Feature Extraction for Pathology Reports Classification with Precise Negation Scope Detection

Olga Zamaraeva, Kristen Howell, & Adam Rhine
University of Washington
Department of Linguistics

DELPH-IN Summit in Oslo, Norway
August 7 2017

Overview

- Broad task: Classify pathology reports as lymph node positive or negative
  - Did the cancer spread to the lymph nodes?
- Core task: find concepts which could be important features (e.g. *malignancy*) in negated context
  - "*No malignancy was found in the lymph nodes.*"
  - "*Metastasis: not identified.*"

RQ 1: How good ERG is for negation detection?

- McKinlay et al (2012)
  - Used ERG to find negation scope
    - Evaluated on sentence level (intrinsic evaluation of negation scope detection)
    - Data: Biomedical literature
    - Events of interest are already identified in the data in previous stages.
      - Feature vectors are constructed for these events, and then the events are classified in terms of negation scope.
- Packard et al. (2012)
  - Negation scope in Sherlock Holmes stories.

RQ2: Is negation important (for classification)?

- …**and if so, can we classify better with ERG?**
- Maybe not:
  - Words that occur negated as well as not negated will likely not be selected as most informative
- Maybe yes:
  - Negated concepts may be good features themselves
  - (Also, machine learning classification is not the only scenario)

Dataset

- SEER program (https://seer.cancer.gov/)

	Total	LN	POS	NEG	UNK
Training	581	298	52	91	155
Test	293	137	26	44	67

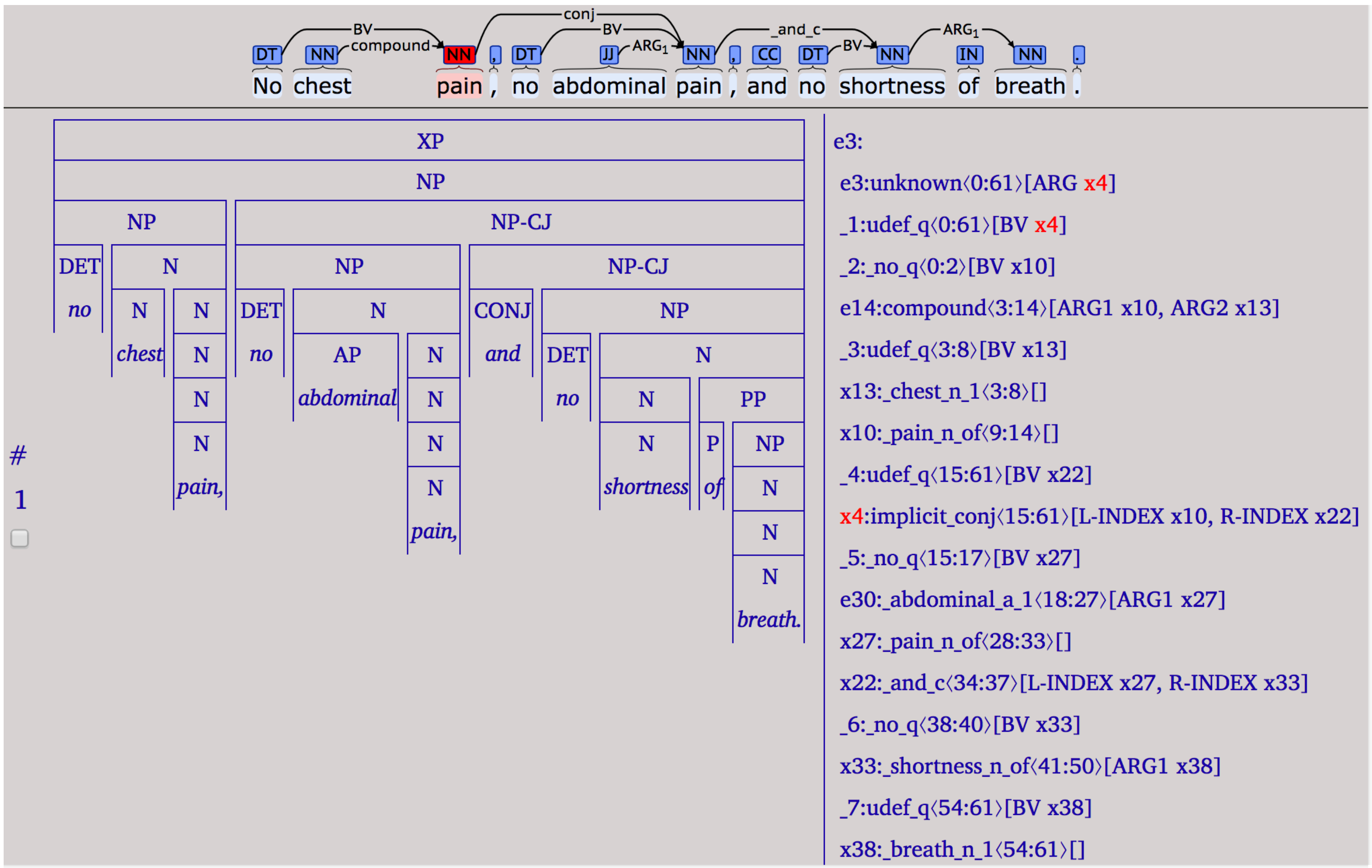
- Annotated on report level
  - Not annotated for negation scope
  - Makes our approach amenable only to extrinsic evaluation

