Toward language inference in medicine

October 29 2018, by Chaitanya Shivade

You will be shown a sentence from the Past Medical History section of a de-identified clinical note. Using only this sentence, your knowledge about the field of medicine, and common sense:

- Write one alternate sentence that is **definitely** a **true** description of the patient. Example, for the sentence "Patient has type II diabetes" you could write "Patient suffers from a chronic condition"

- Write one alternate sentence that **might be** a **true** description of the patient. Example, for the sentence "Patient has type II diabetes" you could write "Patient has hypertension"

- Write one sentence that is **definitely** a **false** description of the patient. Example, for the sentence "Patient has type II diabetes" you could write "The patient's insulin levels are normal without any medications."

Prompt shown to clinicians for annotations. Credit: IBM

Recent times have witnessed significant progress in natural language understanding by AI, such as machine translation and question answering. A vital reason behind these developments is the creation of datasets, which use machine learning models to learn and perform a specific task. Construction of such datasets in the open domain often consists of text originating from news articles. This is typically followed

by collection of human annotations from crowd-sourcing platforms such as Crowdflower, or Amazon Mechanical Turk.

However, language used in specialized domains such as medicine is entirely different. The vocabulary used by a physician while writing a clinical note is quite unlike the words in a news article. Thus, language tasks in these knowledge-intensive domains cannot be crowd-sourced since such annotations demand domain expertise. However, collecting annotations from domain experts is also very expensive. Moreover, clinical data is privacy-sensitive and hence cannot be shared easily. These hurdles have inhibited the contribution of language datasets in the medical domain. Owing to these challenges, validation of high-performing algorithms from the open domain on clinical data remains uninvestigated.

In order to address these gaps, we worked with the Massachusetts Institute of Technology to build MedNLI, a dataset annotated by doctors, performing a natural language inference (NLI) task and grounded in the medical history of patients. Most importantly, we make it publicly available for researchers to advance natural language processing in medicine.

We worked with the MIT Critical Data research labs to construct a dataset for natural language inference in medicine. We used clinical notes from their "Medical Information Mart for Intensive Care" (MIMIC) database, which is arguably the largest publicly available database of patient records. The clinicians in our team suggested that the past medical history of a patient contains vital information from which useful inferences can be drawn. Therefore, we extracted the past medical history from clinical notes in MIMIC and presented a sentence from this history as a premise to a clinician. They were then requested to use their medical expertise and generate three sentences: a sentence that was definitely true about the patient, given the premise; a sentence that was

definitely false, and finally a sentence that could possibly be true.

Over a few months, we randomly sampled 4,683 such premises and worked with four clinicians to construct MedNLI, a dataset of 14,049 premise-hypothesis pairs. In the open domain, other examples of similarly built datasets include the Stanford Natural Language Inference dataset, which was curated with the help of 2,500 workers on Amazon Mechanical Turk and consists of 0.5M premise-hypothesis pairs where premise sentences were drawn from captions of Flickr photos. MultiNLI is another and consists of premise text from specific genres such as fiction, blogs, phone conversations, etc.

Dr. Leo Anthony Celi (Principal Scientist for MIMIC) and Dr. Alistair Johnson (Research Scientist) from MIT Critical Data worked with us for making MedNLI publicly available. They created the MIMIC Derived Data repository, to which MedNLI acted as the first natural language processing dataset contribution. Any researcher with access to MIMIC can also download MedNLI from this repository.

Although of a modest size compared with the open domain datasets, MedNLI is large enough to inform researchers as they develop new machine learning models for language inference in medicine. Most importantly, it presents interesting challenges that call for innovative ideas. Consider a few examples from MedNLI:

Premise	Hypothesis	Annotation
ALT, AST, and lactate were elevated as noted above	Patient has abnormal LFTs	Entailment
The infant emerged with a spontaneous cry	The infant was stillborn	Contradiction
No known sick contacts	No recent travel	Neutral

In order to conclude entailment in the first example, one should be able to expand the abbreviations ALT, AST, and LFTs; understand that they are related; and further conclude that an elevated measurement is abnormal. The second example depicts a subtle inference of concluding that emergence of an infant is a description of its birth. Finally, the last example shows how common world knowledge is used to derive inferences.

State-of-the-art deep learning algorithms can perform highly on language tasks because they have the potential to become very good at learning an accurate mapping from inputs to outputs. Thus, training on a large dataset annotated using crowd-sourced annotations is the often a recipe for success. However, they still lack generalization capabilities in conditions that differ from the ones encountered during training. This is even more challenging in specialized and knowledge-intensive domains such as medicine, where training data is limited and language is much more nuanced.

Finally, although great strides have been made in learning a [language](#) task end-to-end, there is still a need for additional techniques that can incorporate expert curated knowledge bases into these models. For example, [SNOMED-CT](#) is an expert curated medical terminology with 300K+ concepts and relations between the terms in its dataset. Within MedNLI, we made simple modifications to existing deep neural network architectures to infuse knowledge from knowledge bases such as SNOMED-CT. However, a large amount of knowledge still remains untapped.

We hope MedNLI opens up new directions of research in the [natural language](#) processing community.

This story is republished courtesy of IBM Research. Read the original story [here](#).

Provided by IBM