Distributional Semantics for Medical Information Extraction

Lautaro Quiroz[1], Lydia Mennes[2], Mostafa Dehghani[1], Evangelos Kanoulas[1]

[1]University of Amsterdam, [2]CTcue

CLEF eHealth - Medical Information Extraction

- Problem: Fill in a structured form by extracting text-snippets from a free-text report of nursing handover
- Approach: Multi-class classification of each written token
- Methods: (1) A Feed-Forward Neural Network, (2) An Ensemble Method

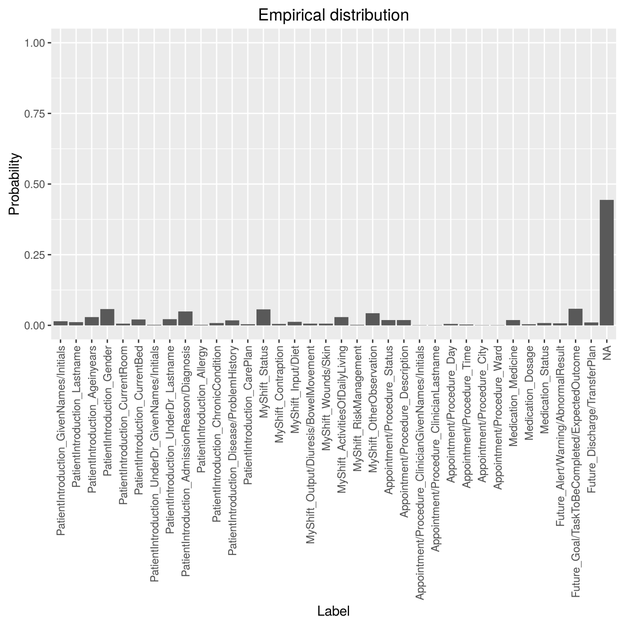
Data Collection

A training, validation, and test set of free-text nursing handover reports, with each word labeled by one of 39 classes.

Dataset	# Docs	# Tokens	# Unique Words	Word overlap w/ stopwords	Word overlap w/o stopwords
Training	101	7451	1347	-	-
Validation	100	6798	1291	645 (49.96%)	560 (43.38%)
Testing	100	5741	1213	527 (43.45%)	453 (37.35%)

Table: Datasets overview (after punctuation removal).

Figure: Training data empirical distribution.



Data Characteristics

- Skewed dataset towards the *"NA"* label
- Large number of output classes
- Not enough data training data.

A Feed-Forward Neural Network

Use a feed forward neural network that processes a context window of tokens and discriminates among the 39 labels.

Figure: Pipeline

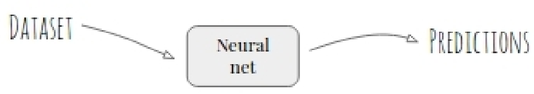
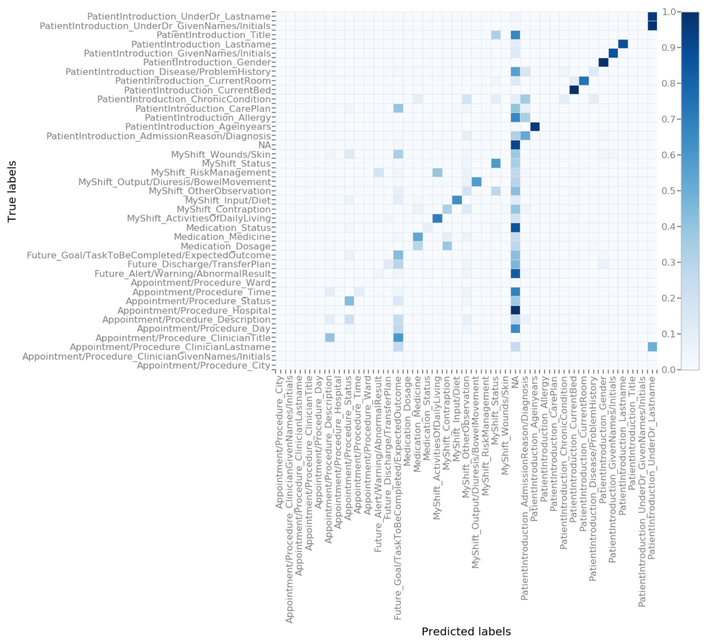


Table: Results

Dataset	Macro average			Micro average			NA		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Training	0.741	0.591	0.624	0.908	0.862	0.884	0.920	0.979	0.948
Validation	0.468	0.344	0.355	0.636	0.495	0.557	0.696	0.920	0.793
Testing	0.411	0.307	0.308	0.563	0.472	0.514	0.723	0.894	0.800

Figure: Validation set confusion matrix.



