Extracting Characteristics of the Study Subjects from Full-Text Articles

Dina Demner-Fushman, MD, PhD, James G Mork, MS
National Library of Medicine, National Institutes of Health, HHS Bethesda, MD, USA

Abstract

Characteristics of the subjects of biomedical research are important in determining if a publication describing the research is relevant to a search. To facilitate finding relevant publications, MEDLINE citations provide Medical Subject Headings that describe the subjects' characteristics, such as their species, gender, and age. We seek to improve the recommendation of these headings by the Medical Text Indexer (MTI) that supports manual indexing of MEDLINE. To that end, we explore the potential of the full text of the publications. Using simple recall-oriented rule-based methods we determined that adding sentences extracted from the methods sections and captions to the abstracts prior to MTI processing significantly improved recall and F_1 score with only a slight drop in precision. Improvements were also achieved in directly assigning several headings extracted from the full text. These results indicate the need for further development of automated methods capable of leveraging the full text for indexing.

Introduction

Retrieval of publications for clinical decision support or database curation often relies on information about study subject characteristics. In publications indexed for MEDLINE®, this information is provided in structured, normalized form in Medical Subject Headings® (MeSH®) and can be easily used to find, for example, articles pertaining to preschool children or inbred mice. When this information is not available in structured form or directly stated in MEDLINE abstracts, it needs to be derived or extracted from the full text of the publications. The U.S. National Library of Medicine (NLM) Medical Text Indexer (MTI) tool[1] that assists manual annotation of MEDLINE citations provided by the NLM Index section is fairly accurate in extracting and deriving some of this information available in the abstracts. For example, MTI achieves 93.84% F-score on assigning the heading "Humans", whereas some other headings pertaining to study subjects are still far from this level of accuracy[2]. Indexers at NLM have access to the full text of an article, while only the title and abstract are provided to MTI. Over the years, Indexers have noted that specific detailed information about the study subjects that MTI missed was included in the full text of an article, but, not in the title or abstract. In this work, we present simple methods for extracting specific Mice and Rat strains, species, age groups, and gender from the full text of publications and explore if information extracted from the full text improves MTI performance for these headings.

Extraction of the subjects' characteristics has been addressed in the past; mostly separately for clinical trials and for identifying animal species studied in the paper. For clinical trials, researchers are extracting such characteristics as the names and time points for primary and secondary outcomes, eligibility criteria and sample size. In the most common approach to extraction of the trials characteristics, the sentences potentially containing this information are identified first using statistical methods, and then knowledge-based methods are used to extract information. Xu et al. found candidate sentences in MEDLINE abstracts using HMM classifiers and then used syntactic parse patterns and rules to extract subject descriptors, such as *man*, *elderly*, *diabetic*, *brain-injured*; the study size (number of participants) and the studied disease[3]. Note that Xu et al. did not further separate subject descriptors by age, gender, or disease class. Similarly, de Bruijn et al. used an SVM trained on 78 full text articles to identify the most promising sentences and then applied simple extraction patterns and rules (called weak extraction rules) to identify over 20 trial characteristics, such as the enrollment start date and the experimental and control interventions[4]. The weak rules implemented in the ExaCT system[5], in which the extracted text for each of the 21 trial characteristics were assessed by curators, corrected and then stored in the database, were sufficient for identifying the majority of trial elements in sentences recommended by SVM classifiers. Zhao et al. used Maximum Entropy classifiers to first assign sentences in a collection of 19,893 medical abstracts and full text articles to one or more of five classes: Patient, Intervention, Result, Study Design, and Research Goal[6]. Then the words in the sentences assigned to these classes were further classified as: Sex, Condition, Race, and Age. Two of these characteristics overlap with our targets: gender and age. In a 5-fold cross validation on 52 words in the Sex class, and 175 words in the Age class, Zhao et al. achieved 90% F-score for gender and 81% F-score for age extraction. Similarly, Kelly and Yang report perfect recall and precision for regular expression-based extraction of gender and age from 17 MEDLINE sentences containing information about the subjects' age and 171 sentences with gender information[7].

