extracting the value from those sentences. For the first we develop **Min/Max age Sentence Classifier** and for the second we propose a QA approach and develop the module **Min/Max Age QA Model**.

Figure 1 illustrated the process associated with our proposed system. In the following sections, we describe how we create training data from *ClinicalTrials.gov* as well as each component of our system in detail.

3.1 Creating Training Data Using Clinical Study Records

We leverage the massive structured clinical study records on *ClinicalTrials.gov* to create training data for *Min/Max Age Sentence classifier* and *Min/Max Age QA Model*. *ClinicalTrials.gov* is one of the largest database of clinical studies conducted around the world. It currently holds registrations around 273,000 trials from 204 countries. Each trial registration record contains a column called "Eligibility Criteria", additionally min and/or max age values are indicated if they are present in the description text of the eligibility criteria. Figure 2 shows an example of a clinical study record from *ClinicalTrials.gov*.

Note that most min/age expressions in eligibility criteria are speculative (e.g., *at least 21 years of age*, or *child must be ages 6-12 years old*), nevertheless they are still reflective of various linguistic forms for factual min/max age (e.g., *aged 6-12 years old* or *age* $\geq$ 18 *years*). Therefore we expect that the models trained on this "noisy" dataset can still (1) identify sentences containing min/max age information and (2) predict the min/max age values.

3.2 Pre-processing

Given a research article in PDF format, we first extract clean text from the PDF file using GROBID (Lopez, 2009). We associate each paragraph to one of the five main sections: *abstract*, *introduction*, *method*, *result* and *discussion*. This step may introduce some noise (e.g., including the content from the table as the main body text) because parsing PDF file in different styles is a challenging task in itself.

The Symptom Experience Study in Persons With Non-Small Cell Lung Cancer (SES)	
Recruitment Information	
Recruitment Status	Completed
Actual Enrollment	74
Original Estimated Enrollment	86
Eligibility Criteria	Inclusion Criteria: - Women and men at least 21 years of age with suspected NSCLC to be confirmed after surgery. - Planned surgical resection, not diagnostics alone, for treatment of suspected non-small cell lung cancer (NSCLC) to include such surgical approaches as open thoracotomy, video assisted thoracic surgery (VATS), and Robotic procedures. - Karnofsky Performance Status score of at least 70%. - Thoracic surgeon approval pre- and post-surgery. - Medically stable co-morbid conditions including cardiovascular disease such as post-myocardial infarction, stable coronary bypass graft surgery, and stable percutaneous transluminal coronary angioplasty; and mild to moderate cardiopulmonary obstructive disease. - Has phone access capability. - Able to speak and write English. - Able to hear and speak for phone interviews. - Owns a television. - Lives within 1.5 hours driving distance of recruitment site.
Mininum Age	21
Maximum Age	null

Figure 2: An example of a clinical study record from *ClinicalTrials.gov*.

3.3 Identifying Sentences Containing Min/Max Age Information:Min/Max age Sentence Classifier

After pre-processing, we identify sentences that contain min/max age information. At the inference stage, we first split paragraphs to sentences using Stanford CoreNLP Toolkit (Manning et al., 2014), then apply a classifier (*MinMaxAgeSentFinder*) to predict sentences containing min/max age information among all sentences containing the word "age/ages/aged" or "year/years".

To train our *MinMaxAgeSentFinder* classifier, we create the training data using the eligibility criteria of the structured clinical study records from *ClinicalTrials.gov*. Text in eligibility criteria can be quite long (for instance, some criteria contain more than 10 clauses/sentences), so we only keep the clause/sentence which contains the annotated min/max age value(s). More specifically, we first split the eligibility criteria into sentences/clauses using the delimiter "-", then choose the clauses/sentences which contain the annotated min/max age values as well as the word "age/ages/aged" or "year/years". For instance, in the example shown in Figure 2, we will keep the sentence "*Women and men at least 21 years of age with suspected NSCLC to be confirmed after surgery.*" as the positive training instance and filter out other sentences/clauses.

We randomly choose 20,000 such sentences/clauses (10,000 for min age and 10,000 for max age) as positive training instances. Negative training instances are sentences which do not contain the word "age/ages/aged" or "year/years" from 60 clinical research articles. Note that these articles are different from the articles in the testing dataset. We use MaxEnt classifier to train *MinMaxAgeSentFinder* with the following features: adjacent word n-grams (n=1-4) and adjacent letter n-grams within words.

