Introduction to NLP (Spring 2022)

Instructor: Dr. Manish Shrivastava

TA Team: Ananya, Hiranmai, Mayank, Sagar, Shivansh

1. **NER Extraction**

   Prepare Named Entity tagged data by utilizing existing annotated knowledge using snorkel.

   Assess the data quality

   Perform Transfer Learning through bootstrapping for F1 improvement

2. **Semantic Textual Similarity**

   Semantic Textual Similarity (STS) measures the degree of equivalence in the underlying semantics of paired snippets of text. Given two sentences, the model should return a continuous valued similarity score on a scale from 0 to 5, with 0 indicating that the semantics of the sentences are completely independent and 5 signifying semantic equivalence. Performance is assessed by computing the Pearson correlation between machine assigned semantic similarity scores and human judgements.

   Work on Cross Lingual Model (Spanish - English) or only English – English?

   Dataset:

   STS 2017 Cross-lingual English-Spanish Data

   STS 2017 Trial Data

   STS 2017 Evaluation Sets v1.1

   More Information and related Datasets can be found in at **Wiki:** STS Wiki

3. **Extracting Keyphrases and Relations from Scientific Publications**

The task deals with extraction of key phrases automatically given a scientific publication.

Moreover, the key phrase needs to be labelled and should be related to other key phrases. PROCESS, TASK and MATERIAL form the fundamental objects in scientific works. Scientific research and practice is founded upon gaining, maintaining and understanding the body of existing scientific work in specific areas related to such fundamental objects. This task aims to address the related fundamental problems in the field.

Corpus Description : https://scienceie.github.io/resources.html

4. **Measure Text Fluency**

Fluency is commonly considered as one of the dimensions of text quality of MT. Fluency measures the quality of the generated text (e.g., the target translated sentence), without taking the source into account. It accounts for criteria such as grammar, spelling, choice of words, and style. A typical scale used to measure fluency is based on the question "Is the language in the output fluent?".

Reference:

https://ieeexplore.ieee.org/document/1244655?arnumber=1244655

https://aclanthology.org/K18-1031.pdf

5. **Words Sense Disambiguation**

The automatic understanding of the meaning of text has been a major goal of research in computational linguistics and related areas for several decades. The task of Word Sense Disambiguation (WSD) consists of associating words in context with their most suitable entry in a pre-defined sense inventory. The de-facto sense inventory for English in WSD is WordNet. For example, given the word "mouse" and the following sentence:

"A mouse consists of an object held in one's hand, with one or more buttons."

we would assign "mouse" with its electronic device sense (the 4th sense in the WordNet sense inventory).

Typically, there are two kinds of approach for WSD: supervised (which make use of sense-annotated training data) and knowledge-based (which make use of the properties of lexical resources).

Supervised: The most widely used training corpus used is SemCor, with 226,036 sense annotations from 352 documents manually annotated. All supervised systems in the evaluation table are trained on SemCor. Some supervised methods, particularly neural architectures, usually employ the SemEval 2007 dataset as development set (marked by *). The most usual baseline is the Most Frequent Sense (MFS) heuristic, which selects for each target word the most frequent sense in the training data.

Knowledge-based: Knowledge-based systems usually exploit WordNet or BabelNet as semantic network. The first sense given by the underlying sense inventory (i.e. WordNet 3.0) is included as a baseline.

6. **Hypernym Discovery**

Hypernymy, i.e. the capability for generalization, lies at the core of human cognition. Unsurprisingly, identifying hypernymic relations has been pursued in NLP for approximately the last two decades, as successfully identifying this lexical relation contributes to improvements in

Question Answering applications (Prager et al. 2008; Yahya et al. 2013) and Textual Entailment or Semantic Search systems (Hoffart et al 2014; Roller and Erk 2016). In addition, hypernymic *(is-a)* relations are the backbone of almost any **ontology**, **semantic network** and **taxonomy** (Yu et al. 2015; Wang et al. 2017), the latter being a useful resource for downstream tasks such as web retrieval, website navigation or records management (Bordea et al 2015).

For each subtask and setting we provide a list of input terms (hyponyms) as well as a large *vocabulary* extracted from each corpus. Team is expected to deliver, for each input term, a *ranked list of candidate hypernyms* (up to **15**) from the provided vocabulary.

