Automatically Linking Registered Clinical Trials to their Published Results with Deep Highway Networks

Travis R. Goodwin, MS[1], Michael A. Skinner, MD[1,2],
Sanda M. Harabagiu, PhD[1]
[1]The University of Texas at Dallas, Department of Computer Science, Richardson, TX;
[2]The University of Texas Southwestern Medical Center, Department of Surgery, Dallas, TX

Abstract

As medical science continues to advance, health care professionals and researchers are increasingly turning to clinical trials to obtain evidence supporting best-practice treatment options. While clinical trial registries such as ClinicalTrials.gov aim to facilitate these needs, it has been shown that many trials in the registry do not contain links to their published results. To address this problem, we present NCT Link, a system for automatically linking registered clinical trials to published MEDLINE articles reporting their results. NCT Link incorporates state-of-the-art deep learning and information retrieval techniques by automatically learning a Deep Highway Network (DHN) that estimates the likelihood that a MEDLINE article reports the results of a clinical trial. Our experimental results indicate that NCT Link obtains 30%-58% improved performance over previously reported automatic systems, suggesting that NCT Link could become a valuable tool for health care providers seeking to deliver best-practice medical care informed by evidence of clinical trials as well as (a) researchers investigating selective publication and reporting of clinical trial outcomes, and (b) study designers seeking to avoid unnecessary duplication of research efforts.

Introduction

Seeking to deliver best-practice medical care, clinicians increasingly rely on information provided by published guidelines and systematic reviews. However, recent analyses have estimated that less than 15 percent of major medical guidelines are supported by high-quality evidence.[1,2] To bridge this gap, health care professionals are increasingly turning to evidence from clinical trials to help evaluate different treatment options.[3] To provide more convenient access to clinical trials for persons with serious medical conditions and to make the results of clinical trial more available to health care providers, the United States Congress mandated in 1997 the development of the online trial registry ClinicalTrials.gov. In 2007, in accordance with the increasing role of evidence-based medicine, the mandate was expanded by requiring the timely inclusion of clinical trial results within the registry for all sponsors of non-phase-1 human trials seeking FDA approval for a new device or drug.[4] Moreover, to further increase the availability of study information to patients, physicians, and investigators, the International Committee of Medical Journal Editors (ICMJE) mandated the registration of trials before considering publication of trial results.[3]

Unfortunately, despite the numerous policies intended to improve the timely accessibility of clinical trial results to clinicians, there remain several barriers hindering effective use of these important data. First, sponsors and investigators have inconsistently complied with the requirement to update the registry with trial results. In an evaluation of eligible human studies registered at ClinicalTrials.gov, Anderson et al. (2015)[5] found that only 13.4% of the trials reported summary results within 12 months of study completion, and only 38.3% of the registered studies reported any results at any time. Moreover, once trial results are published in peer-reviewed literature, the article citation is only provided to the ClinicalTrials.gov registry in about 23%-31% of cases.[6,7] When registered trials with no reported publications were manually reviewed, both Ross et al. (2009)[6] as well as Huser and Cimino (2013)[7] were able to find relevant MEDLINE articles for 31%-45% of reviewed clinical trials. Finally, despite the ICMJE recommendation that pertinent publications of trial results should contain a specific citation of the trial registry number to allow simple retrieval of the article with a MEDLINE search,[8] this information is included in only about 7% of articles presenting trial results.[7]

Recently, Bashir et al. (2017)[9] conducted a systematic review of studies examining "links" between registered clinical trials and the publications reporting their results and found that 83% of studies required some level of manual (i.e., human) analysis (with 19% involving strictly manual analyses, 64% involving both manual and automatic analyses and 17% involving automatic analyses). They also observed that despite the increasing pressures from journal editors to provide information about any clinical trials associated with a publication, the number of articles amenable to being automatically linked to the clinical trials they report has not increased over time. Finally, they found that automatic methods were only able to identify a median of 23% of articles reporting the results of registered trials, leading them to conclude that identifying publications reporting the results of a clinical trial remains an arduous, manual task. Clearly, there is a need for the creation of robust methods to automatically link clinical trials with their results in the medical literature.

In this paper, we present NCT Link,[*] a system for automatically linking registered clinical trials to articles reporting their results. NCT Link incorporates state-of-the-art deep learning techniques through a specialized Deep Highway Network (DHN)[10] designed to determine the likelihood that a link exists between an article and a clinical trial by considering the variety of information about the article, the trial, and the relationships (if any) between them. Our experiments demonstrate that NCT Link provides a 30%-58% improvement over the automatic methods surveyed in Bashir et al. (2017);[9] consequently, we believe that NCT Link will provide a valuable tool for health care providers seeking to obtain timely access to the publications reporting the results of clinical trials. Moreover, we surmise that NCT Link may also benefit (a) researchers investigating selective publication and reporting of clinical trial outcomes and (b) study designers aiming to avoid unnecessary duplication of research efforts.[3]

Linking Registered Clinical Trials to their Published Results

In previous studies examining links between registered clinical trials and published articles, investigators have described at least three ways that a published article may be considered *linked* to a clinical trial. For example, an article may (1) relate in some way to the trial, e.g., by providing supporting evidence for the intervention or highlighting limitations of previous, related studies; (2) be cited in the summary or official description of the trial; or (3) report the results of the trial. In this work, we focus exclusively on the third type of link: articles which report the results of a clinical trial; consequently, we consider a publication to be *linked* to a clinical trial *if and only if* it reports the results of the trial. Moreover, as in Huser and Cimino (2013),[7] we only consider links between clinical trials registered to ClinicalTrials.gov and published articles indexed by MEDLINE.