Several systems for identification of species in biomedical text are publicly available. Gerner et al. have developed a dictionary-based system LINNAEUS that was evaluated using MeSH species headings in MEDLINE citations, among other reference standards[8]. When using MEDLINE abstracts as input to the system, LINNAEUS achieved 52% precision and 68% recall. On the full text of the open access subset of PubMed Central, the system achieved 95% recall with a significant drop in precision to 13%. Naderi et al. have developed a hybrid rule-based/machine learning system, OrganismTagger[9]. Pafilis et al. have developed an open source SPECIES tagger, comparable in performance and faster than LINNAEUS[10]. As part of a large-scale multi-level event extraction effort, Pyysalo et al. achieved over 86% F-score extracting organism mentions, among other entities, using a single model that jointly predicts all entity types.[11]

In the previous work directly concerned with improving MTI performance, Jimeno-Yepes et al. compared extraction of MeSH terms from the abstracts, full text articles, and automatic summaries of different lengths[12]. For the headings addressed in our current work, the authors found that the machine learning methods currently implemented in MTI and trained on the abstracts have higher precision and somewhat lower recall than when summaries or full text are used as input[2]. In this work, we continue exploring if the benefit of high recall offered by the full text of an article could be leveraged without significant losses in precision. Since the rule-base methods have shown high recall and precision in the previous evaluations discussed above, we start our targeted exploration of the study subjects' characteristics with rule-based methods.

Our goals are threefold: first, we still do not have conclusive evidence that full text will significantly improve MTI performance on the headings pertaining to subjects' characteristics; second, we would like to know if focusing on specific sections of the articles, particularly the methods section or captions will be more beneficial; and third, we would like to know if we should augment the original citations with the sentences extracted from the full text and then process these augmented citations using the MTI algorithm, or if we should directly assign the headings to citations using manually prepared lists of mappings of the extracted characteristics to MeSH.

In this work, we explored extraction of the following 29 MeSH Check Tags: *Adolescent*; *Adult*; *Aged*; *Aged, 80 and over*; *Animals*; *Bees*; *Cats*; *Cattle*; *Cercopithecus aethiops*; *Chick Embryo*; *Child*; *Child, Preschool*; *Cricetinae*; *Dogs*; *Female*; *Guinea Pigs*; *Horses*; *Humans*; *Infant*; *Infant, Newborn*; *Male*; *Mice*; *Middle Aged*; *Pregnancy*; *Rabbits*; *Rats*; *Sheep*; *Swine*; and *Young Adult*. We also explored all 51 specific strains of *Mice* and *Rats* under the *Murinae* [B01.050.150.900.649.865.635.505] 2015 MeSH tree, collectively identified as *subject terms* in this study. The specific *Mice* and *Rat* strains were included in our study because they are also typically mentioned in the full text of a paper and not in the title or abstract. Check Tags are a special type of MeSH term that is required to be included for each article and covers species, sex, human age groups, historical periods, pregnancy, and various types of research support (e.g., *Male*)[13].

Methods

We used the 2014 MTI Test Collection that contains 143,658 citations randomly selected from the pool of citations indexed in the last year[14]. Of these, 14,829 (10.32%) full-text articles are available in the Open Access subset of PubMed Central[15]. We downloaded the articles in XML format and used the XML structure to evaluate extraction of the *subject terms* from various sections of the full text.

To identify the Methods sections that are most likely to describe the study subjects, we first extracted all section headings from the XML files. We then manually reviewed the names of the sections and the section titles and created a lookup list of the section names and titles most likely pertaining to Methods.

We then implemented a simple one-pass algorithm that parses the XML files, identifies candidate sentences using trigger words and extracts subjects' characteristics from the candidate sentences. When the XML structure is parsed, information about the current section is stored and assigned to the sentences extracted from the section.

