Three applications of NLP: Drug intoxication, Schizophrenia, and Alzheimer's Disease

Presenter: Elif Eyigoz, IBM Research

Outline

- Semantic Analysis

  Using Automated Metaphor Identification to Aid in Detection and Prediction of Schizophrenia

  E. Darío Gutiérrez, Philip R. Corlett, Cheryl M. Corcoran, Guillermo A. Cecchi, EMNLP 2018

- Acoustic Analysis

  Speech Markers of Oxytocin and MDMA Ingestion

  C. Agurto, R. Norel, R. Ostrand, G. Bedi, H. D. Wit, M. J. Baggott, M. G. Kirkpatrick, M. Wardle, and G. A. Cecchi, "Phonological Markers of Oxytocin and MDMA Ingestion," Interspeech 2017, 2017.

- Syntactic Analysis

  Predicting cognitive impairments with a Mobile Application

  E. Eyigoz, R. Tejwani, G. Cecchi, ICAART 2017

Using Automated Metaphor Identification to Aid in Detection and Prediction of Schizophrenia

Motivation

- Schizophrenia affects 20-70 million worldwide. Global cost over $102 billion per year.
  - Mental health practitioners in short supply
  - Opportunity for AI to assist practitioners
- Can we motivate an algorithm from clinical psychiatry literature?
  - 50 years of observations on “idiosyncratic” speech use among patients (Kuperberg 2010)
  - Examples from Andreasen (1986):
    - **Watches** were called **time vessels**
    - **Gloves** were called **hand shoes**
  - Billow et al (1997) : Patients use more metaphors that healthy controls but they tend to be bizarre

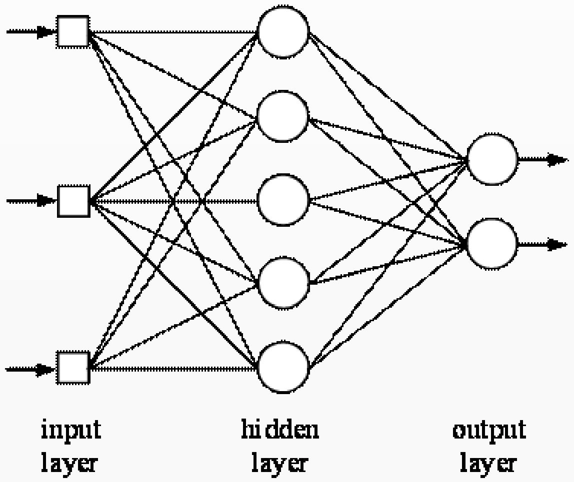
Metaphor Detection

- Based on token level in running text.
- The attorney demolished the prosecution's arguments
- [0 0 1 0 0 0]

- **HYPOTHESIS:** People with schizophrenia produce signficantly more tokens tagged as metaphorical than do healthy controls

Metaphor Detection Algorithm

- Metaphor Detection Algorithm
- Based on the work of Do Dinh and Gurevych (2016).
- Trained on VU Amsterdam Metaphor Corpus (Steen et al. 2010)
- Supervised sequential learning: multilayer perceptron w/ sliding window



text tokens **probabilities**

Classification

Features:

- Token-level metaphoricity
- Additional features:
  - Bizarreness
    - Measured using 2-gram likelihood
  - Token-level sentiment (on 0-5 scale)
    - Stanford sentiment analysis tool

Classifiers:

- RBF support-vector classifier, convex-hull classifier

Experiments and Results

Experiment 1: First-episode Schizophrenia

- 18 patients with schizophrenia, 15 healthy controls.
- Data: Open-ended transcribed interviews
- Test and train using LOO-CV

Variables	F-score	Accuracy
Met+Biz+Sent	0.848	0.833
Met	0.778	0.733
Bedi et al	0.773	0.667
Mota et al	0.733	0.733
Baseline	0.723	0.567

Results:

- Metaphor identification algorithm tags a significantly higher proportion of tokens of schizophrenia patients than in healthy controls.
- Outperform the Mota et al and Bedi et al methods the majority baseline ($p < .005$)
- Combining with bizarreness and sentiment features improve performance

