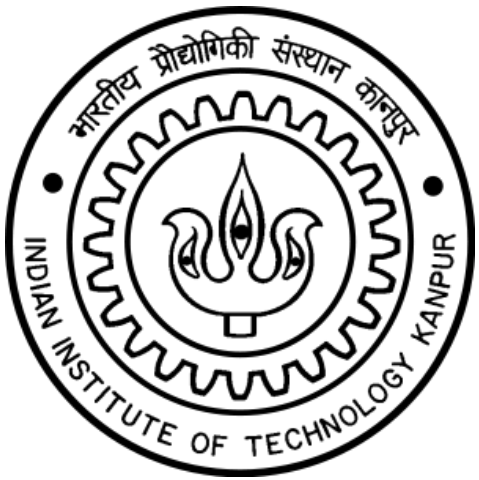
Research Intern'17 at University of Texas at Dallas (UTD)

ABHINAV AGRAWAL

Department of Electrical Engineering
Indian Institute of Technology, Kanpur

December 10, 2017



Outline

Outline

Frame Semantics and FrameNet

- Frame Semantics theorizes that the meanings of most words can be best understood on the basis of a semantic frame, a description of a type of event, relation, or entity and the participants in it.
- For example, the concept of cooking typically involves a person doing the cooking (Cook), the food that is to be cooked (Food), something to hold the food while cooking (Container) and a source of heat (Heating_instrument).
- In the FrameNet (a lexical database), this is represented as a frame called Apply_heat, and the Cook, Food, Heating_instrument and Container are called frame elements (FEs) . Words that evoke this frame, such as fry, bake, boil, and broil, are called lexical units (LUs) of the Apply_heat frame.

SEMAFOR

- SEMAFOR is an open source tool for automated analysis of the frame-semantic structure of English texts.
- It tries to identify words in a sentence that invoke a frame(called target) and then tries to find arguments(words or phrases) for the different Frame Elements of the particular Frame invoked.

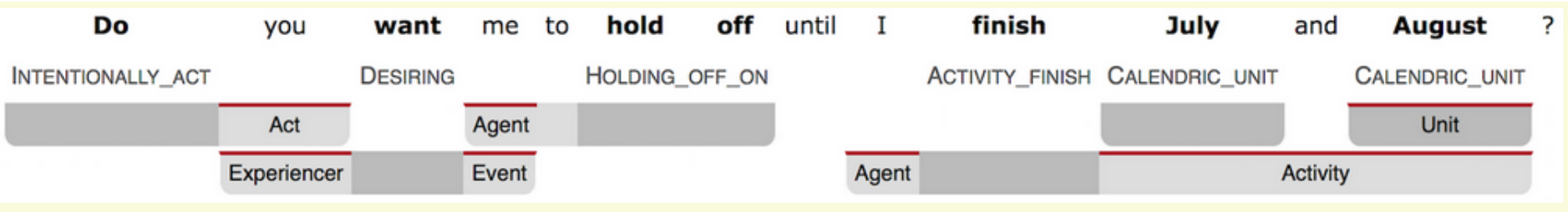


Figure: Output of a typical sentence from SEMAFOR online demo. Frames invoked are written below the LU's that invoked it, along with continuous section showing frame elements and their arguments

Outline

Re-training SEMAFOR

- The version of SEMAFOR available online was trained on an older version of FrameNet (version 1.5), which means that some frames that were added recently are not captured by this tool.
- Frames which were of interest right now, had one such frame, **Protest** which was added to the FrameNet 1.7 version and hence SEMAFOR was unable to capture it.

Re-training SEMAFOR

- The version of SEMAFOR available online was trained on an older version of FrameNet (version 1.5), which means that some frames that were added recently are not captured by this tool.
- Frames which were of interest right now, had one such frame, **Protest** which was added to the FrameNet 1.7 version and hence SEMAFOR was unable to capture it.
- A complete technique to re-train the SEMAFOR so that it starts capturing the any frame of interest was discovered with the help of the documentation provided on original repository of authors. The **detailed instructions** for training a domain specific version of SEMAFOR are now available here

Outline

Motivation, Previous work & Data-set

- The social protest presents itself in various forms such as industrial strikes, violent riots, peaceful sit-ins or long marches. Event coding on such stories has proved to be useful in quantitative social study.
- Due to the abundance of digitized text that is generated, it is very hard to filter document of a particular interest or concepts, before it is used for event coding.
- **Patrick et. al 2017** presented various methods for automated detection of protest stories from digitized text repositories by training supervised machine learning algorithms on a sample of New York Times (NYT) stories from the period 1987-1995.