- Almost every category is being misclassified as *"NA"*.
- *"Expected outcome"*, *"Diagnosis"*, and *"Other observation"* appears as the most conflicting categories, after *"NA"*.
- Categories with best F1 scores are: *"Current room"*, *"Current bed"*, and *"Age in years"*.

An Ensemble Method

Use a random forest to distinguish *"NA"* samples from *"not NA"*. Next, use a feed forward neural network that take those *"not NA"* samples and further classify them among the remaining categories.

Figure: Pipeline.

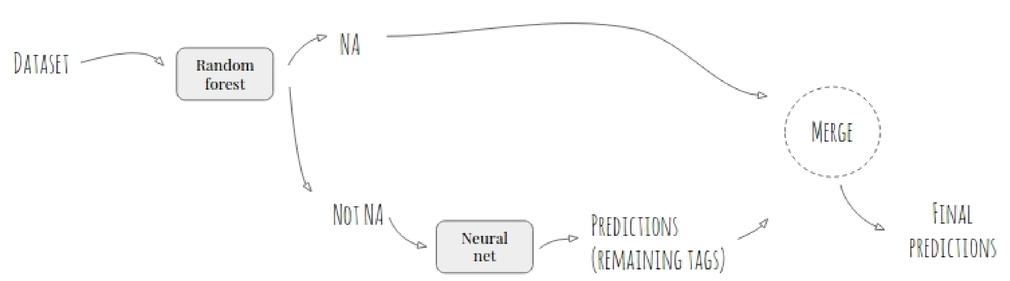
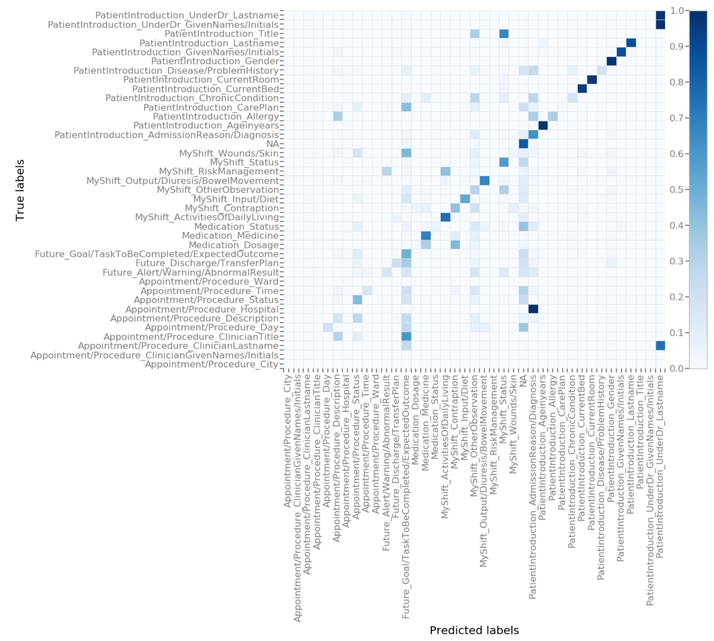


Table: Results

Dataset	Macro average			Micro average			NA		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Training	0.768	0.699	0.718	0.810	0.861	0.835	0.859	0.791	0.824
Validation	0.434	0.397	0.385	0.541	0.546	0.543	0.846	0.835	0.840
Test	0.425	0.383	0.345	0.490	0.517	0.503	0.849	0.779	0.813

Figure: Validation set confusion matrix.



- The most conflicting categories are *"Expected outcome"*, *"Diagnosis"*, and *"NA"*.
- Categories with best F1 scores are: *"Current room"*, *"Current bed"*, and *"Age in years"*.

Conclusions

We made use of two machine learning approaches to classify tokens among 39 nurse handover form labels.

- They make use of words as units of prediction, and
- They avoid hand-crafted feature engineering.

Conclusions from the analysis:

- Some labels are harder to classify than others.
- Distributional semantics help in this classification task.
- The Ensemble method presents many fewer *"NA"* false positives (reduction of 60%), with only a small number of true positives decrement (reduction of 13%). This improves the *"NA"* F1 and the remaining tags macro F1 score.
- Highest F1 gain categories include: *"Bowel movement"*, *"Transfer plan"*, *"Dosage"*, *"Procedure description"*, and *"Other observation"*.
- Highest F1 decrements ($\approx 10\%$) are perceived in categories like: *"Diet"*, *"Status"*, and *"Diagnosis"*.
- The Neural Net's most conflicting category is *"NA"*, while the Ensemble's is *"Expected outcome"*.

UNIVERSITY OF AMSTERDAM