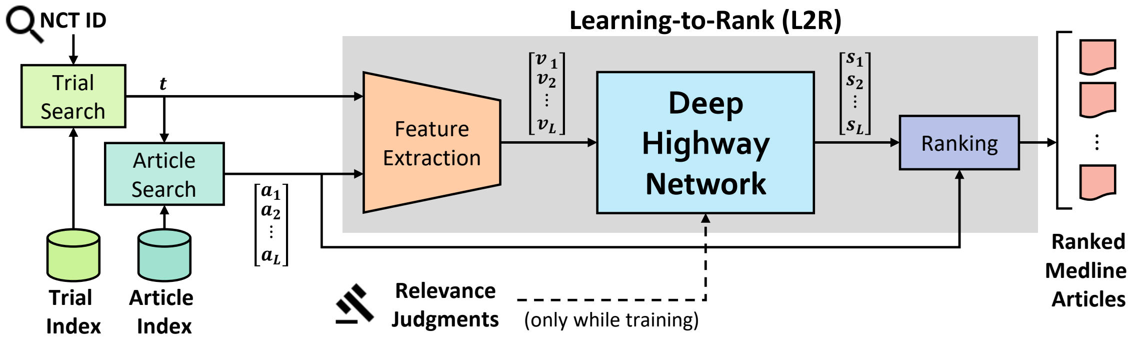


Figure 1: Architecture of NCT Link.

NCT Link, illustrated in Figure 1, operates in five steps:

1. **Trial Search**: given an NCT ID, the (meta)data associated with the trial, denoted as $\boldsymbol{t}$, is obtained from the registry at ClinicalTrials.gov;
2. **Article Search**: the information in $\boldsymbol{t}$ is used to obtain a subset of potentially-linked articles (along with their metadata), denoted as $A = \boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_L$, using a specialized local MEDLINE index (where L is the maximum number of articles considered by NCT Link);
3. **L2R: Feature Extraction**: each article $\boldsymbol{a}_i \in A$ retrieved for $\boldsymbol{t}$ is associated with a feature vector $\boldsymbol{v}_i$ encoding a number of complex features characterizing information about $\boldsymbol{t}$, $\boldsymbol{a}_i$, and the relationship between them;
4. **L2R: Deep Highway Network**: a Deep Highway Network (DHN) is used to infer a *score* s_i for each article $\boldsymbol{a}_i \in A$ quantifying the likelihood that $\boldsymbol{a}_i$ should be linked to (i.e. reports the results of) $\boldsymbol{t}$;

[*] NCT Link is named after the Clinical Trial identifier, NCT ID, used by ClinicalTrials.gov.

5. **L2R: Ranking**: the score s_i associated with each article $\boldsymbol{a}_i$ is used to produce a ranked list of published articles such that the rank of each article corresponds to the likelihood that it reports the results of $\boldsymbol{t}$.

In the remainder of this section, we provide a detailed description each of the five steps listed above.

A. Searching Clinical Trials

NCT Link operates on an NCT ID specified by the user. The NCT ID is used to obtain all the (meta)data stored in the ClinicalTrials.gov registry for the given trial. While the National Library of Medicine (NLM) provides an online interface for programmatically obtaining data about a clinical trial specified by an NCT ID, to reduce the burden on the NLM's servers potentially imposed by our experiments, we instead created and used our own offline index of all clinical trials registered on ClinicalTrials.gov.

Representing clinical trials. Due to the significant variation in the amount of data associated with clinical trials, NCT Link considers only eight key *aspects* of each clinical trial: (1) the set of investigators[*] associated with the trial, (2) the set of unique institutions associated with any investigators, (3) the NCT ID of the trial, (4) the set of interventions studied in the trial, (5) the set of conditions studied in the trial, (6) the set of keywords provided to the registry, (7) the set of Medical Subject Headings (MeSH) terms provided to the registry, and (8) the completion date of the trial[†]. In the remainder of this paper, we use $\boldsymbol{t}$ to simultaneously refer to a clinical trial as well as all eight aspects of information associated with the trial.

B. Searching MEDLINE Articles

Because MEDLINE contains over 14 million articles, rather than applying the learning-to-rank component to process and score every article in MEDLINE, we first obtain a smaller, "high-recall" sub-set of candidate MEDLINE articles that are likely to report the results of $\boldsymbol{t}$. In this section, we describe the MEDLINE searching strategy used for both (1) obtaining this high-recall set of candidate MEDLINE articles as well as (2) feature extraction (described later).

Indexing MEDLINE articles. To search MEDLINE, NCT Link incorporates its own internal, offline index of every article in MEDLINE. This index encodes eight *fields* (i.e., metadata attributes) for each article in MEDLINE: (1) the authors[‡]of the article (if any), (2) the investigators[‡] of the article (if any), (3) the PubMed identifier (PMID) associated with the art icle, (4) the accession numbers (e.g. NCT IDs) of any ClinicalTrials.gov entries in the list of "DataBanks" associated with the article, (5) the full unstructured text of the abstract[§], (6) the title of the article, (7) any MeSH terms associated with the article, and (8) the publication date of the article.

