Extracting Kinship from Obituary to Enhance Electronic Health Records for Genetic Research

Kai He[1,2], Jialun Wu[1,2], Xiaoyong Ma[1,2], Chong Zhang[1,2], Ming Huang[3], Chen Li[1,2*], Lixia Yao[3,*]

[1]School of Electronic and Information Engineering,
Xi'an Jiaotong University, Xi'an, China, 710049
[2]Shanxi Province Key Laboratory of Satellite and Terrestrial Network Technology Research and Development, Xi'an Jiaotong University, Xi'an, China
[3]Department of Health Sciences Research, Mayo Clinic,
Rochester MN, USA, 55905
{hk52025804, andylun96, pixie1997, zc6063}@stu.xjtu.edu.cn
{Huang.ming, Yao.Lixia}@mayo.edu
cli@xjtu.edu.cn

Abstract

Claims database and electronic health records database do not usually capture kinship or family relationship information, which is imperative for genetic research. We identify online obituaries as a new data source and propose a special named entity recognition and relation extraction solution to extract names and kinships from online obituaries. Built on 1,809 annotated obituaries and a novel tagging scheme, our joint neural model achieved macro-averaged precision, recall and F measure of 72.69%, 78.54% and 74.93%, and micro-averaged precision, recall and F measure of 95.74%, 98.25% and 96.98% using 57 kinships with 10 or more examples in a 10-fold cross-validation experiment. The model performance improved dramatically when trained with 34 kinships with 50 or more examples. Leveraging additional information such as age, death date, birth date and residence mentioned by obituaries, we foresee a promising future of supplementing EHR databases with comprehensive and accurate kinship information for genetic research.

1 Introduction

Kinship or family relationship is important for genetic research, particularly for understanding trait and disease heritability, predicting individual disease susceptibility, and developing personalized medicine (Chatterjee et al., 2016). Human genetics started by analyzing pedigrees and twins to understand the roles of heredity and environment in the manifestation of physiological traits and diseases. With the rising of genomics, Electronic Health Records (EHRs) and their integration through biobank, kinship information, if available, can largely augment latest high-throughput computational technologies such as deep phenotyping from medical records (Robinson, 2012) and phenome-wide association study (PheWAS, Denny et al., 2010), and accelerate population-based genetic research (Mayer et al., 2014; Polderman et al., 2015). Unfortunately, neither EHR systems nor claims databases capture kinship information systematically.

A few studies have investigated disease heritability based on inferred kinship information. For example, Wang et al. selected 128,989 families of 481,657 individuals from a large claims database covering 1/3 of the US population, by selecting policyholders and their dependents (e.g., spouse and children) who were on file for at least 6 years, to estimate 149 diseases' heritability and familial environmental patterns (Wang et al., 2017). Similarly, Polubriaginof and colleagues performed a multi-center study based on 3,550,598 patients' medical records from three EHR systems in New York City and used emergency contact information to build more than 595,000 pedigrees, in order to compute the heritability of 500 disease phenotypes (Polubriaginof et al., 2018).

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 1–10
Florence, Italy, August 2, 2019.

However, these studies relied on indirect sources to infer kinship information, which are incomplete and error-prone. First, both the dependents defined by medical insurance and the emergency contacts submitted to EHR systems by patients do not guarantee biological relationships. They do not distinguish adopted relationships or step relationships created through re-marriage from biological relationships. Second, dependents or emergency contact only represents a small portion of a person's whole family relationships. The 2010 Affordable Care Act allows young adults up to 26 to remain on their parents' health insurance plans. Before that, dependent children often "aged out" of their parents' health plan at age 19, or 22 if they were full-time students. Thus adult children older than those ages cannot be identified from claims data. In addition, if married couple work and receive medical insurance through their employers (even with the same employer), they are not usually linked on record. Likewise, most clinics and hospitals list emergency contact as optional (instead of mandatory) information. Most patients provide one or two emergency contacts, but not their entire family when filling the form – The Polubriaginof study (Polubriaginof et al., 2018) collected on average 1.86 emergency contacts per patient.

To address these issues, we propose a new data source (online obituaries) and a special Natural Language Processing (NLP) solution for systematically constructing biological relationships for large families of multi-generations. Obituaries contain rich and high-quality kinship information and are publicly available from the sites of newspapers and funeral services companies. Although obituaries are similar to social media, they are much less studied in biomedicine. One study analyzed obituaries to investigate cancer mortality trends (Tourassi et al., 2016). Another group combined LinkedIn profiles and obituaries to investigate the association between frequent relocation and lung cancer risk (Yoon et al., 2015). In this project, our ultimate goal is to link multiple obituaries by cross-validating name, age, residence and birth/death date information, to build large family trees. For this paper, we aim to investigate if state-of-the-art NLP methods can automatically extract names and kinships from online obituaries with high accuracy.