Experiments and Results

Experiment 2: Clinically high risk youth

- Prodromal Psychosis:
  - 34 youths at clinically high risk for schizophrenia
- Five suffered a first episode of psychosis within 2.5 years of transcribed interview
- Train and test on clinically high risk youth using LOO-CV

Results

- Correctly prognosticated 33 of 34, Bedi et al. predicted 34 of 34

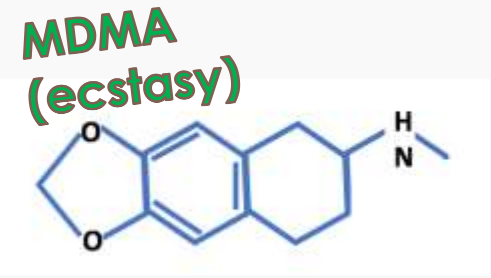
Conclusion

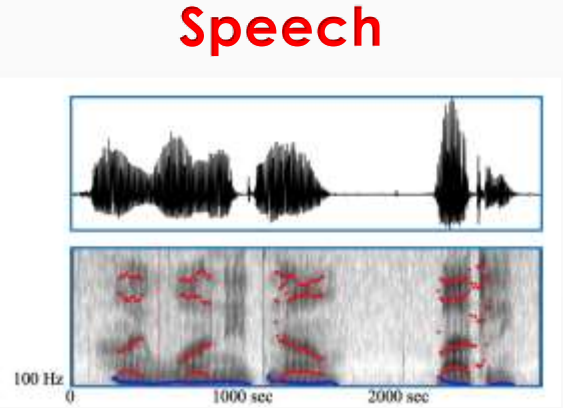
- First demonstration of utility of metaphor identification for detection of schizophrenia
- Supports previous clinical psych research on language-use abnormalities in schizophrenia

Speech Markers of Oxytocin and MDMA Ingestion

Motivation

- Correct assessment of substance use disorders is essential for treatment planning and referral to adjunctive services.
- Clinical reviews are used for this purpose. Issues with objectivity and comprehensiveness.
- Speech data has the potential to provide quantitative information about mental states beyond subjective assessments.





Data

- Subjects: Ecstasy users (at least twice in their lifetime) were recruited and performed different speech tasks.
  - 32 subjects (12 F: 24.6 + 4.7 years, 20 M: 24.1 + 4.5 years)

- Protocol: All participants received, in randomized order, doses of placebo, MDMA at two different concentrations (0.75 mg/kg and 1.5 mg/kg), and Oxytocin.

- Procedure :
  - Participants were asked to perform a monologue speech task of 5-minute durations in each session.
  - Recordings were manually transcribed.

Acoustic analysis

- Mel-Frequency cepstral coefficients (MFCCs).
  - characterize the voice spectrum similar to pitch perception in the human auditory system
- Vowel space (e.g. distribution of formants which measure vocal tract resonances)
- Voice stability (e.g. jitter, shimmer)
- Noise measurements (e.g. harmonics to noise ratio)
- Temporal features (e.g. articulation rate, pause duration distribution)
- Spectral characterization (e.g. slope of frequency spectrum)

Results

- Statistical Analysis: Wilcoxon signed rank test with FDR correction
  - F2 helps distinguish OT from PBO.
    - Positive valence (elation, pleasure, etc.) resulted in higher F2 values.
  - Median pitch distinguish MDMA low dose vs PBO.
- Classification experiments
  - Nested leave-subject-out cross validation approach using Linear SVM and Random Forest.

Classification Accuracy	
MDMA 0.75 vs. PBO	0.85*
MDMA 1.5 vs. PBO	0.71
MDMA 0.75 vs. MDMA 1.5	0.81*
Oxytocin vs. PBO	0.87

* Random Forest

Conclusion

- First study that uses characteristics of speech to identify subjects that are under the influence of MDMA and Oxytocin.
- Most relevant acoustic features correlate with positive valence, which supports previous research of drug effects using subjective analyses.