Query formulation. A clinical trial $\boldsymbol{t}$ is represented by a disjunctive Boolean query in which each aspect corresponds to a clause. Each clause, in turn, is represented by a disjunction of natural language terms (or phrases) encoding the values (e.g., investigators, conditions, etc.) associated with that aspect. Interventions, conditions, and keywords are expanded using synonyms provided by the Unified Medical Language System.[11] To account for variations in the way affiliations were expressed, each affiliation was represented by a sequence of "partial locations" by splitting the text of the affiliation (e.g., "University of California, San Francisco") on occurrences of commas (e.g., "University of California" and "San Francisco"). Likewise, due to differences in how authors and investigators are reported to MEDLINE by various journals, each author/investigator is represented by a sieve consisting of four queries, each less specific than the previous: (1) first name, middle initial, and last name, (2) first initial, middle initial and last name, (3) first name and last name, and (4) the first initial and last name. To account for the progressive loss of specificity, we associated each query with a weight of 1.0, 0.5, 0.3, and 0.2, respectively, which multiplicatively affects the score (described below) of any article retrieved for the query. The clause associated with each aspect is restricted to the set of semantically related fields illustrated in Figure 2.

Scoring MEDLINE articles. When searching our internal MEDLINE index, candidate articles are retrieved (i.e.

[*] Investigators are represented in the registry through three structured fields indicating the investigator's first, middle, and last names.

[†] If the completion date is unspecified, or if the trial is not yet complete, the start date of the trial is used.

[‡] In MEDLINE, authors and investigators are encoded by structured fields corresponding to their first and last names as well as their initials.

[§] For structured abstracts, the content of all sections was combined to create a single unstructured passage of text.

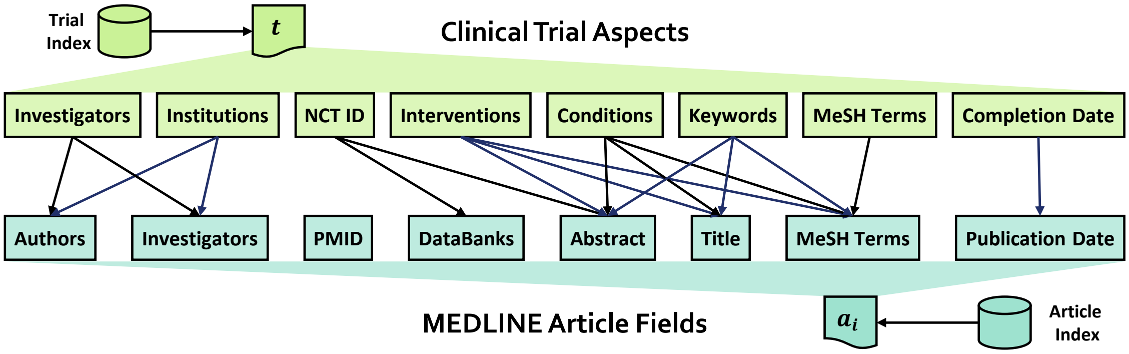


Figure 2: The aspects of medical trials, the fields indexed for MEDLINE articles, and the mapping between them when searching MEDLINE articles. Note: although MEDLINE distinguishes between investigators and authors of published articles, NCT Link currently treats the authors and investigators of MEDLINE articles in the same way.

selected) and scored using the BM25[12] relevance model.[*] This allows the high-recall set of candidate articles A to be defined as the top ranking retrieved articles $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_L$ where L is the number of candidate MEDLINE articles considered by NCT Link. Conceptually, L acts as an upper bound on the number of articles the user of the system might be interested in examining. In our experiments, to ensure thorough evaluations, we used $L = 2,000$. However, for general use, a smaller value of L should be sufficient, e.g., $L = 100$.

C. Learning-to-Rank (L2R)

Given a clinical trial $\boldsymbol{t}$, and a set of articles $A = \boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_L$, the learning-to-rank module is responsible for (1) extracting features encoding the relationship between each article $\boldsymbol{a}_i$ and the clinical trial $\boldsymbol{t}$; (2) training (or using) state-of-the-art deep learning methods – a Deep Highway Network – to score each article based on the likelihood that it reports the results of $\boldsymbol{t}$; and (3) produce a ranked list of MEDLINE articles sorted by their scores.

D. Feature Extraction from MEDLINE Articles and Clinical Trials

Determining whether a link exists between an article $\boldsymbol{a}_i$ and a clinical trial $\boldsymbol{t}$ requires considering a large variety of information which varies from trial to trial and article to article. For this reason, deciding whether an article $\boldsymbol{a}_i \in A$ reports the results of $\boldsymbol{t}$ requires access to a rich set of features. When extracting features, we consider (1) the eight aspects of clinical trial $\boldsymbol{t}$ (described in *Searching Clinical Trials*), (2) the eight fields associated with article $\boldsymbol{a}_i$ (described in *Searching MEDLINE Articles*), and (3) the mapping between aspects of $\boldsymbol{t}$ and the corresponding fields of $\boldsymbol{a}_i$ illustrated in Figure 2. Table 1 lists all the features extracted for each article $\boldsymbol{a}_i$ retrieved for trial $\boldsymbol{t}$ as well as the *domain* (i.e. number and type of values) of each feature, where $\mathbb{N}$ denotes the set of natural numbers, $\mathbb{R}$ denotes the set of real numbers, and the exponent indicates the number of values (e.g. $\mathbb{R}^5$ corresponds to five distinct real numbers).