3.4 Predicting Min/Max Age Values for Each Sentence: Min/Max Age QA Model

We approach the problem of extracting values of min/max age from a question-answering perspective. Specifically, our system first reads a sentence, then answers the questions "what is the min/max age of the participants?".

Various neural network models have been proposed for this task but these models trained on SQuAD do not work well in our scenario, because SQuAD does not contain this type of question-answer pairs. Therefore we create training data for max/min age value extraction by leveraging the massive structured clinical study records from *ClinicalTrials.gov* The training data are 10,000 <eligibility criteria–min age> pairs and 10,000 <eligibility criteria–max age> pairs described previously. Note that we use the whole eligibility criteria instead of choosing the specific sen-

tence/clause which contain the min/max age value. We believe that with the additional min/max age information, the question-answering module can locate the position of the min/max age value and learn various patterns for the target question.

We train our min/max age question-answering module (*MinMaxAgeQA*) using the Bi-Directional Attention Flow (BiDAF) Network (Seo et al., 2017). BiDAF uses attention mechanisms in both directions (i.e., question-to-context and context-to-question) to find a sub-phrase from the input text to answer the question.

BiDAF includes both character-level and word-level embeddings. Most word tokenization models are not robust for numeric expressions in scientific literature. For instance, the Stanford CoreNLP tokenizer tokenizes the clause: "*aged 6-12 years old*" as "{*aged, 6-12, years, old*}" - it does not recognize 6 and 12 as two different tokens. The character-level embeddings in BiDAF can overcome this problem and the module correctly predicts 6 is the value of min age for this example.

3.5 Predicting Min/Max Age Values for Each Article

To predict min/max age values of the study samples for each article, we apply *MinMaxAgeQA* to each predicted sentence containing min/max age information from the *abstract*, *method*, and *result* sections on both questions (i.e., *what is the min age of the participants?* and *what is the max age of the participants?*). Answers that do not represent a valid integer number or answers whose confidence score are less than 0.5 are discarded. For each question, we keep the answer with the highest confidence score.

We do not include sentences from the *introduction* section because it may include other min/max age information which is not related to the study samples (see Example 5 and Example 6). We leave filtering out unrelated min/max age information from introductions as future work.

Finally, if both min and max age values are predicted for an article, we check whether the min age value is smaller than the max age value. Otherwise we keep the answer with the higher confidence score and discard the other one. For instance, as shown in Figure 3, the number 16 is predicted as both the min age value (with the probability of 0.956) and the max age value (with the probability of 0.624) for an article, we keep 16 as the prediction for the min age value and set the prediction of the max age to "Null".

Sent: the target sample comprised all cigarette smokers aged 16 or more who attended the surgeries to see a doctor between 4 and 27 November 1980.

Q: What is the min age of the participants?
A: 16 (confidence score: 0.956)

Q: What is the max age of the participants?
A: 16 (confidence score: 0.624) → **Null**

Figure 3: Conflicting min/max age values.

3.6 Non-factual Age Expression Filter

In this component, we filter out a min/max age value prediction if it is expressed speculatively. We first extract the clause which contains the prediction, then check whether a speculation cue word/phrase is present in the clause using the speculation cues from (Light et al., 2004). These cue words are: {*if, at least, must, had to, has to, have to, need, needs*}.

4 Evaluation

4.1 Testing Dataset

The ground-truth dataset used for evaluation comprises a set of 50 published journal articles in PDF format on smoking cessation. The dataset contains around 432k tokens and 18k sentences. Table 1 shows some statistics about the testing dataset. Overall, we have 843 sentences containing the word "age/ages/aged" or "year/years" and these sentences contain 2,226 numeric tokens.

The articles were annotated by a team of four behaviour science domain experts in the context of a broader project focused on leveraging the scientific literature in behaviour change (Michie et al., 2017). Annotation for a particular document was performed by two human annotators using the EPPI tool[1]. The annotation process involved highlighting relevant pieces of text and then assigning them to the corresponding min/max age attribute. Additionally, in order to disambiguate the highlighted text, the annotators were asked to annotate the entire sentence containing the highlighted piece as the additional context. Conflicts in the annotation process were resolved through discussions. Note that not every document contains a min/max age annotation. This is because not every article reports the min/max age of the overall

[1] http://eppi.ioe.ac.uk/CMS/

Testing Dataset	
# of articles	50
# of sentences	18,417
# of tokens (main text)	432,056
# of sentences containing "age/ages/aged" or "year/years"	843
# of numeric tokens in sentences containing "age/ages/aged" or "year/years"	2,226

Table 1: Statistics for the testing dataset.

study samples. In the testing dataset, 35 papers have min age annotations and 25 papers have max age annotations.