In the first step, the algorithm identifies candidate sentences potentially containing the subjects' characteristics. Aiming for high recall, we qualify a sentence to be a candidate if it contains any members of the lookup lists or matches subject-related regular expressions described below. Using the section label, we determine if the sentence is found in the title/abstract, methods, caption, or anywhere in the body of the paper, excluding the abstract.

In the second step, the algorithm applies the gender and age extraction rules and looks up a MeSH heading corresponding to the list entry or the regular expression found in the sentence. We established the mappings for each entry and expression manually as described below. Finally, using the section label attached to each sentence we generate the following files for our experiments: subjectLinesMethods.txt (sentences extracted from the Methods sections only), subjectLinesMethodsCaptions.txt (sentences extracted from the Methods sections and Figure and Table captions), and subjectLinesBody.txt (sentences extracted from any section in the paper).

The final output of the algorithm consists of the sentences and information extracted from the sentences as shown in Figure 1.

```
23763249|B|8453|Human|Male|Child, Preschool| Preschool participants (n = 52) ranged
in age from 3 years 1 month to 6 years 0 months (mean = 4 years 6 months; SD = 8.04
months), of whom thirty-two were boys (62%).

23637827|B|23050|Chick Embryo; Mice|||The H1N1/177 exhibited an equivalent virus
titer in chicken embryos and mice, and increased virulence and pathogenicity in
mice.

23650499|B|6838|Mice, Inbred BALB C;Mice|Female|| Female BALB/c nude mice (5 - 6
weeks-old; Charles River, Wilmington, MA) were subcutaneously injected with 1.5 *
10 6 BxPC-3 or MIA PaCa-2 cells in 100 u l PBS into each flank.
```

Figure 1 MeSH headings extracted from full-text sentences. The pipe-separated output presents: PMID; section of the full text (B stands for Body); strains and species headings; gender headings; age headings; and the sentence.

Dictionaries and regular expressions

For identifying study subjects, their gender and age, we adapt the dictionaries and algorithms developed previously to identify patients' characteristics in MEDLINE abstracts[16] and clinical text[17]. Briefly, our lookup list for human study participants consists of the manually curated concepts in the UMLS semantic type Population Group[17]. We expanded the study subject list with case-insensitive regular expressions corresponding to MeSH entry terms for the animals in our *subject terms* listed above. For example, for Swine, we added to the list the following terms: `\\WPig[s]?\\W, \\WHog[s]?\\W`, Phacochoerus, Suidae, and Warthogs.

For gender, we use two case-insensitive regular expressions: `\\W(male[s]?|man|men|boy[s]?)\\W` and `\\W(female[s]?|pregnan[tcy]?|women|girl[s]?)\\W`. The second expression also extracts the Pregnancy heading.

Finally, for the subjects' ages, we used both a lookup list of the terms in the UMLS semantic type Age Group and MeSH Age Groups and a set of regular expressions for identifying exact subject ages and ranges, for example, `(?:mean|M)?\\W*age[d|s]?\\W*(?:range|from)?\\W+\\d+\\W*(?:to)?\\W*\\d*\\W*(?:year|day|week) or \\d+[\\s \\-\\d]*(?:year[s]?|month[s]?|week[s]?)[\\s\\-]*old`. We normalized the exact ages to MeSH terms in our list of *subject terms* using MeSH Scope Notes[19], for example, ages >= 65 AND <= 79 map to Aged.

Evaluation

The current abstract-based MTI performance for Check Tags and the specific Mice and Rat strains serves as comparison in all our experiments and the actual human indexing for these citations serves as the reference standard.

To evaluate the contributions of the specific sections, we extracted sentences and headings from the methods sections alone, from the methods sections and captions, and finally, from anywhere in the body of the paper, excluding the abstract.

We evaluated the contributions of the full text under two conditions: 1) adding candidate sentences to the titles or abstracts of MEDLINE citations and processing these extended abstracts using the current MTI algorithm, and 2) normalizing the characteristics extracted from the sentences to MeSH terms and directly assigning these *subject terms* to citations.