Establishing human names and their relations is a Named Entity Recognition (NER) and Relation Extract (RE) task. The NLP community has been working on both for many years. Usually, NER and RE are considered as two separate and sequential tasks (NER precedes RE). Most information extraction systems in biomedicine, including those mining biomedical literature to extract adverse drug events, and molecular interactions between drug, gene and proteins, are built on a battery of pipeline modules integrating NER and RE tasks (Miwa et al., 2012; Kang et al., 2014; Yildirim et al., 2014; Sun et al., 2017; Li et al., 2013; Li et al., 2017). However, pipeline models have inherent limitations: (i) The error from NER will propagate to RE. (ii) Pipeline models cannot fully utilize the internal connections between NER and RE to improve model performance when the separated models finished the two tasks independently. For instance, in a task of extracting adverse drug event, the named entity appeared before the relation keyword of "induce" (non-passive voice) would be a drug and the named entity after "induce" would be an adverse event. NER, which should be finished firstly, definitely would be harder to benefit from this relation information than RE. (iii) Pipeline models are computationally redundant and error-prone because they match up every two named entities to decide their relations, which is not necessary.

In this work, we propose a joint neural model to simultaneously extract names and kinships from obituaries, which combines a two-layer bi-directional Long Short-Term Memory (bi-LSTM) (Hochreiter and Schmidhuber, 1997) and a unique tagging scheme. It, in theory, surpasses pipeline models by overcoming the limitations (i) (Li et al., 2016; Zheng et al., 2017a) and (iii), and by making room for leveraging the contextual information and domain knowledge to address limitation (ii). The rest of the paper is organized into four sections. In the Data and Methods section, we describe how we annotated the obituary corpus, together with the special tagging scheme, the bi-directional LSTM model and evaluation metrics. Then in the Results section, we demonstrate corpus statistics and model performance metrics. After that, we share some discussions regarding the strengths and limitations of our method, before final conclusions and future work.

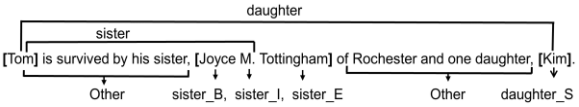

Figure 1: A novel tagging scheme for extracting names and kinships from obituaries

2 Data and Methods

2.1 Corpus preparation

We downloaded obituaries from the websites of three funeral services and one local newspaper in Rochester Minnesota, including: (1) http://www.bradshawfuneral.com, (2) http://www.czaplewskifuneralhomes.com, (3) https://mackenfuneralhome.com, and (4) https://www.postbulletin.com. The downloaded obituaries were published from 10/2008 to 09/2018. After removing those shorter than 290 characters, which is unlikely to contain any mentions of family relationships, messy ones with irregular HTML format or language, and duplicates, we selected 1,809 obituaries for annotation, due to limited resources and labor-intensive annotation described in next subsection.

2.2 Corpus annotation

The success of a machine learning application does not solely depend on the model itself. Most of the time it is more determined by the quality of data, particularly the gold standard dataset for training and testing the model. The challenge for annotating a natural language corpus is that the ground truth is not always obvious, due to the ambiguity and complexity of human language. A detailed annotation guideline and duplicated annotation by multiple people is often necessary to guarantee annotation consistency and corpus quality. Based on two examples of biomedical corpus annotations (Gurulingappa et al., 2012, Roberts et al., 2009), we designed an iterative annotation workflow and revised our guideline three times. All annotations were done at the document level so that the annotators can leverage the context in difficult cases. An open-source software called MAE version 2.2.6 (Kyeongmin, 2016) was used as the annotation tool throughout the entire process.

The corresponding author and three native speakers of English drafted the 1st version of annotation guideline. Then 3 computer science major students were trained for annotation in 2 rounds. In each round, we randomly selected 300 obituaries and asked each student to annotate 200 obituaries. This way each obituary was annotated twice by two different annotators. At the end of each round of training, we evaluated the annotation consistency using inter-annotator agreement (IAA) metrics and improved the annotation guideline. Considering that extracting kinship was actually a NER+RE task, we adopted precision, recall and F1 score rather than Kappa coefficient to report IAA, as suggested by Gurulingappa et al., 2012 and Chinchor, 1992.

