Prediction of Schizophrenia

with *Ngram Tracing* and *Word Graph Analysis*

COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS RESEARCH CENTER CLiPS

Universität Potsdam

lisa.becker@uni-potsdam.de

Lisa Becker and Walter Daelemans

Schizophrenia and NLP?

Disorganized language output is a central symptom of schizophrenia, showing as derailment, neologisms/non-words, schizophasia or alogia. Being able to predict this type of language with machine learning might support diagnosis. We explored two novel methods of authorship attribution and compared them to a supervised machine learning baseline using a bag of words approach with Multinomial Naïve Bayes and Linear Support Vector Machines.

Data

2 written tasks (Dutch), 3 participant groups from Dutch-speaking patients collected by CAPRI (University of Antwerp).

Task 1: "How do you make coffee?"

Task 2: "Please describe this picture."

1. **C**: Controls
2. **PA**: Acute Patients
3. **PR**: Patients in Remission (e.g. under medication)

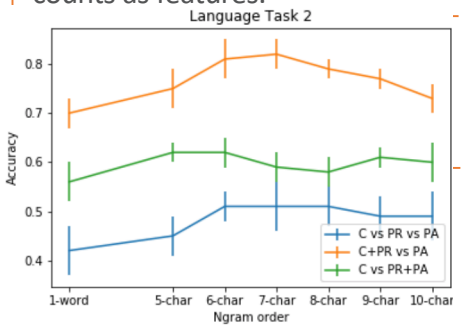
Supervised Machine Learning

The baseline was classification with Multinomial Naïve Bayes and Linear Support Vector Machine using word and lemmata counts as features.

Baseline

Task, Clf	CvsPRvsPA	C+PRvsPA	CvsPR+PA
1, NB	.77 (±.05)	.81 (±.05)	.80 (±.05)
1, SVM	.87 (±.06)	.89 (±.07)	**.90 (±.05)**
2, NB	.67 (±.05)	.95 (±.03)	.72 (±.04)
2, SVM	.83 (±.03)	**.96 (±.01)**	.85 (±.04)
both, NB	.71 (±.04)	.83 (±.03)	.74 (±.05)
both, SVM	.84 (±.04)	**.91 (±.02)**	.89 (±.04)

Table 1: Accuracy with word counts.



1. Method

Figure 1: Accuracy averaged over 10 iterations of calculating the overlap coefficient for each combination of a participant and group.

Ngram Tracing

Grieve and colleagues (2017) developed the method of Ngram Tracing, which computes the overlap of ngrams in the test item (participant, P) and participant group (G) and divides by the total amount of ngrams in the test item:

$$\frac{|P \cap G|}{|P|}$$

This overlap coefficient is computed with every participant group. The test item is classified as the participant group with the highest overlap coefficient.

Word Graph Analysis

Mota and colleagues (2012) developed the freely accessible SpeechGraph Software that represents texts as graphs with each word represented as a node. The resulting parameters linked to the graph representation were fed into a Multinomial Naïve Bayes classifier and a Linear Support Vector Machine.

2. Method

Task	Clf	C vs PR vs PA	C+PR vs PA	C vs PR+PA
1	NB	.68 (± .00)	.56 (±.02)	.56 (±.02)
	SVM	.74 (±.00)	.74 (±.00)	.73 (±.00)
2	NB	.52 (±.02)	.65 (±.03)	.58 (±.03)
	SVM	.79 (±.00)	**.96 (±.00)**	.79 (±.00)
both	NB	.76 (±.01)	.83 (±.05)	.76 (±.05)
	SVM	/	/	/

Table 2: Accuracy with 30-word-windows.

Discussion / Conclusion

Neither Ngram Tracing nor Word Graph Analysis improve over the baseline. Due to the small amount and skew of the data, there was strong overgeneralization towards the majority class (C) which is not represented in the accuracy.

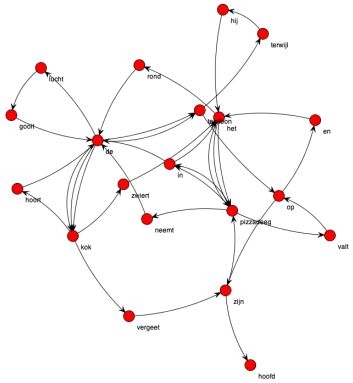


Figure 2: Example of a SpeechGraph representation:
NL: "De kok zwiert het pizzadeeg in het rond. De kok hoort de telefoon terwijl hij het pizzadeeg in de lucht gooit. De kok vergeet zijn pizzadeeg neemt de telefoon op en het pizzadeeg valt op zijn hoofd."
Features include the number of nodes/edges, longest distance, ... etc.