We used recall, precision and F_1 score as evaluation metrics. We computed recall as the proportion of the gold standard *subject terms* that were correctly assigned by our tools and precision as the proportion of the *subject terms* assigned by the tools that were correct. The F_1 score is the harmonic mean of recall and precision.

Results

The corresponding XML tags consistently identified the abstract, captions and the body of the paper. We found significant variations in the section naming, both in terms of the XML structure and the titles themselves. The structure was either providing the section name as an attribute of the section tag: for example, `<sec id="Sec2" sec-type="materials|methods">` or providing the name as title after the section tag, for example, `<sec id="Sec13"><title>Participants</title>`. We collected 163 section type variations for the Methods sections, such as: "intro|methods", "materials", "materials-methods", "materials|methods", "method", "methods", "methods|conclusions", "methods|results", "methods|subjects", "subjects", "subjects|methods" and 116 titles, such as "Methodology", "Methodology and Findings", "Methodology and Principal Findings", "Methodology/Findings", "Methodology/Principal Finding", "Methods", "Methods Findings", "Methods and Findings", "Methods and Results", "Methods and design", "Methods and materials", "Methods/Design".

Table 1 presents the results of the evaluation of full text either added to the titles and abstracts or directly contributing Check Tags from our list of *subject terms*. The type, which sentence file was used, Recall, Precision, F_1, and the number of Check Tags matched to the human indexing are provided for each experiment. Files used for the experiments include: subjectLinesMethods.txt (1), subjectLinesMethodsCaptions.txt (2), and subjectLinesBody.txt (3).

Table 1 Results of the evaluation of Check Tags assignment based on full text articles

Experimental set-up	File	Recall	Precision	F_1	Matched CTs
MTI baseline (currently in use at NLM)	-	74.09%	81.35%	77.55%	31,588
Title expansion with sentences from the Methods section	1	78.19%	77.57%	77.88%	33,339
Abstract expansion with sentences from the Methods section	1	78.19%	78.26%	**78.20%**	33,315
Title expansion with sentences from the Methods section and captions	2	79.70%	76.05%	77.84%	33,983
Abstract expansion with sentences from the Methods section and captions	2	79.60%	76.79%	78.17%	33,937
Title expansion with sentences anywhere in the paper body	3	85.70%	58.07%	69.23%	36,541
Abstract expansion with sentences anywhere in the paper body	3	85.52%	59.86%	70.42%	36,463
Direct assignment of Check Tags with sentences from methods and captions	2	79.28%	74.17%	76.64%	33.802
Direct assignment of Check Tags with sentences anywhere in the paper body	3	86.42%	55.97%	67.94%	36,848

Sentences extracted from the full text consistently increased recall in assignment of Check Tags, independently of the way they were used. In all cases we also observed a drop in precision, however, for the sentences extracted from the Methods section and added to the abstracts (bolded F_1 in Table 1), the drop in precision was relatively small and the F_1-score has increased compared to the MTI baseline. As the scope of the included text increased, the numbers of citations for which candidate sentences were found also increased from 2,696 for 20,813 sentences from the methods sections only, to 9,834 for 79,610 sentences from the methods sections and captions, and to 326,993 sentences from 12,800 citations when the whole paper was considered. The fact that 2,029 articles, for which we had no candidate sentences, include Check Tags in the reference standard indicates that our lookup lists were incomplete.