	Precision (%)	Recall (%)	F1 score (%)
Training round 1	67.93	69.54	62.21
- Last name distribution	70.93	73.63	65.58
- Name with parenthesis	72.35	73.16	66.06
- Name-Residence Pair	69.32	71.11	63.67
- All features	76.84	78.98	71.01
Training round 2	74.61	76.31	68.92
- Last name distribution	77.71	80.51	72.40
- Name with parenthesis	79.03	79.94	72.77
- Name-Residence Pair	76.03	77.94	70.43
- All features	83.66	85.91	77.87
Final annotation	88.46	88.58	82.80
- Last name distribution	89.86	89.96	84.19
- Name with parenthesis	89.26	89.43	83.62
- Name-Residence Pair	88.58	88.68	82.91
- All features	90.94	91.05	85.27

Table 1: IAA scores in different rounds of annotation with different annotation features (- means "without")

After completing the training, 3 qualified annotators finished annotating the rest obituaries with the assistance of a rule-based quality control program written by us. Table 1 demonstrates that the precision, recall and F1 score were steadily improving through training round 1, training round 2 and final annotation. The discrepancy in the final annotation was resolved through group discussions. We warranted that 1,809 obituaries have high-quality annotations before building the models.

2.3 The tagging scheme

Conventional NER and RE are usually formulated as triplet tagging (entity_1, relation, entity_2). But our addressed task is not a general NER+RE task. It is simplified by three factors: (1) There is only one type of named entity to detect (human names); (2) all relations have the same first entity (the deceased); and (3) the first entity is mentioned in the metadata or the first sentence of the obituary, and hence does not need to be

detected most of the times. Therefore in this study, we proposed a novel tagging scheme inspired by Zheng et al (Zheng et. al, 2017b), which extracts names and kinships in relative to the deceased person in one step, as shown in Figure 1. We used the popular "BIESO" (begin, inside, end, single, and other) scheme to mark the position of words in entities, where "O" refers to cases that a word does not belong to an entity. This way we can identify a named entity by simply applying the rule of S or B + n*I + E, where $n \geq 0$. But we added the kinship type into the "BIESO" tags, in order to synchronize the NER and RE annotation. So each tag consists of two parts: the first part indicates the kinship type and the second part illustrates the position of a word in an entity. In an illustrative example shown in Figure 1, "Joyce M. Tottingham" is assigned three tags, including "sister_B" for the word "Joyce", "sister_I" for the word "M.", and "sister_E" for the word "Tottingham". For the single-word entity "Kim", the assigned tag was "daughter_S". All the remaining words were assigned a tag "O". Because we set the decreased as the default first entity for any kinship, triplets were simplified to duplets, like [sister, Joyce M. Tottingham] and [daughter, Kim] for the sentence in Figure 1. "Tom" was the name of the deceased person (inferred from the context or metadata) and we did not annotate it as a named entity. But we annotated other entity types including age, residence, birth date and death date. We plan to use these additional entity types in future work when we build the family trees and link them to EHR database.

2.4 The end-to-end joint neural model

The end-to-end neural model has lately demonstrated effectiveness in various NLP tasks, including NER, RE, part-of-speech tagging and semantic role labeling (Hashimoto et al., 2017, Strubell, 2018). In this study, we adopted an end-to-end neural model (See Figure 2), which contained an embedding layer, two bi-LSTM layers, and a softmax output layer. A rule-based result improver layer was also added to the end for consolidating the tags generated by the softmax output layer. We also used a dynamically weighted loss function to alleviate data imbalance issue.

The input sentences were tokenized and each token was converted to a word vector learned

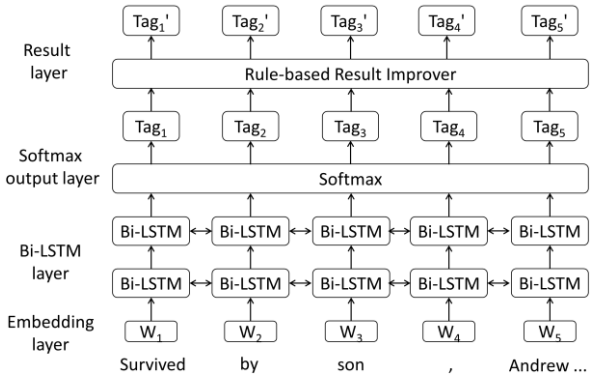

Figure 2: The neural network architecture for jointly extracting names and kinship types

