Making Sense of Clinical Trial Descriptions: A Text Analysis Approach

Munif Ishad Mujib[1], Jake Williams[1], Amy Gottsegen[1], Yuvraj Sharma[1], Anirvan Chatterjee[2], Oksana Gologorskaya[2]
[1]College of Computing and Informatics, Drexel University
[2]Clinical & Translational Science Institute, University of California San Francisco

The Problem

- Clinical trial descriptions are written at 17th-grade reading level.
- Fewer eligible and interested people are able to understand these descriptions.
- This reduces the pool of potential candidates.

```
Temporomandibular disorders (TMD) are characterized by pain and
tenderness in the muscles of mastication and/or the temporomandibular
joint (TMJ), limitations of jaw opening often accompanied by
deviations in mandibular path, and clicking, popping or grating TMJ
sounds. TMD is often found in association with other problems:
depression, anxiety, sleep disturbances, gastrointestinal symptoms,
frequent infections, etc. This project proposes to holistically
address patient symptoms through three different approaches,
Naturopathic Medicine (NM), Traditional Chinese Medicine (TCM), and
usual care at KPNW. We will conduct a pilot test and Phase II trial
to evaluate the two alternative healing approaches, TCM (n=50) and NM
(n=50) delivered by TCM and NM practitioners, are as effective as
usual TMD care (n=50) provided by dental clinicians in the KPNW TMD
Clinic. Subjects will be females 25-55 years of age with multiple
health problems (defined as patients who have had at least 4 organ
system-grouped diagnoses in the past year, not including TMD).
Subjects will be evaluated at baseline, 6 and 12 months after start
of treatment.
```

Figure 1: Example text from a trial description

The data

- 268,000 trial descriptions from ClinicalTrials.gov
- XML format
- Most information of interest contained in unstructured text fields

Keyword indicator matching

Type of information	XML field	Indicator	Frequency
size of study	brief_summary	\d+ patients	3.02
size of study	brief_summary	projected accrual	0.0
size of study	brief_summary	subjects	10.31
size of study	brief_summary	total of \d+ patients	0.23

Figure 2: Example per-document keyword frequencies

Dependency parsing

Sentence: Functional biomechanical outcomes will be measured at 6 months and 12 months using DSX at the Biodynamics Lab.
Extract: measured at 6 months and 12 months
Sentence: Patients who were discharged after an uneventful ERCP were contacted by telephone within 5 days to capture delayed occurrence of the primary end point.
Extract: contacted within 5 days
Sentence: All subjects will receive surgical treatment of their SCCis.
Extract: subjects will receive treatment SCCis
Sentence: Patients receive tacrolimus IV over 24 hours or orally daily on days -3 to 35 and oral mycophenolate mofetil twice daily on days -3 to 28 as graft-vs-host disease (GVHD) prophylaxis.
Extract: Patients receive tacrolimus IV

Figure 3: Extracting information using dependency parsing

Wikification

Key inclusion criteria:. - Males and females, 18 to 65 years of age, with HIV infection and a body mass index of 18.0 to 35.0 kg/m^2. - HIV-infected participants receiving a treatment regimen containing only atazanavir/ritonavir, 300/100 mg once daily (QD) tenofovir, 300 mg QD at least 1 other nucleotide reverse transcriptase inhibitor continuously for at least 3 months prior to study day 1. - Plasma HIV RNA levels of CD4 count >200 cells/mm^3. - No history of virologic failure on a protease inhibitor (PI), documented phenotypic PI resistance, or primary PI mutations, according to International AIDS Society recommendations. - No documented phenotypic resistance to atazanavir or primary genotypic mutations causing resistance to atazanavir. - History of Gilbert's syndrome, hemophilia, chronic pancreatitis, hypochlorhydria, achlorhydria, clinically relevant gastroesophageal reflux disease, hiatal hernia, or peptic/gastric ulcer disease. - Intractable diarrhea (≥ 6 loose stools/day for at least 7 consecutive days) within 30 days prior to study day 1. - Recent (within 6 months prior to study day 1) drug or alcohol abuse. - Evidence of organ dysfunction or any clinically significant deviation from normal in physical examination, vital signs, electrocardiogram (ECG).

Figure 4: Automatic linking of Wikipedia articles

Supervised learning: multiclass regression

- Development of annotation codebook
- Regression models to estimate: number of visits, number of interventions, active period, evaluation period, total study period, and whether or not the trial is of indefinite length
- Tf-idf feature generation

Model performance

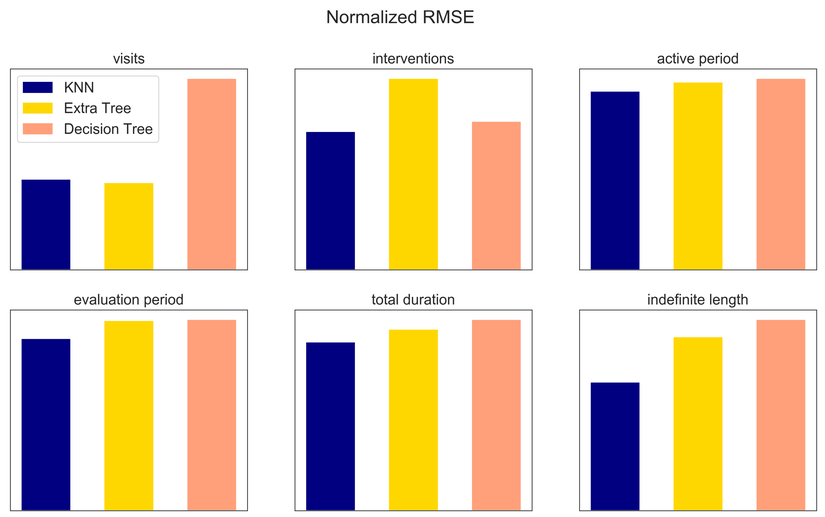

Figure 5: Normalized error rates for regression models: KNN, Extra Tree, and Decision Tree

Future directions

- Transformer language model-based feature generation for regression
- Simplification and/or abstract summarization using neural models