Outline

Challenges and Models

- The NYT stories data-set used for training our classification algorithms had an imbalance ratio of **1:92** with

Class	Number of Stories
Protest	1645
Non-Protest	151402

- Such an imbalance ratio and only few thousand stories for positive class(Social Protest), with diversity in itself, poses a great challenge for any algorithm.
- It was decided to try two deep learning based models, one with **Convolutional Neural Networks** and the other one with **Bi-directional Recurrent Neural Networks** as the main architectures, both being considered current state of the arts for document classification.

Improving baseline models

- Previous Results

Model	Accuracy	PPV Protest Precision	NPV Non-Protest Precision	TPR Protest Recall	TNR Non-Protest Recall
tf-idf	88.18	12.97	99.58	82.36	88.31
Semafor	63.09	4.24	99.15	75.31	62.82
Word2Vec	56.05	3.74	99.17	78.86	55.56

- Improved Baseline Results

Model	Accuracy	PPV Protest Precision	NPV Non-Protest Precision	TPR Protest Recall	TNR Non-Protest Recall
CNN	**96.18**	**16.48**	99.57	61.82	**96.55**
Bi-dir. LSTM	**94.24**	**12.02**	**99.63**	67.88	**94.53**

- Note:
  - An important hyper-parameter was **class-weight ratio** for Optimization Loss, which accounted for the imbalance in the data-set
  - Training data was down-sampled so that the imbalance ratio became 1:10 as compared to 1:92 earlier.
  - The best results before overfitting are reported alongside previously published results for this task.

Outline

Retraining SEMAFOR with human annotated data

- The SEMAFOR model was trained using only the exemplar sentences provided on the FrameNet site. These do not cover all the important lexical units or frame elements of the Protest frame.
- It was hypothesized that if more sentences are annotated by human experts which cover the diversity of the *social protest*, then a better tool would be developed.
- To make the work human annotators easy, an instance of **brat** software has been set up for our particular task. After the human experts have finished the annotation part, the output of **brat** will be converted to the format required by SEMAFOR as training input. The code for this is also complete.

Outline

Where two rivers finally meet ?

- *"Data beats algorithms"*
  With this intuition behind us, we wish to use the SEMAFOR output for the NYT stories as input to classifiers in a novel way. SEMAFOR annotated text contains information which if presented in the right form shall increase the accuracy from current levels.
- This interesting experiment which uses the SEMAFOR output with current state of the art algorithms, is awaiting the annotated output files from a domain specific version of SEMAFOR.

Summary

- With collective efforts, a retrained model of SEMAFOR, an automated frame semantic labelling tool was developed.
- The domain agnostic process used to retrain SEMAFOR can be extended to any field of interest, thus exploiting its complete potential.
- With use of current state of the art methods performance was increased on the classification task, which shall provide a better baseline for future work.

- Outlook
  - Still need to test novel models on the SEMAFOR output.
  - Try other state of the art models that solve the classification problem and use them as baselines for comparisons.

Acknowledgements

- I am grateful for the opportunity provided to me by **Prof. Gopal Gupta** and **Prof. Shyam Karrah**, and my advisor **Prof. Latifur Khan** to accept in his lab as a research intern. Their constant guidance and support made my stay pleasant and allowed me to focus on work.
- I would like to extend an special thank to **Prof. Jey Veerasamy**, who provided for our day meals, helping us cope with the new environment.
- I would like to extend my gratitutde to **Mr. Sanjiv Khosla**, who arranged for this program from New York Office of IIT Kanpur.
- A big thanks to all the **friends** in Prof. Latifur's lab and at my accomodation for making this stay a wonderful experience.

Thank you.