Impressions from EMNLP 2018 by Aditya Joshi
(The notes were written during the conference, with minimal post-editing. I may have skipped removing some rants, apologies.)

Workshop: Social media monitoring 4 Health (SMM4H) (on day -2)

1) Opening session by Graciela G.

- 35,000 (?) tweets per minute reporting health
- Past work seems to focus on getting positive datasets. Are we paying enough attention to getting negative instances? What should they really comprise?

2) Football and beer

Football and beer… people drink beer when football matches are on. The paper was a social media analysis of tweets around football matches.
FIFA: When a match is on, download tweets containing beer/alcohol emojis. Then, analyse them in terms of counts. Observations were qualitative in terms of …more tweets when their team is playing. Etc.

3) Stance taking in topics extracted from vaccine-related tweets

They download vaccine-related discussion posts. Run a topic model. Get the topics. Then, classify topics in their prevalence in the posts, to understand what proportion of tweets take stance

4) Identifying depression on reddit: The effect of training data

Interesting observations. They said that positive instances are about depression self-reporting on reddit. What should the negative instances be? They essentially experiment with different negative instance-containing datasets. They show that depending on what your negative class semantically represents, the performance of your classifier will change.

5) Overview of shared task

Pictures showing scores, etc. of shared task participation.

6) Changes in psycholinguistic attributes

Get tweets reporting flu symptoms. Look at tweets by the same author before and after 3 week interval. Then classify these tweets on psycholinguistic attributes – and then, see what behavioural changes happen during or after a flu.

They observe that people become more emotional and engaging on social media after flu

7) Sentiment analysis of medical online forums

Posts on a medical forum. Sentences manually annotated. Set of classifiers with features

8) Blood donation request generation

Indian context. People ask for blood donation when there is an emergency/accidents. They create set of tweets containing ‘blood’ etc. and predict as to whether a tweet contains

request for blood donation or not. They also use community-based handles which are dedicated to making blood donation requests.

9) Medical non-adherence detection on twitter
Adverse drug reactions can happen if people do not take their meds properly. Can such non-adherence be detected? Classification/entity extraction of sorts.

10) … Shared task related papers. The paper that won the first and third task used a series of self-attention layers. This was followed by other shared task papers. Another paper used stacked bidirectional LSTM. There was another paper that used character n-grams. Other approaches include bagging (sampling instances for a committee of classifiers), use of qualitative features, etc.

Main Conference, Day 1

(It seems the "Coding for NLP Research" tutorial held on day -2 was pretty popular. The slides are available online.)

Inauguration/Welcome address:

1. 100% **increase** in number of registrations :-O
2. Acceptance rate remains similar

Plenary talk 1: Julia H. "deception detection"

The projects have been for government entities in the US. She talked about deception detection. She motivated the problem by describing how professionals who need deception detecting abilities do not do well in detecting deception.

Described her experiments in detail. Their expts involved native and non-native speakers talking to each other. They first fill biographical questionnaires and randomly lying. Monetary rewards if you could fool your interviewer. The participants fill biographical questionnaires to start with.

Linguistic and Acoustic features.... Opensmile and other things. She reported Random Forests, Naïve Bayes… in addition to deep learning architectures.

ML models beat human performance

Session on Social Applications 1:

1) Privacy preserving neural representation

Can private information in text be uncovered from neural representation? Adversarial learning where the adversary tries to act as an attacker.

2) Can demographic attributes be predicted using neural reps?

Similar to (1). They have a classifier that detects emojis from text. They use the last state representation of the LSTM to train an attacker that tries to predict demographic attributes.. gender?

3) Fake news detection

Input is a sentence (claim), evidence is a set of web pages extracted using a web search API. Then, they try to validate if the claim is true based on evidence. The architecture uses claim and evidence embeddings, followed by self-attention.. Article embeddings get augmented.. Finally a series of dense layers.

4) It's going to be okay

Is it better to reveal gender online to get support?

Seems like a set of assorted classification tasks. Polite or not, supportive or not, etc. Predict gender using usernames. They download a set of posts, predict them as polite response or not. Supportive response or not. Then they break it down in terms of gender. Did the post get support?

Conclusion: women get more support but also face more attacks

5) detect aggression and loss from gang-involved youth

Nice social application paper.

Gangs on social media reporting aggression and loss. They classify tweets as detection (yes/no), loss (yes/no). Features comprising of Domain specific word embeddings

Poster session

1) Out of domain detection for sentences, using a GAN

2) BLEU is no good for text simplification.

A manual corpus is created for text simplification.. and then correlated with BLEU. Shows poor correlation

3) dataset of well-formed natural questions... And then a detection system, to detect if a question is well-formed or not

4) Coherence-aware neural topic modelling

Optimise topic coherence during training

5) Topic intrusion to evaluate topic models
automatic method to evaluate topic intrusion (rogue word inserted in topic words). Input is document and a set of topics

Main Conference, Day 2

Session: Semantics & Summarization
1**) Idiom usage detection**
Given a sentence, can you detect if it is idiomatic or literal.

Without labelled data, they aim to train a classifier. A heuristic uses semantic similarity between idiomatic words and context and assigns a soft label. Then uses a classifier. Comparable performance with sota

2) **Coming to your senses**

Sense embeddings are trained. What are the right ways to evaluate them? Is it better to be polysemy-aware?