from the GloVe method (Pennington et al., 2014), when fed into the embedding layer. Padding, which was a common programming trick, was performed in a way that all sentences were aligned to the longest sentence in a batch using padding tags for parallel computation. They would not impact the model performance as the output of those padding tags were masked out in the backward layer of the Bi-LSTM model. The Bi-LSTM architecture consisted of a forward layer and a backward layer, which was supposed to capture sequential context information bi-directionally. Both layers consisted of blocks made up of a forget gate, an input gate and an output gate. The forget gate decided how much information from the previous block would be dropped at the current block, considering the current input and the previous hidden representation. The input gate took the output of the forget gate and the previous cell state to update the current cell state. The output gate was designed to create a hidden representation for each token based on all the information from the forget gate and input gate. Finally, the outputs of both forward layer and backward layer were concatenated by Bi-LSTM as final representation. The softmax function served as the classifier for computing final normalized probabilities for each tag. After that, each token was classified into one of (m*5+1) tags, where m was the total number of kinship types. We tried m=57 and 34, according to the number of annotated examples in our experiment (See Table 4). In the end, a rule-based result improver was added to make sense of the sequence of the classified tags. For example, if the softmax output layer tagged two neighboring words as "sister_B" and "sister_I" without "sister_E" nearby, the improver would correct the second tag to "sister_E".

Corpus	Count	Deceased Person	Count	Special Language Patterns	Count
Sentences	30,035	Name	1,711	Last name distribution	5,186
Names	29,938	Mention of Age	1,517	Name with parentheses	8,118
Kinship	27,227	Mention of Death Date	1,712	• Nickname	84
Mention of Residence	8,476	Mention of Birth Date	1,522	• Previous last name	1,607
Name-Residence Pair	9,189	Mention of Residence	1,331	• Spouse's name	6,427

Table 2: Summary Statistics of the Corpus

Language Pattern		Example	Explanation
Last name distribution		Preceded in death by her grandparents, Ellen and Everett Uebel.	Uebel is also the last name for Ellen.
Name with parenthesis	Nickname	Kay is also survived by her daughter Maureen (Mo) Bahr of Rochester	Mo is the nickname of Maureen Bahr.
	Previous last name	Paul was born April 18, 1942 in Rochester to Boyd and Fern (Miller) Kinyon.	Miller is the maiden name for Fern Kinyon.
	Spouse's name	Survived by daughter, Sydney (Sam) Davis; granddaughter, Autumn Ellen.	Sydney Davis's husband is Sam Davis.

Table 3: Examples of unique language patterns in obituaries

Hierarchy	Kinship type
Generation 0	ex-husband (18), ex-wife (32), married to (1,457), spouse (18), husband (586), wife (690), sibling (718), cousin (91), **brother (2,106)**, **sister (2,156)**, half-brother (13), half-sister (7), sister-in-law (344), sibling-in-law (28), cousin-in-law (1), brother-in-law (251)
Generation 1	**child (2,658)**, daughter (1,445), son (1,713), niece (242), nephew (297), step-child (175), step-daughter (60), step-son (65), child-in-law (25), daughter-in-law (114), son-in-law (103), niece-in-law (20), nephew-in-law (25)
Generation 2	grandson (310), **grandchild (4,413)**, granddaughter (231), grandnephew (24), grandniece (24), grandson-in-law (13), grandchild-in-law (11), granddaughter-in-law (12), step-grandchild (98), step-grandson (7), step-granddaughter (6)
Generation 3	great grand-child (1,293), great granddaughter (46), great grandson (65), great grand-nephew (2), great grand-niece (6), great grandchild-in-law (4),
Generation 4	great-great grand-child (27), great-great granddaughter (1), great-great grandson (1)
Generation -1	**born to (2,332)**, son of (132), daughter of (172), parent (720), mother (155), father (139), step-mother (16), step-father (24), step-parent (2), aunt (49), uncle (54), parent-in-law (43), mother-in-law (30), father-in-law (26), aunt-in-law (6), uncle-in-law (3)
Generation -2	grandparent (210), grandmother (44), grandfather (29), grand uncle (1), grandmother-in-law (1)
–	Other* (987)

Table 4: 71 kinship types in annotated obituaries. Top 5 common relationships are highlighted in red.
* Other relationships refer to kinships not included in previous 6 categories, such as fiancé, guardian, and friend.