4.2 Evaluation Metric

We use recall, precision and F-score for evaluation. Recall is calculated as the number of articles where the min/max age values are correctly predicted divided by the number of articles where min/max age values are annotated. Precision is calculated as the number of articles where the min/max age values are correctly predicted divided by the number of articles where the system makes a min/max age value prediction. F-score is the harmonic average of the precision and recall.

4.3 Baseline 1: *PassageRetrievalBasedMinMaxAgeExtractor*

We developed a passage retrieval based IE system to extract min/max age values (Ganguly et al., 2018). The first step is to retrieve the passages containing 10, 20, and 30 words using the query "*(age OR ages OR aged OR year OR years)*". The intention of retrieving passages is to restrict extraction of factoid answers to potentially relevant small semantic units of text rather than the text of the whole document.

The next step is to use validation criteria to select the likely answer candidates. We use the min/max age patterns from (Summerscales, 2013) as the validation criteria to choose the likely answer candidates from each retrieved passage for min age and max age respectively. These patterns can be viewed as rules which are carefully designed by humans to extract min/max age values. For instance, a rule can be: *if a passage contains the phrase "**greater than X**" or "**older than X**" and* X *is an integer number between 10 to 100, then choose* X *as an answer candidate*. It is worth noting that (Summerscales, 2013) is the only previous work targeting the same task according to our best knowledge. We integrate all the heuristic rules for min/max age value extraction from (Summerscales, 2013) into our passage retrieval based IE system.

Finally, we score the answer candidates by a term proximity function that takes into account the differences in position between the query terms and the candidate answers (Zhao and Yun, 2009). The function is formally defined in the following Equation:

$$sim(c, Q) = \frac{1}{|Q|} \sum_{q \in Q} exp(-(p_c - p_q)^2/\sigma) \quad (1)$$

Equation 1 describes the proximity based ranking function between a candidate answer c and a query Q, denoted by $sim(c, Q)$. Practically, for each word in the passage that matches the query terms (q), the similarity function increases the score of that candidate by an amount that depends on the distance between that matched word and the candidate answer $(p_c - p_q)$. Specifically, we use a Gaussian function centered at each query term to determine the increase in similarity score. The parameter σ controls the bandwidth of the Gaussians and is set to 1 in our experiments.

4.4 Baseline 2: *CRFBasedMinMaxAgeExtractor*

We also developed the second baseline using CRF (Sutton and McCallum, 2012). The training dataset contains the clauses/sentences which contains the annotated min/max age value(s) from the eligibility criteria of the clinical studies registered in *ClinicalTrials.gov*. For each clause/sentence, we use Stanford CoreNLP Toolkit (Manning et al., 2014) to obtain the tokens as well as the POS tags, then we create the corresponding training instance using BIO labels (i.e., Beginning/Inside/Outside

of a min/max age). Table 2 shows the training instance for the example illustrated in Figure 2.

Token	POS tag	MinAgeAnnotation
Women	NNS	O
and	CC	O
men	NNS	O
at	IN	O
least	JJS	O
21	CD	B
years	NNS	O
of	IN	O
age	NN	O
with	IN	O
suspected	VBN	O
NSCLC	NNP	O
...	...	...

Table 2: A training instance for the min age extraction CRF model.

We train two CRF models for min age and max age extraction respectively, using 10,000 training instances for each model. We use words as well as POS tags as features. More specifically, for the word type features, we consider the current word w_i, the surrounding words (w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}), as well as bi-grams (w_{i-1} + w_i, w_i + w_{i+1}) and tri-grams (w_{i-2} + w_{i-1} + w_i, w_{i-1} + w_i + w_{i+1}, w_i + w_{i+1} + w_{i+2}) created from words. We create similar unigram, bi-gram and tri-gram features using the automatically predicted POS tags as well. We also include the combinations of the previous prediction and the current prediction as bi-gram features.

At the inference stage, for each article, we first extract all sentences containing the word "age/ages/aged" or "year/years". We then apply the min/max age CRF model on these sentences and extract all tokens with the predicted label "B". In the end, among all predicted words, we choose the word which represents a valid integer number and has the highest confidence score as the predicted min/max age value for the article.