As shown in Table 1, three types of features are extracted: (1) trial features (F_1 - F_4), encoding information about $\boldsymbol{t}$ which is independent of $\boldsymbol{a}_i$; (2) dynamic features (F_5 - F_{29}), encoding information about the relationship between $\boldsymbol{a}_i$ and $\boldsymbol{t}$; and (3) article features (F_{30} - F_{33}), encoding information about $\boldsymbol{a}_i$ which is independent of $\boldsymbol{t}$. Features F_1 - F_3 allow the model to account for that fact that the more investigators, interventions, or conditions associated with $\boldsymbol{t}$, the more likely it is that an article will have an investigator, intervention, or condition in common with $\boldsymbol{t}$. Features F_6 - F_{29} adapt four commonly used *relevance models* to act as similarity measures between an aspect of $\boldsymbol{t}$ and an article $\boldsymbol{a}_i$. Specifically, we used: (1) the Best Match 25[12] (BM25), (2) Dirichlet-Smoothed language model probability[13] (LMD), (3) Axiomatic relevance[14] (F2Exp), and (4) Divergence from Independence[15] (DFI). To account for the significant variance in the number of investigators, as well as the prevalence of common names, conditions, or interventions, F_{18} - F_{29} measure five *statistics* capturing the similarity between ***each*** investigator, condition, or

[*] To reduce the impact of abstract length on the ranking of candidate MEDLINE articles, we specified the BM25 document-length normalization term as $k_1 = 0.25$ rather than standard value of 0.75.[12]

Table 1: Features extracted for each article $\boldsymbol{a}_i$ retrieved for trial $\boldsymbol{t}$.

	Feature Description	Domain
F_1	number of **investigators** in $\boldsymbol{t}$	$\mathbb{N}$
F_2	number of **interventions** in $\boldsymbol{t}$	$\mathbb{N}$
F_3	number of **conditions** in $\boldsymbol{t}$	$\mathbb{N}$
F_4	**completion date** of $\boldsymbol{t}$	$\mathbb{N}$
F_5	days elapsed between the publication date of $\boldsymbol{a}_i$ and the completion date of $\boldsymbol{t}$	$\mathbb{N}$
F_6	BM25 from the **NCT ID** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_7	F2Exp from the **NCT ID** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_8	DFI from the **NCT ID** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_9	LMD from the **NCT ID** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_{10}	BM25 from all **keywords** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_{11}	F2Exp from all **keywords** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_{12}	DFI from all **keywords** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_{13}	LMD from all **keywords** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_{14}	BM25 from all **MeSH terms** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_{15}	F2Exp from all **MeSH terms** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_{16}	DFI from all **MeSH terms** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
F_{17}	LMD from all **MeSH terms** of $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}$
⋮	⋮	⋮
F_{18}	BM25 *statistics* from each **investigator** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{19}	F2Exp *statistics* from each **investigator** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{20}	DFI *statistics* from each **investigator** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{21}	LMD *statistics* from each **investigator** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{22}	BM25 *statistics* from each **intervention** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{23}	F2Exp *statistics* from each **intervention** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{24}	DFI *statistics* from each **intervention** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{25}	LMD *statistics* from each **intervention** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{26}	BM25 *statistics* from each **condition** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{27}	F2Exp *statistics* from each **condition** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{28}	DFI *statistics* from each **condition** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{29}	LMD *statistics* from each **condition** in $\boldsymbol{t}$ to $\boldsymbol{a}_i$	$\mathbb{R}^5$
F_{30}	number of **authors** in $\boldsymbol{a}_i$	$\mathbb{N}$
F_{31}	number of **investigators** in $\boldsymbol{a}_i$	$\mathbb{N}$
F_{32}	**publication date** of $\boldsymbol{a}_i$	$\mathbb{N}$
F_{33}	**publication type**(s) of $\boldsymbol{a}_i$	$\{0,1\}^{38}$

intervention in $\boldsymbol{t}$ and $\boldsymbol{a}_i$, namely, the mean, minimum, maximum, variance, and sum. Feature F_{33} encodes the MeSH publication type(s) associated with $\boldsymbol{a}_i$ (in our experiments, we encountered only 38 different types of publications). The values features F_1 - F_{33} are concatenated together to form a single vector $\boldsymbol{v}_i$ allowing the Deep Highway Network to consider and combine a variety of different interactions between the aspects of $\boldsymbol{t}$ and the fields of article $\boldsymbol{a}_i$.

E. The Deep Highway Network

Owing to the lack of clear and exact criteria for determining whether a link exists between $\boldsymbol{t}$ and $\boldsymbol{a}_i$, we were interested in applying deep learning techniques to automatically learn contextual high-level and expressive "meta"-features by combining the elements of $\boldsymbol{v}_i$. However, a common problem when designing deep learning networks is that the there are no clear criteria or guidelines for deciding the number (and configuration) of internal or "deep" layers in the network. Fortunately, by taking advantage of recent advances in deep structure learning, we were able to define a deep neural network which automatically tunes the number of internal layers used. Specifically, we implemented a Deep Highway Network[16] (DHN). Unlike traditional deep networks, in which information flows through each layer of the network sequentially, DHNs allow information to "skip" layers in the network by traveling along a so-called "information highway"[*]. Thus, in a DHN, the number of specified internal layers acts as an upper bound on the number of layers used by the model. In fact, the information highway allows DHNs to be constructed with hundreds of intermediate layers – for example DHNs with over 1,000 intermediate layers have been reported.[10] The DHN we have implemented within NCT Link considers a maximum of 10[†] internal layers and is illustrated in Figure 3.

As shown, the main component of each intermediate layer, l, is a Rectified Linear Unit (ReLU),[17] i.e.,

$$\boldsymbol{x_{l+1}} = \mathrm{ReLU}(\boldsymbol{x}_l) = \max\left(\boldsymbol{x}_l, \mathbf{0}\right),$$

where $\boldsymbol{x}_l$ indicates the output of layer l and $\mathbf{0}$ denotes a zero-vector. In the DHN, each ReLU layer is augmented with a highway mechanism composed of two gates: (1) a *transform* gate, $T \in [0, 1]$, which learns a weight that is applied to the output of the ReLU, and (2) a *carry* gate, $C = 1 - T$, which learns whether to skip, apply, or partially apply

[*] Additionally, and perhaps more importantly, the information highway allows the gradient to directly influence each layer during back propagation, effectively eliminating the vanishing gradient problem and allowing very deep networks to be trained.