Dynamic weighted Loss function: We trained our joint model with weighted log-likelihood function, and used RMSprop (Tieleman and Hinton, 2012) for optimization. The objective function was defined as follows:

$$L = -\sum_{s=1}^{B}\sum_{t=1}^{L_s}\left(\log\left(p_t^{(s)} = \hat{y}_t^{(s)}\middle|x_s\right)\cdot\left(1 - P(O)\right)\right. \\ \left. + f_\omega \cdot \log\left(p_t^{(s)} = \hat{y}_t^{(s)}\middle|x_s\right)\right. \\ \left.\cdot P(O)\right) + \frac{\lambda}{2}\|\theta\|_2^2 \quad (1)$$

Kinship filter	Method	Average method	Precision (%)	Recall (%)	F-measure (%)
n≥10	Pipeline	macro	68.60 (4.81)	69.52 (4.98)	68.43 (4.90)
		micro	87.10 (0.57)	89.46 (0.82)	87.80 (0.78)
	Joint	macro	72.69 (3.96)	78.54 (3.85)	74.93 (3.95)
		micro	95.74 (0.98)	98.25 (0.43)	96.98 (0.60)
n≥50	Pipeline	macro	81.11 (3.70)	79.51 (2.62)	79.18 (3.22)
		micro	85.42 (0.98)	92.80 (0.43)	88.18 (0.60)
	Joint	macro	85.27 (3.90)	94.35 (2.09)	88.97 (3.18)
		micro	96.06 (0.64)	98.12 (0.37)	97.08 (0.46)

Table 5: Comparing the performance of pipeline model versus joint model. The values in brackets represent the standard deviation during 10-fold cross validation.

where B was the batch size, $L_s$was the length of input sentence x_s. ${\hat{y}_t}^{(s)}$and ${p_t}^{(s)}$ were the true tag and the normalized probability of the predicted tag for word t. λ was the hyper-parameter for L2 regularization. $P(O)$ was the indicator function to determine if the current tag was “O" (other), which was formulated as:

$$P(O) = \begin{cases} 0, & if\ tag = "O" \\ 1, & if\ tag \neq "O" \end{cases} \quad (2)$$

f_ω was dynamic weighted loss function, which assigned the tag ω different weights in different sentences, aiming to alleviate influence caused by too much “O" tag. It was defined as:

$$f_\omega = \frac{\frac{\sum_{j \in T} N_{D_i}^j}{N_{D_i}^\omega} - N_{Y}min}{N_{Y}max - N_{Y}min} \quad (3)$$

where T was the union of all possible tags, D_i referred to a sentence i in a batch of the training set, $N_{D_i}^\omega$ was the total count of all tags in D_i, $N_{D_i}^j$ was the number of a specific tag ω in D_i, and $N_{Y}max$ and $N_{Y}min$ were the maximal and minimal hyper-parameters for normalization respectively.

2.5 Evaluation metrics

A recognized named entity mention was considered true positive (TP) if both its boundary and type matched with the annotation. A relation extraction was considered as TP if both the NER and RE tasks were correctly captured. A recognized entity or relation was considered as false positive (FP) if it did not exactly match with the manual annotation in terms of the boundaries and relation types. The number of false negatives (FN) instances was computed by counting the number of named entities or relations in the manual annotation that had been missed by the model.

We performed 10-fold cross validation in our experiment, where 10% of the annotated data were randomly selected for validation, and the remaining for training the model. We evaluated the model performance using macro- and micro-averaged Precision, Recall and F-measure. A macro-averaged metric treats all classes equally by computing the metric independently for each class and then taking the average. In contrast, a micro-averaged metric aggregates the TP, TN, FP, and FN counts of all classes to compute an average metric.

Our corpus and codes could be downloaded at https://github.com/qw52025804/Obituary.git.