4.5 Results and Discussion

Table 3 shows the performance of the baselines (*PassageRetrievalBasedMinMaxAgeExtractor* and *CRFBasedMinMaxAgeExtractor*) as well as our system (*QABasedMinMaxAgeExtractor*, described in Section 3) for extracting min/max age values of the study samples.

For *MinAge*, the first baseline (*PassageRetrievalBasedMinMaxAgeExtractor*) achieves a very high precision score (90.9%) but suffers from low recall (28.6%). The second baseline (*CRFBasedMinMaxAgeExtractor*) improves the recall by 21.4 points but only achieves a precision score of 42.5%. Compared to the first baseline, our system manages to improve recall by 37.1 points and still achieves a reasonable level of precision (79.3%). Overall, our system improves the results over the two baselines by a large margin regarding F-score (71.9% vs. 43.5%, and 71.9% vs. 45.9%).

The similar pattern is also observed for *MaxAge*: Our system improves the results over the first baseline by a substantial margin on recall (60.0% vs. 32.0%) and F-score (66.7% vs. 44.4%) respectively.

It might seem surprising that *CRFBasedMinMaxAgeExtractor* performs much worse than *PassageRetrievalBasedMinMaxAgeExtractor* for *MaxAge*. This is because many max age values in scientific articles are not correctly recognized as a single token by Stanford CoreNLP Toolkit. For instance, the tokenization model predicts that "*18-60*" or "*<=60*" as single tokens. In contrast, our system is more robust for parsing such numeric expressions.

In addition, it seems that the carefully designed min/max age patterns in the first baseline only cover a few forms of min/max age expressions. On the contrary, our min/max age question-answering module (*MinMaxAgeQA*, Section 3.4) trained over a large-scale dataset can capture various linguistic expressions of min/max age in natural language, for instance, "$\geq$ 18 *years of age*" or "*age >= 18 years*".

4.6 Analysis

To better understand the roles of different components in our system, we carried out a few experiments:

- —*WO MinMaxAgeSentFinder*: instead of using *MinMaxAgeSentFinder* to find the sentences containing min/max age information, we pass all sentences containing the word "age/ages/aged" or "year/years" from the abstract, method, and result sections to the next component *MinMaxAgeQA*.
- —*WO MinMaxAgeQA*: we use the most common min/max age expression pattern in clini-

	MinAge			*MaxAge*		
	R	P	F	R	P	F
Baseline 1: PassageRetrievalBasedMinMaxAgeExtractor	28.6	90.9	43.5	32.0	72.7	44.4
Baseline 2: CRFBasedMinMaxAgeExtractor	50.0	42.5	45.9	25.0	18.2	21.1
This work: QABasedMinMaxAgeExtractor	65.7	79.3	**71.9**	60.0	75.0	**66.7**

Table 3: Experimental results. Bold indicates statistically significant differences over the baseline using randomization test ($p < 0.01$).

	MinAge			*MaxAge*		
	R	P	F	R	P	F
This work: QABasedMinMaxAgeExtractor	65.7	79.3	**71.9**	60.0	75.0	**66.7**
—WO MinMaxAgeSentFinder	68.6	68.6	**68.6**	52.0	56.5	54.2
—WO MinMaxAgeQA	31.4	84.6	45.8	40.0	71.4	51.3
—WO Non-factualSentFilter	68.6	70.6	**69.6**	60.0	71.4	**65.2**

Table 4: Contribution of each component to the overall system performance.

cal trial studies "*X-Y*" (e.g., *18-23 years old*) to predict min and max age values from the first sentence contain such a pattern.

- *—WO Non-factualSentFilter*: Non-factual age expression filter is not used.

The results of these experiments are shown in Table 4. It seems that *MinMaxAgeQA* has the most impact on the performance while *Non-factualSentFilter* has less of an impact. In addition, *MinMaxAgeSentFinder* has more impact on the results of *MaxAge* compared to *MinAge*.

We also performed some error analysis on our full system. We noticed that the noise introduced in the pre-processing step (e.g., missing some paragraphs) is the main reason to cause our system to predict "Null" for articles with min/max age annotation. For cases where a wrong min/max age value is predicted, they are often embedded in the speculative expressions which are not captured by our current *Non-factualSentFilter*. For instance, the system predicts **24** as the max age for one article in which **24** appears in a speculative sentence (see *speculative expression* in Example 7). For this article, the annotation for max age is **23** (see *factual expression* in Example 7).

(7) (*speculative expression*) Eligibility for this study included being a student (full or part time), smoking at least 1 cigarette/day in each of the past 7 days, being aged 18-**24** years, and being interested in quitting smoking in the next 6 months. (*factual expression*) Participants were 83 smokers, who were 18-**23** years old and undergraduate students at a university.