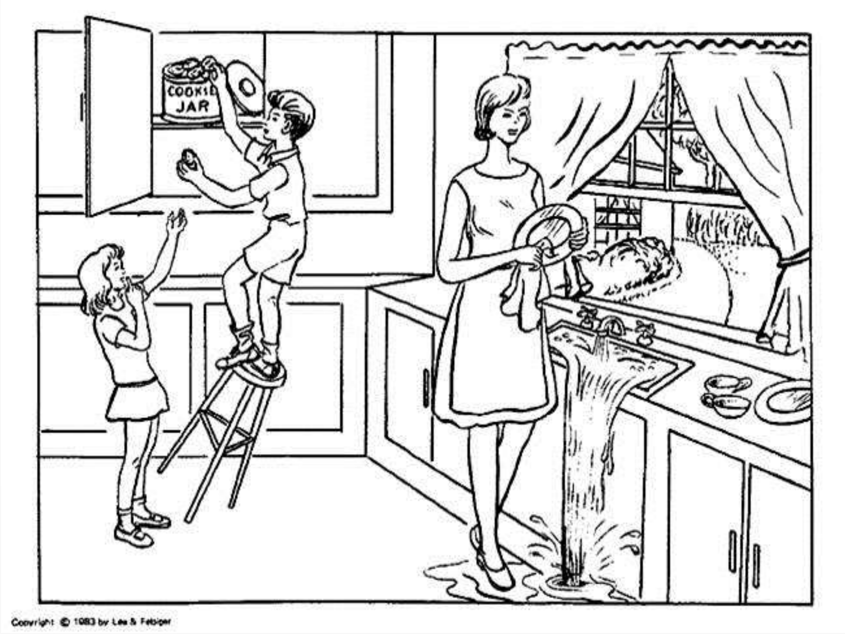
Predicting Cognitive Impairments with Syntactic Analysis

Motivation

- In 2016, about 47 million people worldwide were affected by dementia
  - 131 million by 2050.
- Demented subjects have difficulties with both comprehension and production of syntactically (grammatically) complex utterances.
- Utterances of the demented adults were shown to be shorter and syntactically (grammatically) simpler than those produced by the nondemented adults.

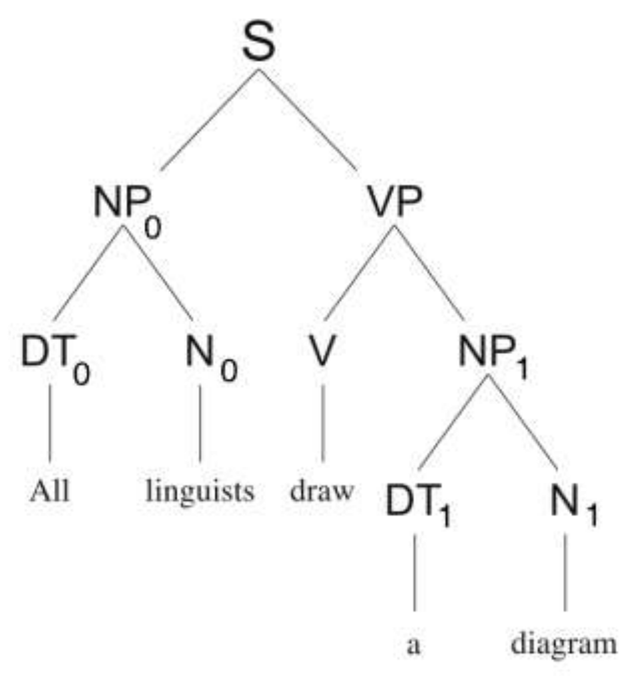
Data

- DementiaBank Pitt Corpus
- Cookie-theft picture description task
- Mini mental state examination (MMSE) score for each sample