The paper examines several questions: (a) if averaging similarity over senses is a good idea... (b) If sense embeddings need to be evaluated on the right task... (c) If quality of annotations important for quality evaluation of embeddings
The speaker is speaking too softly. I cannot hear him clearly. I am in the fourth row.

2) Abstractive summarisation

“Don't give me the details, tell me a one sentence summary.”

Dataset: BBC news articles have a one line heading... And a body. So, the body is the input and one line heading is the output

New dataset and a topic aware convolutional model. Convolutional representation for the document. Uses LDA Distributions to encode a document

Evaluation on... Informativeness, well-formedness

3) Adjective intensity
Using backoff for the similarity scores.. to compute graded sentiment strength between adjectives. Paraphrasing can provide a mechanism to compute adjectival intensity

Session: Question answering

4) Exam dataset

CLOTH dataset. Thoda arbit hai.. I don’t understand this.

5) emrQA: QA for medical records

Questions on medical records... With rules to generate answers. Rules have been generated by physicians. Use schema to generate patterns of q and a.

6) QA dataset using Wikipedia. Along with supporting facts. The questions aree of diverse types, contain different kinds of answers. The baseline model is an attention based model. They call their dataset HotpotQA.

7) Multi hop question answering using an "open book" of scientific facts.
The example question shown in the talk seems difficult. OpenBookQA. A dictionary fo open-book facts and supervision from KB such as wordnet – used collectively to answer questions. Baselines reported.

8) validating theory of mind using question answering
Interesting validation of beliefs based on a set of sentences. Milk.. is it in the pantry or the fridge? Nice paper from Facebook AI.

Session: Social Applications

9) gender bias in abusive language detection
Abusive is profane, offensive language. Not necessarily against minorities.

Although a good f score was obtained, classifiers tend to return hateful speech positive when words like female or Muslim were used in a sentence.

They generate dummy sentences.... Sexist and abusive tweets...

Measure model bias due to gender
Different pre trained embeddings or architecture can have gender bias. They propose three mitigation approaches.

10) sexual harrassment reports
Forms to report SH involve a text area and a set of labels to indicate type of classification.

Safecity... Is a website where people describe and also self-label.

Single label models (based on svm, CNN and rnn) and multi label models (based on svm, CNN and rnn)

Lots of interesting analyses of the output. What words are important for labels, what are word similarity, time/abuser of harassment etc.

11) multi-view embeddings for dementia
Dementia affects memory and language. We can use language for early detection of dementia

Two kinds of participants.. patients and healthy. Each patient participant has a score.

Participants participate in three tasks: (a) describes in words .. think of words that begin with a letter... Think of words that belong to a category (e.g. animal)...

Two tasks: detect healthy or not.. and then predict score value

Generalized canonical correlation analysis... Turn different features into embeddings. Random forest with 100 decision trees.

12) Wikipedia conversations dataset

Dataset of Wikipedia conversations.. then they analyse it in terms of forum used, nature of conversations and hate speech, etc.

Plenary talk 2: Bloomberg's head of Data Science "News move markets"

This seems to be related to vizie

Finance technology... Pre-historic methods of book-keeping.
News move markets. News articles cause stock market changes
First to break news also matters.

Within milliseconds of publishing a news, the stock market changes. The author described four steps:

Detection: Detecting a breaking news... Impact of fake news on stock market....
(Explosions in white house and Obama injured)
(Geolocation credibility.. building trajectory of where someone has been to predict if the person is likely to have been at a location.)

Extraction: entity extraction. Finance markets is difficult because it can be fine grained and can have point in time granularity. He showed an ontology example.
Interpretation: Will detection/extraction step result in increase or decrease?

Presentation... Visualisation

Sentiment analysis posters

Hyperbole detection: Dataset available. Simple sentence vectors and classifiers, along with qualitative features. Nice paper and poster.

Multi-modal Sentiment analysis: Tri-modal sentiment analysis

Learning a language model from code-mixed data: Uses a switch variable and learns separate LSTMs for each language. Mandarin-English

Native language identification: Based on a person's 'chunk of reddit posts', can you predict their native language? Datasets consist of European users.

Automatic detection of ‘vague’ words in policies: Modeled like a WSD task
Demo: Given the abstract and title of a paper, generate the title of a science journalism article.

Demo: Visualisation of attention weights for a trained model.

Main Conference, Day 3

Session: Sentiment Analysis

1) Sentiment classification towards QA
Question-answer snippets where sentiment is sought and expressed. They use a bidirectional matching layer to allow specific aspect words to be included.

2) Cross-topic argument mining
   Integration of topics into bilstm to allow cross-domain/topic argument mining. They use CLSTM, a contextual LSTM cell that allows to store topic information.

3) Summarizing opinions

Dataset of reviews, along with human-annotated summaries and aspect-specific scores. Multi-task learning for domain-specific extraction of aspects.

4) CARER: Contextualized affect representations for emotion recognition
Generate graphs based on sentences in an emotion-annotated dataset. A graph-based algorithm that seems to generate emotion representation in terms of aspect/sentiment words. Non-deep learning paper, compares with DL algos