Using Deep Learning to Discover Deontic Logic Statements

Cooper M. Stansbury

Winter 2019 CIS 5700

MIDAS

MICHIGAN INTEGRATED CENTER FOR HEALTH ANALYTICS & MEDICAL PREDICTION
UNIVERSITY OF MICHIGAN

UNIVERSITY OF MICHIGAN

Background:

- Donating biospecimen samples or clinical data requires informed consent
  - Research study
  - Clinical procedure
- Regulatory frameworks that prescribe **‘conditions of use’** for samples and data are complex and conflicting
- Different **‘conditions of use’** for samples and data are prescribed by:
  - Federal laws
  - State laws
  - Institutional policies
  - Local IRB(s)
  - Individual informed consent forms

Problem Statement:

1. Records of past consents are stored **in paper copies** that cannot easily be linked to the data or samples whose use they govern
2. ‘Permissions’ are specified by highly variable and conditional language
3. Current information systems do not account for an individual’s choices w.r.t donor consent
4. Managing individual records of **permissions** is not scalable

Research Question: Is it possible to use machine learning to automate the discovery of *statements of permission* in informed consent forms?

UNIVERSITY OF MICHIGAN

Data and Task:

Data:

- 752 unsigned, publicly available informed consent forms and informed consent form templates (and growing)
- Most available in PDF and were converted to text via OCR processes
- NOTE: Medical institutions do not want to share specific informed consent forms due to liability reasons, so these forms are hard to obtain

Goal:

- Train a model to predict if an input sentence contains language that ***permits some action*** if the form is signed in the affirmative
- Supervised classification task:

$$\hat{y} = F'(\mathbf{X}) + \epsilon$$

$\hat{y}$: *output, ideally a probability*

$F'(\mathbf{X})$: *latent structure approximation*

$\mathbf{X}$: *feature vector*

ϵ : *error*

UNIVERSITY OF MICHIGAN

Motivating Examples:

Simple:

- *“yes, i will donate a blood sample to be stored at the biorepository for future, undesignated research which may include genetic research.”*
- *“we may continue to use and disclose your information as described above indefinitely.”*

Harder:

- *“i do not wish to take part in the research”*
- *“i understand that there is a risk that my personal information may be accessed by people with no affiliation to the university”*

Initial Challenges:

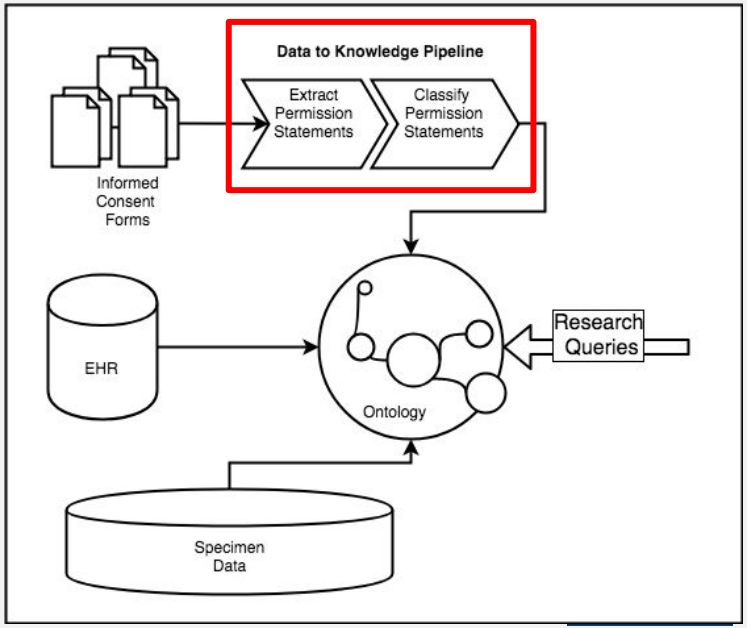
Data:

- High potential for bias in corpus
- Raw data is unlabeled
- Information loss during OCR conversion

Study:

- Unit of analysis (word, document, sentence?) (trade-offs)
- **Permissions can be semantically ambiguous**

Figure 1:
Proposed data infrastructure to allow for research queries subject to ‘permissions’ constraints.