[†] We also experiment with 100 internal layers and observed no discernible change in performance.

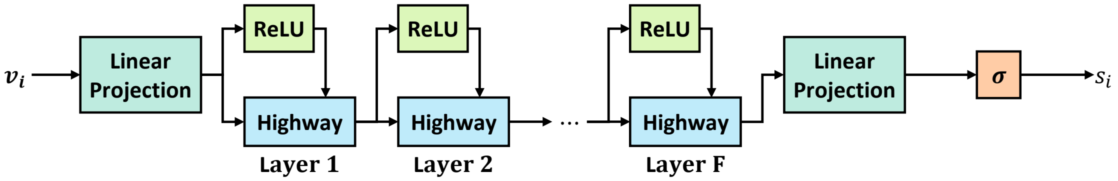
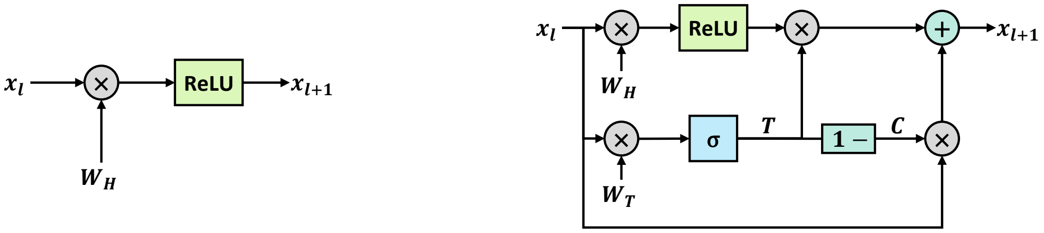


Figure 3: Architecture of the Highway Network used in NCT Link.

the ReLU in the layer to $\boldsymbol{x}_l$. Thus, the highway mechanism enables the network to learn how many (and which) layers should be applied. Formally, we define each layer in our DHN as follows:

$$T(\boldsymbol{x}_l) = \sigma\left(\boldsymbol{x}_l \cdot \boldsymbol{W}_T^{(l)}\right) \qquad \boldsymbol{x}_{l+1} = T(\boldsymbol{x}_l) \cdot \text{ReLU}\left(\boldsymbol{W}_H^{(l)} \cdot \boldsymbol{x}_l\right) + (1 - T(\boldsymbol{x}_l)) \cdot \boldsymbol{x}_l$$

where $\boldsymbol{W}_T^{(l)}, \boldsymbol{W}_H^{(l)} \in \boldsymbol{\theta}$ correspond to the learned weights of the transform gate and ReLU used in layer l. Figure 4 illustrates the difference between a standard ReLU layer and a ReLU layer incorporating a highway mechanism.



(a) ReLU without a highway mechanism.

(b) ReLU with a highway mechanism.

Figure 4: Comparison of ReLU layers with and without highway mechanisms.

To produce the score $s_i \in [0, 1]$ (i.e., the likelihood that $\boldsymbol{a}_i$ reports the results of $\boldsymbol{t}$) associated with $\boldsymbol{a}_i$, the output of the final highway layer is projected down to a single element projected into the range $[0, 1]$ by a sigmoid layer:

$$s_i = \sigma\left(\boldsymbol{W}_s \cdot \boldsymbol{x}_F\right)$$

where $\sigma(x) = {}^{e^{\boldsymbol{x}}}/_{e^{\boldsymbol{x}}+1}$ is the logistic sigmoid function, $\boldsymbol{W}_s \in \boldsymbol{\theta}$ corresponds to the learned weights of the final project layer, and $\boldsymbol{x}_F$ indicates the output of the final ReLU layer.

Training the Deep Highway Network. Training the DHN was achieved by finding the parameters $\boldsymbol{\theta}$ most likely to predict the *correct* score s_i for every article $\boldsymbol{a}_i \in A$ retrieved for each clinical trial $\boldsymbol{t}$ in the training set $\mathcal{T}$ (details about the training set and relevance judgments used in our experiments are provided in the *Experiments* section). Formally, let $y_{t,i}$ indicate the relevance judgment of article $\boldsymbol{a}_i$ with respect to trial $\boldsymbol{t}$ such that $y_{t,i} = 1$ if $\boldsymbol{a}_i$ is relevant to (i.e., reports the results of) $\boldsymbol{t}$ and $y_{t,i} = 0$, otherwise. We minimize the entropy loss between the score s_i assigned by the DHN and the relevance judgment $y_{t,i}$:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{(\boldsymbol{t},L) \in \mathcal{T}} \left(\sum_{i=1}^{L} s_i \log y_{t,i} + (1 - s_i) \log (1 - y_{t,i}) \right)$$

The model was trained using Adaptive Moment Estimation[18] (ADAM) (using the default initial learning rate $\eta = 0.001$).

F. Ranking MEDLINE Articles

After producing a score s_i for each article $\boldsymbol{a}_i \in A$ retrieved for trial $\boldsymbol{t}$, the final ranked list of articles is produced by sorting the articles $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_L$ in descending order according to their scores $s_1, s_2, \cdots, s_L$.