The direct assignment of the Check Tags produced mixed results. Table 2 shows the results for the individual terms compared to the current MTI baseline. Only 13 of the 29 Check Tags showed moderate improvements offset by significant degradation in performance for the remaining Check Tags. Notably, for all of the Check Tags that rely on simply finding a term in the sentence, there was a significant drop in the F_1 score. The gender terms are an exception to this observation with a slight improvement for both *Female* and *Male* tags. There were three cases (bolded results in Table 2) where Precision, Recall, and F_1 all improved for a Check Tag (Aged; Aged, 80 and over; and Cricetinae). The age tags that rely on both the dictionary terms and patterns showed improvement for the terms that do not occur in the text often and have to rely more on extracting the ages and mapping the numeric values to headings. For example, given the sentence "Separate analyses were conducted for children (age 6-11y), adolescents (age 12-19y), and for younger (age 20-50y) and older adults (>= 51y)", our algorithm extracts the following tags: `23656639|B|6100|Humans||Young Adult; Middle Aged; Adolescent; Adult; Child`. While Child (children), Adolescent (adolescents), and Adult (older adults) are mentioned directly in the text, based on the MeSH age range rules, we also included Young Adult (19-24) and Middle Aged (45-64).

Table 2 includes Precision, Recall, and F_1 for both the MTI Baseline and the Full Text Check Tag results and a column showing the differences between F_1 scores. Improvements with Full Text are highlighted in tan, and major negative results are highlighted in yellow.

Table 2 Individual directly assigned tags compared to the current MTI baseline

	MTI Baseline			Full Text			F_1 Diff
MH	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**	
Adolescent	62.14%	28.89%	39.44%	56.24%	42.49%	48.41%	8.96%
Adult	66.34%	67.56%	66.95%	56.12%	73.02%	63.47%	-3.48%
Aged	68.67%	59.76%	63.90%	**68.70%**	**64.15%**	**66.35%**	**2.45%**
Aged, 80 and over	49.82%	18.38%	26.85%	**52.86%**	**26.22%**	**35.05%**	**8.20%**
Animals	91.77%	82.54%	86.91%	86.19%	87.63%	86.90%	-0.01%
Bees	75.00%	100.00%	85.71%	69.23%	100.00%	81.82%	-3.90%
Cats	62.79%	96.43%	76.06%	58.33%	100.00%	73.68%	-2.37%
Cattle	75.82%	79.31%	77.53%	69.65%	80.46%	74.67%	-2.86%
Cercopithecus aethiops	60.00%	34.62%	43.90%	50.00%	40.38%	44.68%	0.78%
Chick Embryo	100.00%	10.53%	19.05%	83.33%	52.63%	64.52%	45.47%
Child	62.07%	55.99%	58.88%	42.42%	67.54%	52.12%	-6.76%
Child, Preschool	70.30%	42.69%	53.13%	40.91%	51.37%	45.55%	-7.58%
Cricetinae	56.41%	37.93%	45.36%	**56.82%**	**43.10%**	**49.02%**	**3.66%**
Dogs	84.68%	73.44%	78.66%	79.03%	76.56%	77.78%	-0.88%
Female	82.37%	79.30%	80.81%	80.04%	85.35%	82.61%	1.80%
Guinea Pigs	94.44%	94.44%	94.44%	80.95%	94.44%	87.18%	-7.26%
Horses	66.00%	97.06%	78.57%	27.97%	97.06%	43.42%	-35.15%
Humans	89.86%	91.65%	90.74%	84.06%	93.01%	88.31%	-2.44%
Infant	60.69%	45.69%	52.13%	41.98%	51.15%	46.11%	-6.02%
Infant, Newborn	70.21%	41.60%	52.24%	30.00%	50.42%	37.62%	-14.63%
Male	79.34%	78.41%	78.87%	77.34%	84.95%	80.97%	2.10%
Mice	92.96%	77.96%	84.80%	89.54%	87.56%	88.54%	3.73%
Middle Aged	74.75%	70.05%	72.32%	74.67%	73.91%	74.29%	1.96%
Pregnancy	79.76%	88.50%	83.90%	66.73%	90.11%	76.68%	-7.23%
Rabbits	89.23%	63.04%	73.89%	72.16%	76.09%	74.07%	0.19%
Rats	93.95%	74.69%	83.22%	88.61%	80.90%	84.58%	1.36%
Sheep	55.17%	88.89%	68.09%	20.37%	91.67%	33.33%	-34.75%
Swine	74.03%	89.76%	81.14%	19.32%	93.70%	32.03%	-49.11%
Young Adult	57.73%	20.00%	29.71%	56.03%	29.81%	38.92%	9.21%