Methods (I): Data Collection and Preparation

(1) Identify candidates:

- String cleaning (lowercase, force encoding, ect…)
- Parse (sentences) documents in parallel
- Strip short sentences
- Find sentences with *any* of the defined lexical clues
- *# Find sentences with any pronoun.*
- Find sentences above .95 similarity to those containing a lexical clue
- Remove duplicates

Identified ~24.7K candidates out of ~71K sentences

(2) Annotation:

- Annotated ~3.9K of these with binary labels
- (1 = ‘permission’; 0=’not permission’) using DataTurks
- Free tier only allows upload of 20,000

(3) Feature extraction:

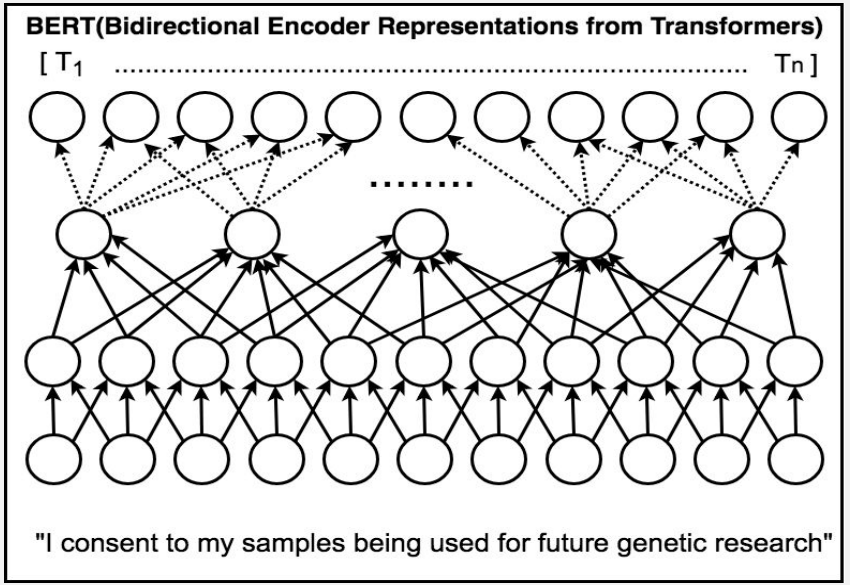
- Simple word counts, density, punctuation counts, ect…
- POS counts
- GloVe vectors
- *# Noun chunk counts*
- *# tf-idf features*

Methods (II): Classification

Permissions only account for ~0.22 of the training data. Used synthetic oversampling (SMOTE) to bring ratio to 1:1

Train classifiers:

- Started with baseline classifiers
- Moved to more advanced CNN using Keras
- Compared against transfer learning model: BERT, Google pre-trained bidirectional DNN
- Predict on new sentences



Results:

Model	Accuracy	Precision	AUCROC
KNN Classifier	0.480	0.261	0.569
Logistic Regression	0.828	0.719	0.664
Simple Decision Tree	0.698	0.323	0.566
Random Forest	0.843	0.635	0.789
Random Forest-2	0.837	0.807	0.664
Bagging Classifier	0.762	0.476	0.729
Bagging Classifier-2	0.779	0.750	0.505
AdaBoost	0.842	0.639	0.780
Gradient Boosted	0.814	0.572	0.755
Naive Baseline: (Majority Class)	0.777	0.000	0.500
Simple ANN (Keras)	0.731	0.731	0.786
CNN (Keras)	0.851	0.851	0.972
BERT	0.859	0.745	0.859