Corpus: https://drive.google.com/file/d/14_RgB3_it7a_1mLXeRCyzwY5BHdWgnlP/view

General-purpose corpora. For the first subtask we use the 3-billion-word **UMBC corpus** (Han et al. 2013), which is a corpus composed of paragraphs extracted from the web as part of the Stanford WebBase Project. This is a very large corpus containing information from different domains.

7. **Neural Unsupervised Paraphrasing**

   Paraphrasing is expressing a sentence using different words while maintaining the meaning. In this project teams will be implementing unsupervised approaches to generate paraphrases for Indian Languages.

8. **Anaphora & Coreference Resolution**

   In discourse, anaphora may be defined as a reference back to a word used earlier in a text or conversation, to avoid repetition. In this task, teams are expected to create an algorithm that resolves anaphora based on the dataset provided.

   In discourse, coreference may be defined as the phenomenon when two or more expressions in a text refer to the same person or thing; they have the same referent. In this task, teams are expected to create an algorithm that resolves coreference based on the dataset provided.

9. **Domain Terms Extraction**

   In this project, teams are expected to scrape multiple articles from multiple domains, and based on the available monolingual data, to identify and extract multiword expressions. A multiword expression may be defined as a single semantic concept represented in multiple words. In domain terminology, a multiword expression is a subject specific term that represents a concept particular to that subject. Teams are expected to use the corpus that they have scraped to construct a dictionary of domain specific terms based on how frequently used words occur with each other.

10. **Natural Language Inference**

    A fundamental phenomenon of natural language is the variability of semantic expression, where the same meaning can be expressed by or inferred from different texts. Textual entailment recognition is the task of deciding, given two text fragments, whether the meaning of one text is

entailed (can be inferred) from another text. Given two text fragments, one named text (t) and the other named hypothesis (h), respectively. The task consists in recognizing whether the hypothesis can be inferred from the text. TE has a three-class balanced classification problem over sentence pairs:

a. Contradiction
b. Entailment
c. Neutral

Dataset: multinli , SICK , SNLI

11. **Neural Dependency Parser**

Dependency parsing is a popular grammar formalism used for better understanding a sentence structure. A dependency parsing mechanism can give a parsed dependency tree for a given sentence, which involves a set of relations over its words such that one is a 'dependent' of the other. Training such dependency parsers involves making use of annotated data to learn the way these trees are constructed, which can be modeled to give good performance by effectively using neural networks. Based on the approach taken to model the problem, the parsing can be either **transition-based or graph-based**. In this project, the students will implement a neural model which will be trained to perform dependency parsing on a sentence, using a method of their choice.

Resources:

Universal Dependencies Project

Non-Projective Dependency Parsing in Expected Linear Time

Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task

12. **Textual Coherence**

An important aspect for a textual discourse, coherence measures the readability, clarity, and consistency of ideas expressed in a passage. Within a discourse, the coherency is exhibited at both - local and global levels. Whether to model the problem on a global level or to decompose it into a series of local decisions remains a modeling choice. In this project, students will experiment with neural models in measuring textual coherence.

Resources:

A Cross-Domain Transferable Neural Coherence Model

Neural Net Models of Open-domain Discourse Coherence

Following are the submission deadlines and criteria for the projects.

Following are the deadlines for project submissions.

project selection :10 February 2022

Project outline submission: 15 February 2022

Interim Report :15th March 2022

Final Report Submission: 21st April 2022

Evaluation: After Final Exams

For both Project Outline and Interim Report each team is expected to submit a recorded presentation (one video) of duration no more than 15 mins. Final submission will contain all the deliverables (incl report, code etc as detailed below). **All the submission should be done with due approval of your mentor.**

Project outline should contain :

- problem statement;
- Short Summary of papers(or methods) and datasets you would be implementing (Baseline and Baseline +);
- Timeline (workplan) ; interim and Final Deliverables.

Interim report should contain :

- description of work done so far.
- compare the work done against the timelines as mentioned in your project outline
- mention work plan till the final submission deadline.

Final submission should contain :

- Presentation file that will be used during your evaluation
- Project Report : detailing description of your project, the relevant literature you read up on for the project, your ideas/experiments, results and error analysis.
- Code : can submit the code as zip or git repo link. If you are submitting git repo link, please ensure no commits are made after deadline, till course grades are out.
- Data: if data is huge post git repo or drive link
- Saved Model Checkpoints: can attach a drive link
- ReadMe File: complete details and structure of submission (code location, code structure, data location, checkpoints), instructions to reproduce the results, input & output format.

One submission per team is enough.