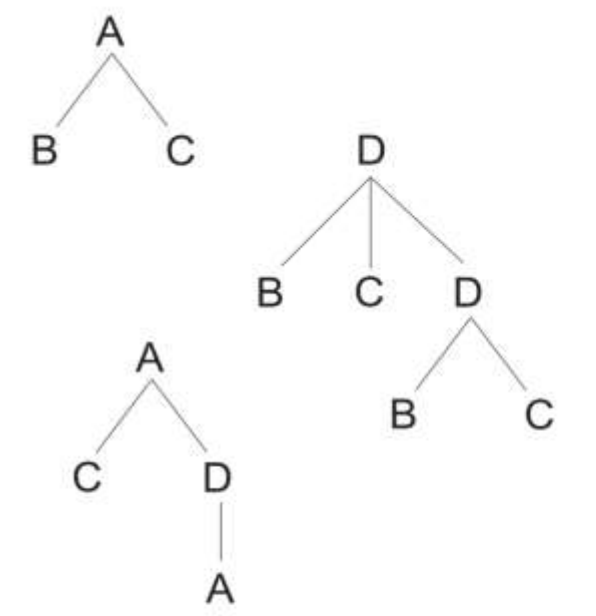
Syntactic Analysis

- Syntax trees obtained from Stanford parser
- Subtree patterns (node relations) in parse trees
  - Context-free-grammar (CFG) Rules (Prior work use only subset)
  - Sister, Dominance
  - Node label, C-command
- Feature extraction
  - Collect statistics over all observed subtree patterns
  - Unlike prior studies, we developed a language independent algorithm



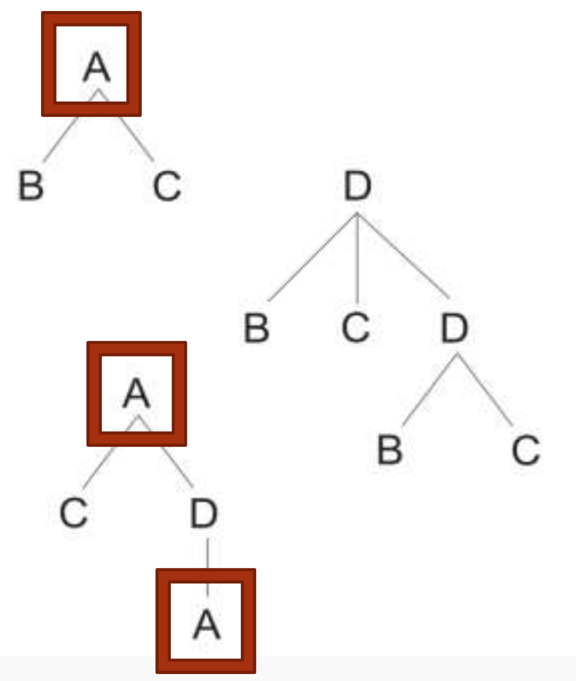
Feature Extraction

- Each sample consists of multiple utterances, therefore multiple parse trees



Feature Extraction

- Each sample consists of multiple utterances, therefore multiple parse trees
- Multiple instances of the same node label
  - Node label A occurs 3 times
  - Node label C occurs 4 times
  - Total nodes = 13
- Rate of node label A = 3/13



Feature Extraction

Each sample consists of multiple utterances, therefore multiple parse trees

Multiple instances of the sister-relation in the sample

- (B,C), (C,D), (B,D)