Experiments

Each clinical trial in ClinicalTrials.gov was manually registered by a Study Record Manager (SRM) and may be associated with two types of publications corresponding to distinct fields in the registry: (1) "related articles", articles the

SRM deemed related to the trial (typically references) and (2) "result articles", articles the SRM indicated as reporting the results of the trial. To evaluate NCT Link, we randomly selected 500 clinical trials which were each associated with at least one "result article" in the registry. In our experiments, we used a standard 3:1:1 split for training, development, and testing. Relevance judgments for all 500 trials were automatically produced using the "result articles" encoded for each trial. Specifically, for each trial $\boldsymbol{t}$, we assigned a judgment of RELEVANT to all MEDLINE articles listed as "result articles" for $\boldsymbol{t}$. We considered two strategies for producing IRRELEVANT judgments. Initially, we applied the Closed World Assumption[19] (CWA) by judging every MEDLINE article not explicitly listed in the "result articles" of $\boldsymbol{t}$ as IRRELEVANT to $\boldsymbol{t}$. We refer to this judgment strategy as CLOSED.

However, it has been shown that the SRM of a clinical trial does not always update the registry as new articles are published.[9] Under the CWA, these articles would be mistakenly labeled IRRELEVANT. To account for this, we considered a secondary judgment strategy intended to minimize the likelihood of assigning an IRRELEVANT judgment to a MEDLINE article that may report the results of a trial despite not being included in the "result articles" of the trial. Formally, for each trial $\boldsymbol{t}$ we obtained a list A of $3,000$ MEDLINE articles using the search strategy described in the *Searching MEDLINE Articles* section (without applying the learning-to-rank component of NCT Link). We determined the set of IRRELEVANT articles for $\boldsymbol{t}$ as: (1) articles which were not listed in the "result articles" of $\boldsymbol{t}$ but were listed in the "result articles" of any other trial in the registry, (2) 10 randomly selected articles between ranks 10 and 100, (3) 10 randomly selected articles between ranks 1000 and 2000, (3) 10 randomly selected articles between ranks 2000 and 3000, and (4) 10 randomly selected articles from MEDLINE not in A. We refer to this second judgment strategy as OPEN. We report the performance of NCT Link when trained using the OPEN strategy and evaluated using both strategies*.

Due to the paucity of published automatic systems for linking clinical trials to their results in the literature, we measured the performance of NCT Link against two baseline systems as well as four alternative configurations of NCT Link:

1. **Exact Match:** A system in which an article is considered to be linked to a clinical trial if it specifically mentions the NCT ID of the trial in its abstract or metadata – this is an extension of the automatic approach described by Bashir et al. (2017)[9] which considers only metadata.
2. **IR:BM25:** An information retrieval (IR) system which represents all aspects of the clinical trial as a single disjunctive Boolean query relying on the BM25 similarity function.
3. **NCT Link: BM25** A configuration of NCT Link in which no learning-to-rank is performed; that is, the system returns the ranked list of candidate articles described in the *Searching MEDLINE Articles* section.
4. **NCT Link:Linear Regression.** A configuration of NCT Link that replaces the Deep Highway Network (DHN) with a linear regression model to determine article scores.
5. **NCT Link:Random Forests.** A configuration of NCT Link that replaces the DHN with a Random Forest[20] model to determine article scores.
6. **NCT Link:Gradient Boosting.** A configuration of NCT Link that replaces the DHN with a Gradient Boosting[21] model to determine article scores; Gradient Boosting can be viewed as modern extension to Random Forests that incorporates boosting rather than bagging to combine the scores predicted by each decision tree in the forest.

Quality Metrics. We measured the quality of ranked MEDLINE articles produced by all systems using standard metrics for evaluating the performance of information retrieval systems. Formally, let $\mathcal{X}$ indicate the test set, consisting of pairs of a clinical trial, $\boldsymbol{t}$, and the final ranked list of L articles produced for $\boldsymbol{t}$, $B_{1:L}$. To measure the overall ranking produced by each system, we measured the Mean Average Precision (MAP):

$$\mathrm{MAP}\left(\mathcal{X}\right) = \sum_{(\boldsymbol{t},B_{1:L}) in \mathcal{X}} \mathrm{AP}\left(B_{1:L};\boldsymbol{t}\right); \quad \mathrm{AP}\left(B_{1:L};\boldsymbol{t}\right) = \frac{1}{L}\sum_{k=1}^{L}\left(\mathrm{P}\left(B_{1:k};\boldsymbol{t}\right)\cdot \mathrm{Rel}\left(\boldsymbol{b}_k;\boldsymbol{t}\right)\right); \quad \mathrm{P}\left(B_{1:k};\boldsymbol{t}\right) = \sum_{k=1}^{K}\frac{\mathrm{Rel}\left(\boldsymbol{b}_k;\boldsymbol{t}\right)}{\mathrm{Num\,Rel}\left(\boldsymbol{t}\right)}$$

where $\mathrm{AP}\left(B_{1:L},\boldsymbol{t}\right)$ indicates the Average Precision of $B_{1:L}$ with respect to $\boldsymbol{t}$, $\mathrm{P}\left(B_{1:k};\boldsymbol{t}\right)$ represents the precision of the top-K ranked articles retrieved for trial $\boldsymbol{t}$, $\mathrm{Rel}\left(\boldsymbol{b}_k;\boldsymbol{t}\right)$ is an indicator function returning the value 1 if article $\boldsymbol{b}_k$ was judged as RELEVANT for trial $\boldsymbol{t}$ and returning 0, otherwise, and $\mathrm{Num\,Rel}\left(\boldsymbol{t}\right)$ returns the number of articles judged RELEVANT for $\boldsymbol{t}$. In addition to the MAP, we report the Mean Reciprocal Rank (MRR) which is the average of the multiplicative inverse of the rank of the first relevant article produced for each trial. The MRR captures how many

*We found that training with the CLOSED strategy degraded performance on the test set in all cases.

irrelevant reports are ranked, on average, above the first relevant article for each trial. We also report the average precision over all clinical trials at three different ranks: the R-Precision (R-Prec) which is the precision of the first R-ranked articles, where for each trial $\boldsymbol{t}$, $R = \operatorname{Num}\operatorname{Rel}(\boldsymbol{t})$; the Precision of the top-five ranked articles (P@5) and the Precision of the top-ten ranked articles (P@10).