Figure 1: PC-1 vs PC-2 from tf-idf for predictions on 20K sentences

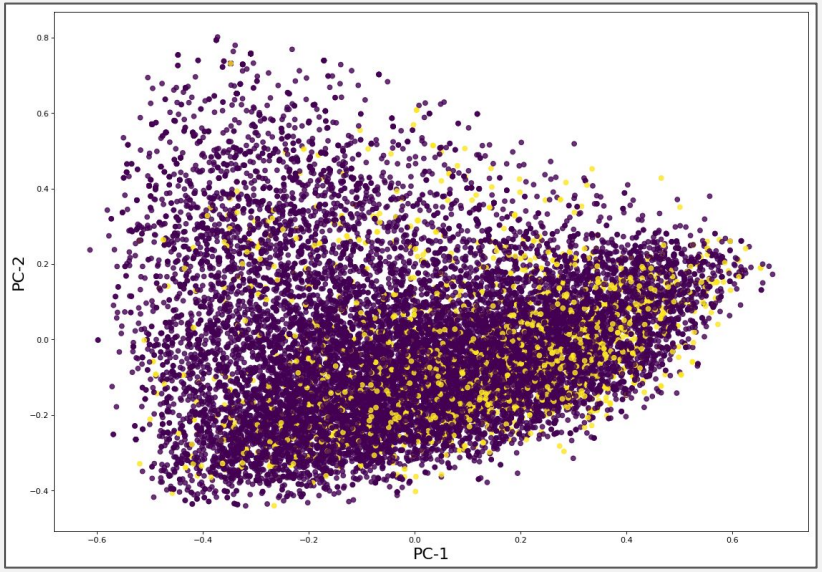
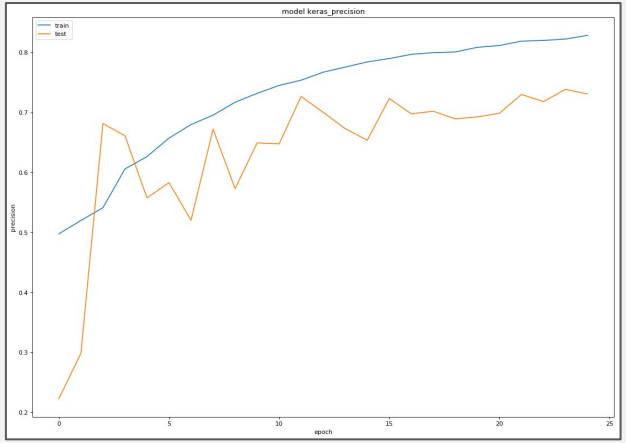


Figure 2: Simple ANN learning curve (precision)



Conclusions:

Limitations:

- Time (as always)
- Bias in the data
- Too few annotations
- Too few annotators
- Predictions low quality
- Unit of analysis not always appropriate

Implications:

- Possible to use deep learning to learn latent understanding of ‘permissible acts.’
- Facilitate greater, more ethical access to data and biospecimen samples

Future Work:

- Expand corpus
- Recruit more annotators and annotations
- Move algorithm execution to big data platforms (Spark, UM hadoop cluster)

Selected References:

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv:1810.04805 [cs], Oct. 2018.

[2] R. R. Faden, T. L. Beauchamp, and N. E. Kass, “Informed Consent, Comparative Effectiveness, and Learning Health Care,” New England Journal of Medicine, vol. 370, no. 8, pp. 766–768, Feb. 2014.

Acknowledgements:

Thanks to Michigan Institute for Data Science (MIDAS), Michigan Integrated Center for Health Analytics and Medical Prediction (MiCHAMP)

Dr. Marcy Harris, Elizabeth Umberfield, Krishan Amin

NIH/NHGRI 5U01HG009454-03

Links:

Source Code:

https://github.com/CooperStansbury/permission_statement_extraction

Data Repository:

https://github.com/CooperStansbury/InformedConsentForms

Appendix: Tooling and Technology

DEARBORN
COLLEGE OF ENGINEERING & COMPUTER SCIENCE

Primary language: Python 3.7.0 (+ some bash)

Python Libraries:

- **Data Structures:** pandas/numpy/native
- **Parallel Processing:** Dask
- **NLP Algorithms:** Explosion AI's spaCy (sentence parsing, POS tagging, similarity scoring)
- **Baseline Classifier Algorithms:** scikit-learn
- **Clustering Algorithms:** scikit-learn
- **Deep Neural Networks:** Tensorflow/Keras
- **Transfer Learning:** BERT
- **VC:** git