Table 3 shows the results for 27 of the 51 specific strains of Mice and Rats where changes in performance were noted. The remaining 24 strains were not identified in the full text in this study and not included in Table 3. Similarly to the Check Tag performance in Table 2, we have mixed results for the specific strains of Mice and Rats as shown in Table 3. Only 12 of the 27 specific strains of Mice and Rats showed moderate improvements offset by significant degradation in performance for the remaining strains. Both "Rats, Inbred Lew" and "Rats, Zucker" (bolded results in Table 3) are cases where Precision, Recall, and F_1 all improved from the use of full text. In the case of "Mice, Inbred SENCAR"*, the results in Table 3 show no change in performance, but, in fact the full text provided 16 new cases of this term which were all incorrect.

Table 3 includes Precision, Recall, and F_1 for both the MTI Baseline and the Full Text specific strains of Mice and Rats results and a column showing the differences between F_1 scores. Improvements with Full Text are highlighted in tan, and major negative results are highlighted in yellow.

Table 3 Individual assigned specific strains of Mice and Rats compared to the current MTI baseline

	MTI Baseline			Full Text			F_1
MH	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**	**Diff**
Mice, Inbred BALB C	61.36%	24.77%	35.29%	58.88%	57.80%	58.33%	23.04%
Mice, Inbred C3H	50.00%	50.00%	50.00%	33.33%	100.00%	50.00%	0.00%
Mice, Inbred C57BL	74.73%	28.96%	41.74%	63.54%	65.00%	64.26%	22.52%
Mice, Inbred CBA	100.00%	30.00%	46.15%	37.50%	60.00%	46.15%	0.00%
Mice, Inbred CFTR	100.00%	100.00%	100.00%	50.00%	100.00%	66.67%	-33.33%
Mice, Inbred DBA	100.00%	55.56%	71.43%	60.00%	66.67%	63.16%	-8.27%
Mice, Inbred ICR	50.00%	5.56%	10.00%	43.75%	38.89%	41.18%	31.18%
Mice, Inbred NOD	77.78%	60.00%	67.74%	56.86%	82.86%	67.44%	-0.30%
Mice, Inbred SENCAR*	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Mice, Knockout	69.11%	31.84%	43.59%	58.57%	46.07%	51.57%	7.98%
Mice, Nude	71.15%	31.36%	43.53%	70.59%	30.51%	42.60%	-0.93%
Mice, SCID	76.00%	36.54%	49.35%	47.89%	65.38%	55.28%	5.93%
Mice, Transgenic	76.60%	37.50%	50.35%	75.79%	37.50%	50.17%	-0.18%
Rats, Inbred ACI	0.00%	0.00%	0.00%	2.63%	100.00%	5.13%	5.13%
Rats, Inbred BB	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Rats, Inbred BN	50.00%	100.00%	66.67%	9.09%	100.00%	16.67%	-50.00%
Rats, Inbred BUF	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Rats, Inbred F344	50.00%	25.00%	33.33%	44.44%	50.00%	47.06%	13.73%
Rats, Inbred Lew	50.00%	20.00%	28.57%	**55.56%**	**100.00%**	**71.43%**	42.86%
Rats, Inbred OLETF	100.00%	100.00%	100.00%	50.00%	100.00%	66.67%	-33.33%
Rats, Inbred SHR	83.33%	83.33%	83.33%	62.50%	83.33%	71.43%	-11.90%
Rats, Inbred WF	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Rats, Inbred WKY	60.00%	75.00%	66.67%	50.00%	100.00%	66.67%	0.00%
Rats, Long-Evans	50.00%	28.57%	36.36%	37.50%	42.86%	40.00%	3.64%
Rats, Sprague-Dawley	83.70%	37.93%	52.20%	75.45%	62.07%	68.11%	15.90%
Rats, Wistar	83.05%	38.28%	52.41%	75.65%	67.97%	71.60%	19.20%
Rats, Zucker	83.33%	83.33%	83.33%	**85.71%**	**100.00%**	**92.31%**	8.97%