Table 2: Quality of ranked list of MEDLINE articles retrieved for each clinical trial.

	CLOSED Strategy					OPEN Strategy				
System	**MAP**	**MRR**	**R-Prec**	**P @ 5**	**P @ 10**	**MAP**	**MRR**	**R-Prec**	**P @ 5**	**P @ 10**
Exact Match	0.001	0.001	0.0000	0.0000	0.000	0.236	0.260	0.2203	0.0620	0.031
IR: BM25	0.011	0.016	0.0032	0.0040	0.003	0.258	0.294	0.1793	0.1220	0.095
NCT Link: BM25	0.017	0.021	0.002	0.004	0.004	0.586	0.610	0.549	0.210	0.115
NCT Link: Linear Regression	0.269	0.302	0.236	0.082	0.046	0.656	0.723	0.620	0.264	0.161
NCT Link: Random Forests	0.196	0.219	0.154	0.072	0.051	0.734	0.808	0.709	0.298	0.185
NCT Link: Gradient Boosting	0.143	0.162	0.102	0.046	0.030	0.717	0.791	0.684	0.282	0.182
⋆**NCT Link: DHN**	**0.308**	**0.342**	**0.244**	**0.123**	**0.082**	**0.824**	**0.873**	**0.920**	**0.358**	**0.221**

Results. Table 2 depicts the performance of all baseline systems as well as all configurations of NCT link when evaluated according to both judgment strategies and measured with all five metrics. As expected, all systems obtained higher performance when using the OPEN judgment scheme than when using the CLOSED scheme. The poorer performance of all systems when using the CLOSED judgment scheme supports the notion that many relevant articles were incorrectly labeled as IRRELEVANT. Consequently, the OPEN judgment scheme may be viewed as an upper bound while the CLOSED judgment scheme may be viewed as a lower bound of each system's performance. Regardless of judgment scheme, NCT Link using the Deep Highway Network (DHN) obtains the highest performance, followed by the three other NCT Link configurations employing learning-to-rank. It is interesting to note that the complexity of these three models coincided with an increase in performance when using the OPEN judgment scheme and a decrease in performance when using the CLOSED judgment schemes. This indicates that the more complex models may have over-fit on the OPEN judgment scheme used for training. The lowest performance was exhibited by the exact match baseline, reinforcing the observations reported by Bashir et al. (2017)[9] that considering the NCT ID alone is not sufficient to determine links between MEDLINE articles and clinical trials. Likewise, the disparity in performance between the basic information retrieval system (IR: BM25) and the BM25 configuration of NCT Link clearly indicates that the search criteria described in the *Searching MEDLINE Articles* section obtains higher quality results than when using a naïve retrieval strategy. Moreover, the increase in performance when incorporating learning-to-rank within NCT Link suggests that the features extracted by NCT Link are able to capture useful criteria for determining whether a link exists between an article and a trial. When comparing the performance of NCT Link using the DHN against the performance when using Random Forest, Linear Regression, and Gradient Boosting, it is clear that the DHN obtains superior performance, suggesting that our DHN is able to successfully extract "meta"-features capturing additional semantics about the relationship between a MEDLINE article and a clinical trial.

Discussion

We manually analyzed the MEDLINE articles retrieved by NCT Link for 30 clinical trials in test set and found four main sources of error.

The most common source of errors we observed was the result of investigator and author names. Specifically, we found that, in general, clinical trials represented investigator names with three fields: first name, middle name, and last name. However, many journals in MEDLINE only report the authors' last names and the initials of first and sometimes middle names. This resulted in scenarios in which the system incorrectly concluded that the investigator of a trial was the same as the author of a paper. This error was most prevalent for common last names (e.g., *Lin*, *Brown*), common first initials (e.g., *J*, *M*, *S*, *D*), and when the middle initial was unspecified. Moreover, we observed that in several cases, the first and middle names in the clinical trial registry were blank and the last name contained the full name of the primary investigator. In four cases case, the first and middle names were blank and the last name appeared to refer to the sponsoring company. In future work, we believe some of these errors could be at least partially addressed by incorporating some degree of citation analysis to help (1) disambiguate initials and/or (2) infer unspecified names from previous work. The second most common source of errors were mismatched affiliations. We found many cases in which the same institution was referenced in multiple ways, (e.g. “UCLA” and “University of California, Los Angeles”). Moreover, addresses were often specified with different levels of detail (street names, cities, states, country). Unfortunately, resolving this kind of ambiguity is a difficult problem involving world, spatial, and geographical knowledge as well as prior knowledge about known institutions and their standard abbreviations.

The third most common source of errors was inconsistencies in the way clinical trial completion dates were provided to the registry. Because the completion date was represented in natural language, completion dates were represented in a wide variety of formats. For example, some SRMs preferred formatting the dates in the European fashion (day-month-year), while others preferred the American notation (month-day-year). In some cases, only the month and the year and year were indicated. Individual months were specified using digits (e.g. “01”), the full name (e.g., “January”) as well as a variety of abbreviations (e.g., “J”, “Jan”, and “Jan.”). Years were specified in both two and four digit varieties (e.g., “07”, and “2007”). In our study, we investigated applying automatic tools for recognizing time expressions (e.g., SUTime[22]) but found it increased processing time by two orders of magnitude.