3 Results

3.1 Corpus annotation

Table 2 lists the detailed summary statistics of our corpus. There were 1,711 mentions of deceased names in 1,809 obituaries. Some obituaries mentioned the names of the deceased people in the title (metadata) rather than the main body of obituaries. In those cases, we directly linked the deceased names in the title of obituaries with their main body of free text. On average, each obituary

Examples of Correct Classification	
Sentence	**Extracted Relation**
On May 8, 1982 he married Madonna Oleson & became a proud dad of Ryan and Kelly.	Madonna Oleson： wife Ryan : child Kelly : child
He is survived by his brother Richard R. Arend (Carol) of Rochester, his beloved children and their mother, Kristy.	Richard R. Arend (Carol): brother Kristy : wife
One brother, Gordon "Scotty" Hyland of LaMirada, CA. and many nieces and nephews.	Gordon "Scotty" Hyland : brother
Examples of Wrong Classification	
Sentence	**Extracted Relation**
Craig is also survived by the boy's mother, Jolene Stock, sister Dianna Povilus; ...	Jolene Stock : mother
Survivors include Mary, his wife of 44 years and three children. Kristen (Matt) Asleson of Fountain, MN, and ...	Kristen (Matt) Asleson : grand child
Wooing Cecelia Stevens by serenading the words from the musical Carousel, "If I loved you, words wouldn't come in an easy way" - he proposed and on July 6, 1955, they began sixty-one years of marriage.	Cecelia Stevens : missing

Table 6: Correctly classified examples and wrongly classified examples

contains 16.6 sentences, or 1,809 obituaries contain 30,035 sentences in total. We extracted and annotated 29,938 names, 27,227 family relations and 8,476 residences for the deceased and their families. We were able to pair up a name and a residence for 9,189 times. For the deceased people, we also annotated their age, death date, birth date, and residence when available.

We noticed two interesting language patterns in obituaries, namely last name distribution and name with parentheses (See Table 3). These patterns might be due to the word limitation in the old time when the family paid for publishing an obituary on printed newspapers. In total, we annotated 71 kinships (See Table 4). Among them, 57 kinships have ≥10 examples, 34 kinships have ≥50 examples, and 28 relationships have ≥ 100 examples. The most populated five relationships were grandchild (4,413), child (2,658), born to (equivalent to parent, 2,332), sister (2,156) and brother (2,106).

It is worth noting that we kept "married to" and "spouse", "born to" and "parent" as separate kinship types in our experiment. This is because the syntax, co-occurred words and their order near "married to"/"born to" are subtly different from "spouse"/"parent". Keeping them as separate kinship types might help to improve the model performance. We will group them in the next step when we build the family trees, as they are semantically equivalent.

3.2 Model performance

Table 5 illustrates the final performance of the baseline method (pipeline model) versus our proposed joint neural model for extracting names and kinships from obituaries. The baseline model consists of two one-layer bi-LSTMs. The first bi-LSTM is for NER with simple BIESO tagging scheme, and its outputs were used as the inputs of the second bi-LSTM for RE. The general architecture is the same as that of the joint model, but the tagging scheme is different for NER, and NER and RE worked in a pipelined way. It is shown that the joint model outperformed the pipeline model by 4.09%, 9.02% and 6.5% for Precision, Recall and F measure at macro level using 57 kinships with 10 or more examples. The joint model outperformed the pipeline model by even bigger margins for Precision, Recall and F measure (4.16%, 14.84% and 9.79% respectively) at macro level when considering 34 kinships with 50 or more examples. The micro-level evaluation metrics demonstrated even better results of similar trends, due to the nature of an imbalanced multi-class classification problem. Table 6 showed some correctly classified examples and wrongly classified examples, which demonstrated the challenges in this project.

4 Discussions

The proposed joint neural model seemed capable of extracting the human names and relations with

high performance. For common kinship types with large number of examples in the training dataset, such as grandchild, child, parent (born to), sister and brother, the model's performance were close to perfect: Precision> 96.06%, Recall>98.12% and F measure> 97.08%. It could also recognize multiple variations of family relationships such as "marry" and "dad of", thanks to the high quality annotated corpus we created.

As shown in Table 6, the model was able to tell that "Kristy" was the wife of the deceased person (the second example of correct classification), but could not figure out "Jolene Stock" was the wife of the deceased "Craig" (the first example of wrong classification). It seems that the model was confused by the relationships between the deceased, "the boy's mother" and Jolene Stock. For the second example of wrong classification, the incorrect punctuation might have led to the error. The period before "Kristen (Matt) Asleson" should be a comma instead. The last example in Table 6 was an extremely difficult and rare case. Common kinship keyword indicating wife was missing. Without properly understanding the semantic meaning of 'propose' and 'marriage' in the sentence, our model failed to pick up "Cecelia Stevens" as a name.

One limitation of this study was that we built the Bi-LSTM model on sentences, and therefore lost the context information beyond a sentence. More sophisticated LSTM model would be helpful to parse the entire document of obituaries. Another challenge was that we could not afford to annotate more obituaries, which led to 14 kinship types had less than 10 examples (e.g., grandmother-in-law, grand uncle, great-great grandson and great-great granddaughter). Our model, or any supervised models, would not perform well on such small size of training data.