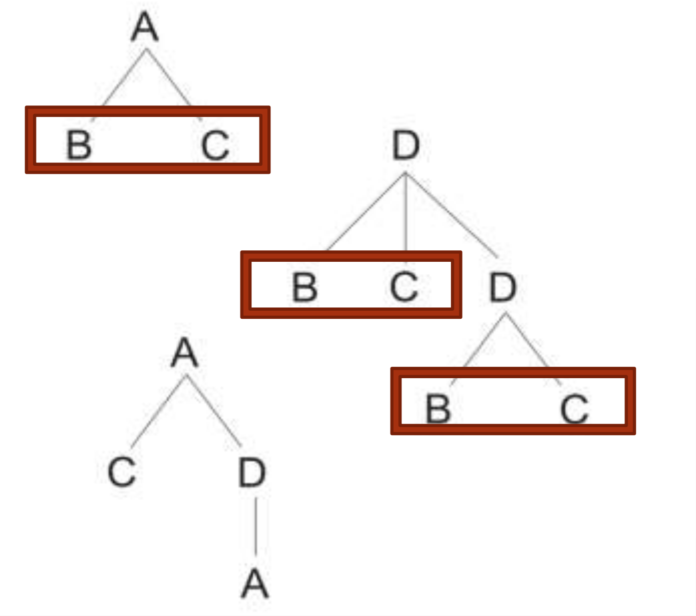
sister(B,C) occurs 3 times

sister(B,D) occurs once

sister(C,D) occurs twice

<u>Total sister relations = 6</u>

- Rate of **sister(B,C)** = 3/6



Feature Extraction

- A rate feature was obtained for each instance of a relation by dividing
  - the count of that instance
    - e.g. sister(B,C) = 3, node(A) = 3
  - by the sum of the counts of all instances of that relation
    - e.g. total sister relations = 6, total nodes = 13
  - Example features = Rate sister(B,C) = 3/6, Rate node-label(A) = 3/13
- **Subtree-patterns:** cfg-rule, sister, dominate, c-command, c-command-via-node, dominate-via-node, node-label

Feature Extraction

Node scores

- Statistical parsing algorithms compute a score between 0 and 1 for each node
  - indicating how grammatical the yield of a node is within the context of the entire sentence
  - We obtained the node scores from Stanford Parser's data structures
- For each node-label:
  - maximum, minimum, standard deviation, skewness and kurtosis
  - e.g. max(NP), min(VP), std(N)

Feature Extraction

- We extracted a rate feature for each observed instance of all relations
    - Thousands of distinct instances!
- We also computed 5 different statistics over node scores for all observed node labels

Feature Selection

- Feature-selection within leave-one-subject-out cross-validation folds without observing the entire data set
- Univariate selection methods
  - Pearson *r*
    - Compute Pearson *r* between each feature and MMSE score
    - Eliminate those features whose Pearson *r* is p-val > 0.01

Feature Selection

- Feature-selection within leave-one-subject-out cross-validation folds without observing the entire data set
- Univariate selection methods
  - Pearson *r*
  - ANOVA(Analysis of Variance) f-test
    - P-values are modelled as an exponential decay curve and those at the tail of the curve are eliminated.

Feature Selection

- Feature-selection within leave-one-subject-out cross-validation folds without observing the entire data set
- Univariate selection methods (Pearson *r* and ANOVA f-test)
- Stability Selection
  - Model the feature scores as an exponential decay curve, and eliminate the features at the tail of the curve

Feature Selection

- Feature-selection within leave-one-subject-out cross-validation folds without observing the entire data set
- Univariate selection methods (Pearson *r* and ANOVA f-test)
- Stability Selection
- Recursive Feature Elimination
  - Estimator is trained on initial set of features and weights are assigned to each.
  - Features with lowest weights are eliminated
  - The process is recursively performed until the pruned set of features are exhausted.

Data and Experiments

- Data:
  - DementiaBank Pitt Corpus
- Experiments:
  - Baseline: CFG rules

	Dementia	Control
Number of samples	278	182
Number of subjects	192	96
Age (years)	72 (8.66)	64 (7.48)
Gender (male/female)	101/177	66/115
MMSE	20 (5.7)	29 (1.1)

Data and Experiments

- Data:
  - DementiaBank Pitt Corpus
- Experiments:
  - Baseline: CFG rules
  - All subtree patterns including CFG rules
  - Only Nodescores

	Dementia	Control
Number of samples	278	182
Number of subjects	192	96
Age (years)	72 (8.66)	64 (7.48)
Gender (male/female)	101/177	66/115
MMSE	20 (5.7)	29 (1.1)

Results: Feature selection

- Number of features drop after application of each selection method
- Column 2 shows median number of features across folds

Baseline: CFG Rules

Total # Features	978	Pearson r	MAE
1 Pearson r <0.01	65	0.60	4.08
2 ANOVA f-test	20	0.61	3.94
3 Stability	18	**0.61**	**3.93**
4 RFE	16	0.60	3.95

All subtree patterns

Total # Features	4297	Pearson r	MAE
1 Pearson r <0.01	468	0.62	3.97
2 ANOVA f-test	99	**0.66**	**3.86**
3 Stability	35	0.64	3.91
4 RFE	35	0.64	3.91

Results: Feature selection

- Number of features drop after application of each selection method
- Column 2 shows median number of features across folds
- Pearson *r* and ANOVA reduce number of features significantly
- RFE has minimal effect as it comes last

