**NLP – Project Proposition**

Group Members : Meghna Bhaugeerutty, Jean-Baptiste Del Rio & Gabrielle Morin.

Project : determining the inference relation between two (short) texts.

Data source : Stanford Natural Language Inference (SNLI) corpus

This database consists of sentence pairs annotated with labels indicating the relationship between the two sentences: contradiction, neutral, or entailment. It contains 570k human-written English sentence pairs which were manually labelled.

Examples :

|  |  |  |
| --- | --- | --- |
| Text | Judgments | Hypothesis |
| A man inspects the uniform of a figure in some East Asian Country. | Contradiction | The man is sleeping. |
| An older and younger man smiling. | Neutral | Two men are smiling and laughing at the cats playing on the floor. |
| A soccer game with multiple males playing. | Entailment | Some men are playing a sport. |

We anticipate three main challenges by using this database :

* Ambiguity : language is ambiguous and certain sentence pairs may be open to multiple interpretations.
* Subjectivity : real persons annotated the 570k pairs, which can influence the labelling process. For instance, different annotators may interpret the relationship between sentences differently.
* Data imbalance : one label may appear far less than the other two.

Project :

We plan to use this dataset with a Neural Network. The model should output one of the three labels for each sentence pair, corresponding to the possible outcomes in the database. The model accuracy will be computed on how often its prediction matches the true label.

As of today, we plan on separating the work as follows :

* Data preparation : Gabrielle
* Modeling : Meghna
* Baseline comparison : Jean-Baptiste