5 Conclusions and Future Work

In this work, we built an annotated corpus of >30,000 sentences (from 1,809 obituaries written in English) and proposed a two-layer Bi-LSTM model to simultaneously extract human names and kinships. Our joint neural model achieved macro-averaged Precision, Recall and F measure of 72.69%, 78.54% and 74.93%, and micro-averaged Precision, Recall and F measure of 95.74%, 98.25% and 96.98% using 57 kinships with 10 or more examples during 10-fold cross validation experiment. The model performance improved dramatically when trained with 34 kinships with 50 or more examples. We shared our corpus and codes on GitHub for the convenience of researchers.

Given such promising results, we will continue to improve our joint model to recognize other types of entity and relation, including the age, residence, birth date and death date. We will further parse names with parenthesis; resolve last name distributions; and leverage existing knowledge to infer the gender of names. Only when we complete theses tasks with high quality, could we build large family trees and link people to our EHR database. We are cautiously optimistic because almost all residents in Rochester MN have been patients at Mayo Clinic at some time of their life and population mobility rate in Rochester MN is far less than major metropolitan areas in the U.S. With the massive obituary data freely available on the Internet, our ultimate goal is to accelerate large-scale disease heritability research and clinical genetics research.

6 Ethics

In this study, we mined only publicly available information from 4 websites, without interacting with, intervening, or manipulating/changing the website's environment. The study does not include "human subject" data and is approved by the Office of Research and Compliance without IRB requirement at Mayo Clinic.

7 Acknowledgements

Funding for KH, JW, XM, CZ and CL are provided by the National Key Research and Development Program of China (2018YFC0910404); National Natural Science Foundation of China (61772409) and the consulting research project of the Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for "The Belt and Road" Training in MOOC China). Funding for MH and LY are provided by the National Center for Advancing Translational Sciences (UL1TR002377) and the National Library of Medicine (5K01LM012102).

8 References

Alvaro, N., Miyao, Y., Collier, N.: TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations. *JMIR*

Public Health and Surveillance (2017). doi:10.2196/publichealth.6396

Bekoulis, G., Deleu, J., Demeester, T., Develder, C.: Adversarial training for multi-context joint entity and relation extraction, 2830{2836 (2018). 1808.06876

Chatterjee, N., Shi, J., García-Closas, M.: Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* 17(7), 392{406 (2016). doi:10.1038/nrg.2016.27

Chinchor, N.: MUC-4 evaluation metrics. In: Proceedings of the 4th Conference on Message Understanding MUC4 '92 (1992). doi:10.3115/1072064.1072067. arXiv:1011.1669v3

Cohen, K.B., Fox, L., Library, D., Ogren, P.V., Hunter, L.: Corpus design for biomedical natural language processing. Technical report (2005). http://compbio.uchsc.edu/corpora

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, ACL 2002 (2002). doi:10.3115/1073083.1073112

Denny J C, Ritchie M D, Basford M A, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 2010, 26(9): 1205-1210.

Ge, T., Chen, C.-Y., Neale, B.M., Sabuncu, M.R., Smoller, J.W.: Phenome-wide heritability analysis of the UK Biobank. *PLOS Genetics* 13(4), 1006711(2017). doi:10.1371/journal.pgen.1006711

Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O 'connor, K., Sarker, A., Smith, K., Gonzalez, G.: Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark. In: *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing* (2014). doi:10.1590/S1516-35982012000500024

Gurulingappa, H., Rajput, A.M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L.: Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics* 45(5), 885{892 (2012). doi:10.1016/j.jbi.2012.04.008

Hashimoto, Kazuma, Yoshimasa Tsuruoka, and Richard Socher. "A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

Herrero-Zazo, M., Segura-Bedmar, I., Mart´ınez, P.: Annotation Issues in Pharmacological Texts. Procedia-Social and Behavioral Sciences 95, 211{219 (2013). doi:10.1016/j.sbspro.2013.10.641

Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* 9(8), 1735{1780 (1997)

Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E.M., Kors, J.A.: Knowledge-based extraction of adverse drug events from biomedical text. *BMC Bioinformatics* 15(1) (2014). doi:10.1186/1471-2105-15-64

Kyeongmin Rim: MAE2: Portable Annotation Tool for General Natural Language Use (May), 75{80 (2016)

