Detection & Disambiguation in Free Text Using GDELT Data

Submitted by Michael Johns for CSCI E-89 Deep Learning Final Project, Spring 2019

Problem Statement: It is asserted that deep learning approaches can be applied to free text to detect and disambiguate entities and themes.

Approach: Use GKG (Global Knowledge Graph) metadata source from GDELT (Global Database of Events, Language, and Tone) to build and apply Deep Learning concepts for disambiguation.

Use / Benefits: GDELT metadata include URLs to crawl as well as extracted semantic content which can be used to augment and label crawl data to some extent.

Drawbacks / Challenges:

* A lot of data prep work given this was free text that was crawled from myriad sources.
* The combination of free text with disambiguation is challenging, e.g. lack of label data; most models are only able to predict within the vocabulary on which they have been trained.
* The need to stay with context-preserving approaches is important.
* The semantic content can drift if the corpus is too varied.

Results: The predicted cosine scores for the top 20 words in the Word2Vec vocab were near perfect; however, that declined to accuracy of ~.85 as we tested words towards the end of the word vectors. More comprehensive training, hyperparameter tuning, and possibly working with the established sentence vectors would potentially help.

Working Example Description: This Demonstration involves six main notebooks which show the following:

1. **Notebook-01 GDELT Intro:** Quick background on GDELT and the focus on UK Attack.
2. **Notebook-02 GDELT GKG Metadata:** Characterize and filter to records with Manchester, UK (the site of the attack).
3. **Notebook-03 GDELT Crawl Data:** Clean data crawled from URLs references in records.
4. **Notebook-04 GDELT Normalize Crawl Data:** Use ML Pipeline to prep for featurization.
5. **Notebook-05 GDELT Doc Finalize Data:** Establish Master Data for Disambiguation, join Crawl and Metadata DataFrame.
6. **GDELT Doc Model Data:** Word2Vec with Gensim, convert to matrix for input to Keras model.

Big Data Sources: Uses GDELT Event file dumps, from 20170522-20170524. Here is the link:

* <http://data.gdeltproject.org/events/index.html>.

Videos

* 2-minute summary at <https://youtu.be/0EWWrrIVta4>
* 15-minute overviews at <https://youtu.be/zYKnljbv9zM>