Baseline: CFG Rules

Total # Features	978	Pearson *r*	MAE
1 Pearson *r* <0.01	65	0.60	4.08
2 ANOVA f-test	20	0.61	3.94
3 Stability	18	**0.61**	**3.93**
4 RFE	16	0.60	3.95

All subtree patterns

Total # Features	4297	Pearson *r*	MAE
1 Pearson *r* <0.01	468	0.62	3.97
2 ANOVA f-test	99	**0.66**	**3.86**
3 Stability	35	0.64	3.91
4 RFE	35	0.64	3.91

Results: Regression

- Significant improvement over only CFG rules

Baseline: CFG Rules

Total # Features	978	Pearson r	MAE
1 Pearson r <0.01	65	0.60	4.08
2 ANOVA f-test	20	0.61	3.94
3 Stability	18	**0.61**	**3.93**
4 RFE	16	0.60	3.95

All subtree patterns

Total # Features	4297	Pearson r	MAE
1 Pearson r <0.01	468	0.62	3.97
2 ANOVA f-test	99	**0.66**	**3.86**
3 Stability	35	0.64	3.91
4 RFE	35	0.64	3.91

Results: Regression

- Significant improvement over only CFG rules
- Stability Selection decreases number of features significantly but decrease the performance slightly

Baseline: CFG Rules

Total # Features	978	Pearson r	MAE
1 Pearson r <0.01	65	0.60	4.08
2 ANOVA f-test	20	0.61	3.94
3 Stability	18	**0.61**	**3.93**
4 RFE	16	0.60	3.95

All subtree patterns

Total # Features	4297	Pearson r	MAE
1 Pearson r <0.01	468	0.62	3.97
2 ANOVA f-test	99	**0.66**	**3.86**
3 Stability	35	0.64	3.91
4 RFE	35	0.64	3.91

Results: Regression

- Significant improvement over only CFG rules
- Stability Selection decreases number of features significantly but decrease the performance slightly
- State-of-the-art performance comparable to human inter annotator reliability scores

Baseline: CFG Rules

Total # Features	978	Pearson r	MAE
1 Pearson r <0.01	65	0.60	4.08
2 ANOVA f-test	20	0.61	3.94
3 Stability	18	**0.61**	**3.93**
4 RFE	16	0.60	3.95

All subtree patterns

Total # Features	4297	Pearson r	MAE
1 Pearson r <0.01	468	0.62	3.97
2 ANOVA f-test	99	**0.66**	**3.86**
3 Stability	35	0.64	3.91
4 RFE	35	0.64	3.91

Results: Regression

- Significant improvement over only CFG rules
- Stability Selection decreases number of features significantly but decrease the performance slightly
- State-of-the-art performance comparable to human inter annotator reliability scores
- **Node scores alone:**
  - Pearson r: 0.56
  - MAE: 4.28

Baseline: CFG Rules

Total # Features	978	Pearson r	MAE
1 Pearson r <0.01	65	0.60	4.08
2 ANOVA f-test	20	0.61	3.94
3 Stability	18	**0.61**	**3.93**
4 RFE	16	0.60	3.95

All subtree patterns

Total # Features	4297	Pearson r	MAE
1 Pearson r <0.01	468	0.62	3.97
2 ANOVA f-test	99	**0.66**	**3.86**
3 Stability	35	0.64	3.91
4 RFE	35	0.64	3.91

Conclusion

- A novel method for syntactic analysis for assessing cognitive impairments:
  - does not rely on pre-determined set of tree labels, or CFG rules
  - we applied our method to Spanish and German with no modification at all
- Unlike a large number of related studies, feature-selection performed in each cross-validation fold without observing the entire data set.
- State-of-the-art performance comparable to human inter annotator reliability scores

Summary

Three studies on using NLP on clinical interviews

- Semantic analysis
  - First demonstration of utility of metaphor identification for detection of schizophrenia
- Acoustic analysis
  - First study that uses characteristics of speech to identify subjects that are under the influence of MDMA and Oxytocin.
- Syntactic analysis
  - A novel method for syntactic analysis that is language and formalism independent
  - Validated by performing regression to predict the MMSE score