Li, F., Yue, Z., Meishan, Z., Ji, D.: Joint models for extracting adverse drug events from biomedical text. *International Joint Conference on Artificial Intelligence* 2016-Janua, 2838{2844 (2016)

Li, C., Liakata, M., Rebholz-Schuhmann, D.: Biological network extraction from scientific literature: State of the art and challenges. *Briefings in Bioinformatics* 15(5), 856-877 (2013). doi:10.1093/bib/bbt006

Li, F., Zhang, M., Fu, G., Ji, D.: A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics* 18(1), 1-11 (2017). doi:10.1186/s12859-017-1609-9

MacKinlay, A., Aamer, H., Yepes, A.J.: Detection of Adverse Drug Reactions using Medical Named Entities on Twitter. *AMIA Annual Symposium Proceedings* 2017, 1215{1224 (2017)

Mayer, J., Kitchner, T., Ye, Z., Zhou, Z., He, M., Schrodi, S.J., Hebbring, S.J.: Use of an Electronic Medical Record to Create the Marshfield Clinic Twin/Multiple Birth Cohort. *Genetic Epidemiology* 38(8), 692-698 (2014). doi:10.1002/gepi.21855

Miwa, M., Bansal, M.: End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures (2016). doi:10.18653/v1/P16-1105. 1601.00770

Miwa, M., Thompson, P., McNaught, J., Kell, D.B., Ananiadou, S.: Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics* (2012). doi:10.1186/1471-2105-13-108

Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532-1543 (2014)

Polderman, T.J.C., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M., Posthuma,D.: Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics* 47(7), 702-709 (2015). doi:10.1038/ng.3285

Polubriaginof, F.C.G., Vanguri, R., Quinnies, K., Belbin, G.M., Yahi, A., Salmasian, H., Lorberbaum, T., Nwankwo, V., Li, L., Shervey, M.M., Glowe, P., Ionita-Laza, I., Simmerling, M., Hripcsak, G., Bakken, S., Goldstein, D., Kiryluk, K., Kenny, E.E., Dudley, J., Vawdrey, D.K., Tatonetti, N.P.: Disease Heritability Inferred from Familial Relationships Reported in Medical Records. *Cell* 173(7), 1692{170411 (2018). doi:10.1016/j.cell.2018.04.032

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., Setzer, A.: Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics* 42(5), 950{966 (2009). doi:10.1016/j.jbi.2008.12.013

Robinson P N. Deep phenotyping for precision medicine. *Human mutation*, 2012, 33(5): 777-780.

Strubell, E., Verga, P., Andor, D., Weiss, D., McCallum, A.: Linguistically-informed self-attention for semantic role labeling. arXiv preprint arXiv:1804.08199 (2018)

Strubell, Emma, et al. "Linguistically-Informed Self-Attention for Semantic Role Labeling." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.

Sun, T., Zhou, B., Lai, L., Pei, J.: Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* (2017). doi:10.1080/01418639108224439

Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 4(2), 26{31 (2012)

Tourassi G, Yoon HJ, Xu S. A novel web informatics approach for automated surveillance of cancer mortality trends. Journal of biomedical informatics. 2016 Jun 1;61:110-8. doi: 10.1016/j.jbi.2016.03.027

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, pp. 5998{6008 (2017)

Verga, P., Strubell, E., McCallum, A.: Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction (2018). doi:10.18653/v1/N18-1080. 1802.10569

Wang, K., Gaitsch, H., Poon, H., Cox, N.J., Rzhetsky, A.: Classification of common human diseases derived from shared genetic and environmental determinants. *Nature Genetics* 49(9), 1319{1325 (2017). doi:10.1038/ng.3931

Yildirim, P., Majnari´c, L., Ekmekci, O.I., Holzinger, A.: Knowledge discovery of drug data on the example of adverse reaction prediction. *BMC Bioinformatics* (2014). doi:10.1186/1471-2105-15-S6-S7

Yoon HJ, Tourassi G, Xu S. Residential mobility and lung cancer risk: Data-driven exploration using internet sources. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction 2015 Mar 31 (pp. 464-469). Springer, Cham. doi: 10.1007/978-3-319-16268-3_60

Zheng, S., Hao, Y., Lu, D., Bao, H., Xu, J., Hao, H., Xu, B.: Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing* 257, 59{66 (2017a). doi:10.1016/j.neucom.2016.12.075

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., Xu, B.: Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme (2017b). doi:10.24963/ijcai.2018/620. 1706.05075