Discussion

Our first question – whether full text holds promise for improving MTI performance was answered positively, as we see significant improvements for thirteen Check Tags and twelve strains of Mice and Rats. The results also hint at the huge potential of full text with our highest recall of 86.42% compared to our MTI Baseline recall of 74.09%. Our results indicate however, that the simple methods that have been reported to show excellent results for extracting study subject characteristics lead to drop in precision for about half of the Check Tags and the specific strains of Mice and Rats. An example of such drop is the Check Tag *Horses* that relies on finding words *horse*, *horses* or *equus* in the text. It turns out, that the word *horse* occurs very frequently in the context of *horse serum* that

is added as supplement to culture media. We will need to look at ways to improve our selection criteria and add filtering to reduce the false positives.

With respect to the second question, whether we should focus on specific sections of the articles, the differences in recall achieved when using the whole text and the Methods sections indicate that it might be more promising to focus on finding good candidate sentences anywhere in the text. Although the recall for candidate sentences was fairly high, our experiments indicated that our lookup lists are incomplete. Inspection of the full text of some of the articles from which no sentences were extracted showed that we systematically missed the age groups in the case reports for single patients because the terms *man*, *farmer* and *woman* were not included in our trigger terms. Generally, singular nouns were intentionally excluded from our trigger lists tailored for extracting characteristics of the subjects of clinical trials because for that task they often triggered false positives. Similarly, some age patterns are missing, for example, *aged between \\d and \\d years.* In some cases however, our algorithm will not be able to assign a Check Tag even if the trigger list is exhaustive. For example, the Check Tag *Female* was assigned to MEDLINE citation 23552690, however, the paper presents a study conducted on breast cancer cells and does not contain any gender-specific terms. It is quite possible that the specific cell lines discussed in the paper indicate that the tissues were female. If this assumption is correct, to assign the Check Tag to this article, we will either need to incorporate more domain knowledge or hope to find a sufficient number of examples for a machine learning algorithm. In the future, we will expand the lists using machine learning methods or manually inspecting citations that are indexed with Check Tags, but for which no candidate sentences were extracted using the current lookup lists. We also need to focus on identifying the higher quality candidate sentences to improve precision.

Our third question was whether we should add candidate sentences to the title or abstracts prior to MTI processing, or attempt to extract the *subject terms* directly using the simple mapping rules. The current results show that the answer for some of the age groups, strains of Mice and Rats, and gender might be to assign the *subject terms* directly, but for the majority of the tags augmenting the abstracts appears to be safer at the moment. Although the F_1 score is 1% higher than the current MTI baseline that supports the NLM Index section when the full text sentences are added to the abstracts, the corresponding 3% drop in precision indicates that we need to further explore how to use the full text.

Our work has the typical limitations of a feasibility study: we focused on testing the hypotheses rather than making sure that our trigger term lists are complete and all our extraction rules take into account the context surrounding the trigger terms. We chose to explore the hypotheses "breadth-first" exploring all parts of the full text, rather than following up "in-depth" with the very promising Methods sections results. These limitations also clearly define the future work that we plan to conduct shortly: expand the lists of trigger terms; maximize the benefits of using the Methods section; refine our extraction rules; and use the extracted sentences in weakly supervised machine learning experiments.

An additional consideration for pursuing methods based on the full text is its availability. MTI does not have access to the full text of an article at this time due to contractual reasons and only utilizes the title and abstract to produce its recommendations for the manual indexing performed at NLM. One possibility is to utilize the full text while it is available in memory for a relatively short period of time while it is being processed in the NLM Document Management System. To use the text in this short time, our algorithms need to be fast and at the same time offer significant benefits to justify the substantial efforts needed for including the full text in MTI processing. We hope that work like the research detailed in this paper will provide incentive for publishers to grant MTI access to full text in order to provide more complete recommendations for the MeSH indexing of their articles.