5 Conclusions

This paper aims to extract factual min/max age values of the study samples from clinical research papers. We leverage the large-scale records from the *ClinicalTrials.gov* database to provide distant supervision for our system. We also explore "speculative cues" and the structure of the scientific papers to extract information from factual statements about the target study. We show that our approach outperforms a passage retrieval based IE system and a CRF-based model by a large margin on a testing dataset consisting of 50 journal articles and around 18,000 sentences.

In the future, we plan to extend our framework to extract other types of numeric values from the clinical research papers, such as the outcome values of the different intervention groups and the control group (e.g., *40% of PP abstinence rates*), as well as the time frame of the follow up (e.g., *52 weeks* or *6 months*).

Acknowledgments

This work was supported by a Wellcome Trust collaborative award as a part of the Human Behaviour-Change Project (HBCP): Building the science of behaviour change for complex intervention development (grant no. 201,524/Z/16/Z).

References

Berry de Bruijn, Simona Carini, Svetlana Kiritchenko, Joel Martin, and Ida Sim. 2008. Automated information extraction of key trial design elements from clinical trial publications. In *AMIA 2008 Annual Symposium, Washington DC, USA, November 8-12, 2008*, pages 141–145.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Shared Task of the 14th Conference on Computational Natural Language Learning,* Uppsala, Sweden, 15–16 July 2010, pages 1–12.

Debasis Ganguly, Léa A Deleris, P Aonghusa Mac, Alison J Wright, Ailbhe N Finnerty, Emma Norris, Marta M Marques, and Susan Michie. 2018. Unsupervised information extraction from behaviour change literature. *Studies in health technology and informatics*, 247:680–684.

Marie J Hansen, Nana Ø Rasmussen, and Grace Chung. 2008. A method of extracting the number of trial participants from abstracts describing randomized controlled trials. *Journal of Telemedicine and Telecare*, 14(7):354–358. PMID: 18852316.

Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. 2014. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of Biomedical Informatics*, 49:159 – 170.

Ken Hyland. 1998. *Hedging in scientific research articles*. John Benjamins, Amsterdam, The Netherlands.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: A linguistically motivated perspective. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing,* Columbus, Ohio, 19 June 2008, pages 46–53.

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.

Svetlana Kiritchenko, Berry de Bruijn, Simona Carini, Joel Martin, and Ida Sim. 2010. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Medical Informatics and Decision Making*, 10(1):56–73.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of Bioscience: Facts, speculations, and statements in between. In *Proceedings of the HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases,* Boston, Mass., 6 May 2004, pages 17–24.

Sein Lin, Jun-Ping Ng, Shreyasee Pradhan, Jatin Shah, Ricardo Pietrobon, and Min-Yen Kan. 2010. Extracting formulaic and free text clinical research articles metadata using conditional random fields. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, Louhi '10, pages 90–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Patrice Lopez. 2009. GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications. In *The 13th European Conference on Digital Libraries (ECDL 2009), Corfu, Greece, September 27 - October 2, 2009*, pages 473–474.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. *CoRR*, abs/1708.06075.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the ACL 2014 System Demonstrations,* Baltimore, USA, 22–27 June 2014, pages 55–50.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, pages 992–999.

Susan Michie, James Thomas, Marie Johnston, Pol Mac Aonghusa, John Shawe-Taylor, Michael P Kelly, Léa A Deleris, Ailbhe N Finnerty, Marta M Marques, Emma Norris, et al. 2017. The human behaviour-change project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. *Implementation Science*, 12(1):121.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* Austin, Texas, USA, 1–4 November 2016, pages 2383–2392.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of the 5th International Conference on Learning Representations,* Toulon, France, 2017.

Rodney L. Summerscales. 2013. *Automatic Summarization of clinical abstracts for evidence-based medicine*. Ph.D. thesis, Illinois Institute of Technology, Chicago, Illinois.

Charles Sutton and Andrew McCallum. 2012. An introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.

Simon Šuster and Walter Daelemans. 2018. Clicr: A dataset of clinical case reports for machine reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* New Orleans, Louisiana, 1–6 June 2018, pages 1551–1563.

Byron C. Wallace, 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*, 17(132):1–25.

Jinglei Zhao and Yeogirl Yun. 2009. A proximity language model for information retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* Boston, Mass., 19–23 July 2009, pages 291–298.