The final source of errors appears to result from SRMs providing incorrect information to the registry. We found cases in which the references provided as “result articles” for a clinical trial were published before the trial’s start date (in some cases, decades before). It is unclear whether incorrect citations were given, or whether there was confusion between the “related articles” and “result articles” fields in the registry.

In addition to the errors described above, there were some limitations in our experiments. First, we only considered the clinical trials registered on ClinicalTrials.gov despite the availability of other registries such as the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP)[*]. Second, we limited our system to considering only articles published on MEDLINE and did not consider other databases such as EMBASE[†] or research conference proceedings. Moreover, because MEDLINE itself only provides abstracts, NCT Link did not have access to the full text of articles. In future work, it may be advantageous to consider the full text of articles included in the PubMed Central (PMC) Open Access Subset[‡] (OAS); it should be noted, however, that the PMC OAS contains just over 1 million articles while MEDLINE itself contains over 14 million articles.

Implementation Details

NCT Link was implemented in both Java (version 8) and Python (version 2.7.13). Java was used for (1) parsing the data from MEDLINE as well as ClinicalTrials.org, relying on the Java Architecture for XML Binding (JAXB, version 2.1), (2) indexing and searching clinical trials and MEDLINE articles, relying on Apache Lucene[§] (version 6.6.0), and (3) feature extraction. The Deep Highway Network (DHN) was implemented in Python using Tensorflow[¶] (version 1.3). L2R baselines relied on the implementation provided by the RankLib component of the Lemur Project[‖] using the recommended parameters. Our DHN used 200-dimensional internal layers. When designing the network, we found that changing the dimensionality of internal layers had no discernible effects on performance.

Conclusion

In this paper, we have presented NCT Link, a system for automatically linking clinical trials to MEDLINE articles reporting their results. While traditional approaches for linking trials to their publications rely on arduous, manual analyses,[9] NCT Link learns to automatically determine the likelihood that a published article reports the results of a clinical trial by incorporating state-of-the-art deep learning and information retrieval techniques, obtaining a 30%-58% improvement over previously reported automatic systems.[9] These promising results suggest that NCT Link will provide a useful tool for clinicians seeking to provide timely, evidence-based care. Opportunities for future work include (1) citation analyses to resolve investigator and author names, (2) geo-spatial reasoning to resolve investiga-

[*] http://apps.who.int/trialsearch/

[†] https://www.elsevier.com/solutions/embase-biomedical-research

[‡] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[§] https://lucene.apache.org/

[¶] https://www.tensorflow.org/

[‖] https://www.lemurproject.org/

tor/author affiliations, (3) temporal expression normalization to account for variations in the way trial completion dates are expressed.

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC. Scientific evidence underlying the ACC/AHA clinical practice guidelines. Jama. 2009;301(8):831–841.
2. Lee DH, Vielemeyer O. Analysis of overall level of evidence behind Infectious Diseases Society of America practice guidelines. Archives of Internal Medicine. 2011;171(1):18–22.
3. De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al.. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. Mass Medical Soc; 2004.
4. Congress U. Food and Drug Administration Amendments Act of 2007. Public Law. 2007;p. 115–85.
5. Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J, Califf RM. Compliance with results reporting at ClinicalTrials. gov. New England Journal of Medicine. 2015;372(11):1031–1039.
6. Ross JS, Mulvey GK, Hines EM, Nissen SE, Krumholz HM. Trial publication after registration in ClinicalTrials. Gov: a cross-sectional analysis. PLoS medicine. 2009;6(9):e1000144.
7. Huser V, Cimino JJ. Linking ClinicalTrials. gov and PubMed to track results of interventional human clinical trials. PloS one. 2013;8(7):e68409.
8. International Committee of Medical Journal Editors (ICMJE), et al. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: writing and editing for biomedical publication. Haematologica. 2004;89(3):264.
9. Bashir R, Bourgeois FT, Dunn AG. A systematic review of the processes used to link clinical trial registrations to their published results. Systematic reviews. 2017;6(1):123.
10. Srivastava RK, Greff K, Schmidhuber J. Training very deep networks. In: Advances in neural information processing systems (NIPS); 2015. p. 2377–2385.
11. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl 1):D267–D270.
12. Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M, et al. Okapi at TREC-3. NIST Special Publication (SP). 1995;109:109.
13. Zhai C, Lafferty J. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS). 2004;22(2):179–214.
14. Fang H, Zhai C. An exploration of axiomatic approaches to information retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM; 2005. p. 480–487.
15. Kocabaş İ, Dinçer BT, Karaoğlan B. A nonparametric term weighting method for information retrieval based on measuring the divergence from independence. Information retrieval. 2014;17(2):153–176.
16. Srivastava RK, Greff K, Schmidhuber J. Highway networks. Deep Learning Workshop at the International Conference on Machine Learning (ICML). 2015;.
17. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: International Conference on Artificial Intelligence and Statistics; 2011. p. 315–323.
18. Kingma D, Ba J. Adam: A method for stochastic optimization. ICLR. 2015;.
19. Minker J. On indefinite databases and the closed world assumption. In: 6th Conference on Automated Deduction. Springer; 1982. p. 292–308.
20. Breiman L. Random forests. Machine learning. 2001;45(1):5–32.
21. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001;p. 1189–1232.
22. Chang AX, Manning CD. Sutime: A library for recognizing and normalizing time expressions. In: LREC. vol. 2012; 2012. p. 3735–3740.