Conclusion

Our study shows that the full text of biomedical articles has potential to significantly improve automatic indexing of MEDLINE citations with MeSH headings pertaining to the study subjects' characteristics. Furthermore, we show that simple rule-based methods significantly outperform the current automated indexing provided by NLM's Medical Text Indexer for 25 of the 56 *subject terms* in our study, in some cases significantly better (Chick Embryo **+45.47**, Rats, Inbred Lew **+42.86**, and Mice, Inbred ICR **+31.18**). These encouraging results indicate we should continue exploring how to better use the full text for automated indexing of MEDLINE citations.

Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Mork J, Jimeno Yepes A, Aronson A: The NLM Medical Text Indexer System for Indexing Biomedical Literature. 2013 BioASQ Workshop. Valencia, Spain, September 2013.
2. Jimeno-Yepes A, Mork J, Fushman D, Aronson A. Automatic algorithm selection for MeSH Heading indexing based on meta-learning. In Proceedings of the Fourth International Symposium on Languages in Biology and Medicine. 2011.
3. Xu R, Garten Y, Supekar KS, Das AK, Altman RB, Garber AM. Extracting subject demographic information from abstracts of randomized clinical trial reports. Stud Health Technol Inform. 2007;129(Pt 1):550-4.
4. de Bruijn B, Carini S, Kiritchenko S, Martin J, Sim I. Automated information extraction of key trial design elements from clinical trial publications. AMIA Annu Symp Proc. 2008 Nov 6:141-5.
5. Kiritchenko S1, de Bruijn B, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inform Decis Mak. 2010 Sep 28;10:56.
6. Zhao J, Kan MY, Procter PM, Zubaidah S, Yip WK, Li GM. Improving Search for Evidence-based Practice using Information Extraction. AMIA Annu Symp Proc. 2010 Nov 13;2010:937-41.
7. Kelly C, Yang H. A system for extracting study design parameters from nutritional genomics abstracts. J Integr Bioinform. 2013 Apr 4;10(2):222.
8. Gerner M, Nenadic G, Bergman CM. LINNAEUS: a species name identification system for biomedical literature. BMC Bioinformatics. 2010 Feb 11;11:85.
9. Naderi N, Kappler T, Baker CJ, Witte R. OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. Bioinformatics. 2011 Oct 1;27(19):2721-9.
10. Pafilis E, Frankild SP, Fanini L, Faulwetter S, Pavloudi C, Vasileiadou A, Arvanitidis C, Jensen LJ. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. PLoS ONE. 2013 8(6): e65390.
11. Pyysalo S, Ohta T, Miwa M, Cho HC, Tsujii J, Ananiadou S. Event extraction across multiple levels of biological organization. Bioinformatics. 2012 Sep 15;28(18):i575-i581.
12. Jimeno-Yepes AJ, Plaza L, Mork JG, Aronson AR, Diaz A. MeSH indexing based on automatically generated summaries. BMC Bioinformatics. 2013 Jun 26;14:208.
13. Features of the MeSH Vocabulary. Available at: http://www.nlm.nih.gov/mesh/features2003.html Date accessed: March 10, 2015
14. Datasets & Test Collections. Available at: http://ii.nlm.nih.gov/DataSets/index.shtml Date accessed: March 10, 2015
15. PMC Open Access Subset. Available at: http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ Date accessed: March 10, 2015
16. Demner-Fushman D, Lin J. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. Computational Linguistics. 2007;33(1):63-103.
17. Demner-Fushman D, Abhyankar S. Syntactic-Semantic Frames for Clinical Cohort Identification Queries. DILS June 2012, LNBI 7348 proceedings. Lecture Notes in Computer Science. 2012;7348:100-112
18. UMLS Current Semantic Types. Available at: http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html Date accessed: March 10, 2015
19. PubMed Help [Internet]. Available at: http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Ages Date accessed: March 10, 2015.