(Extrinsic) Evaluation

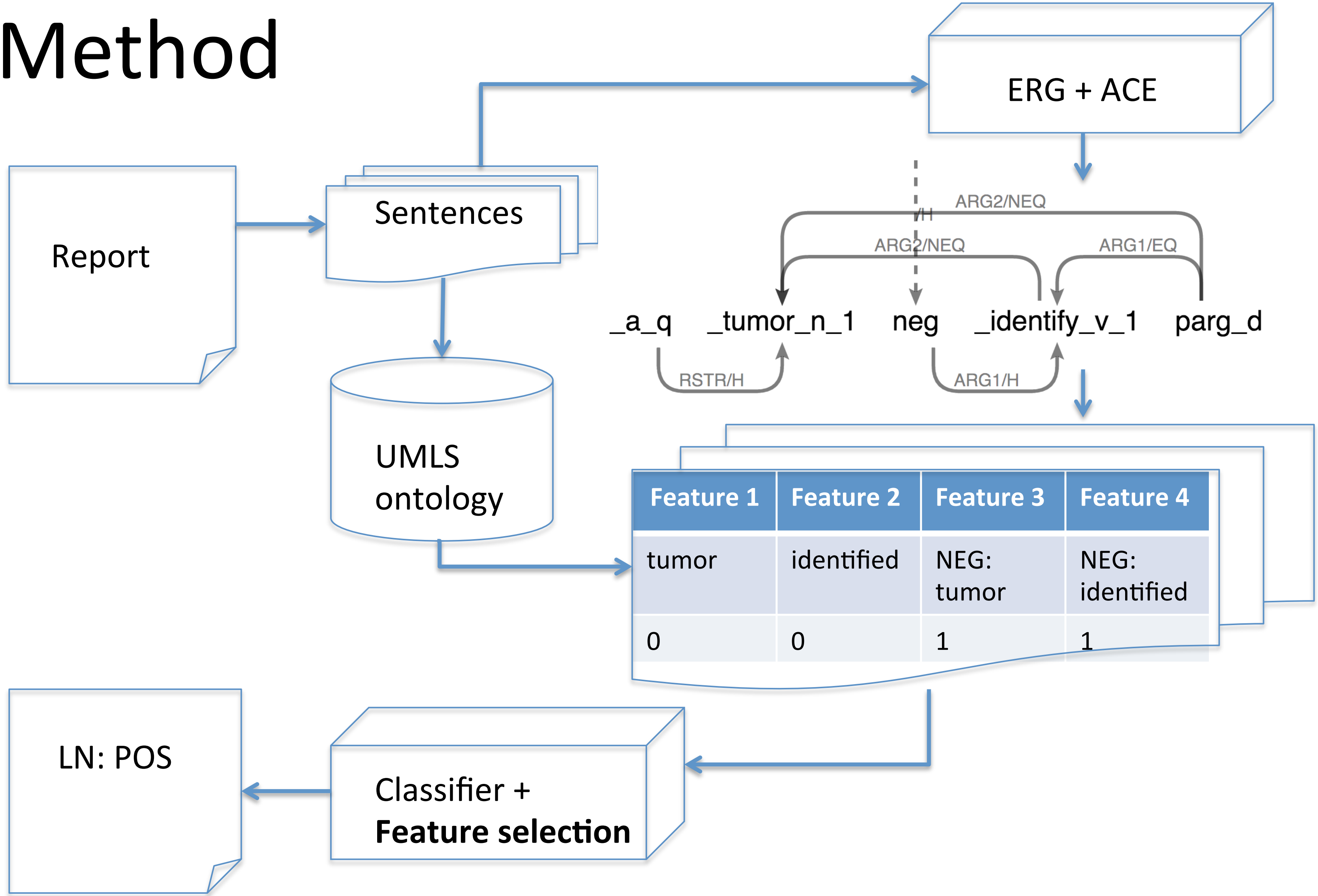
- Compare to NegEx
  - Chapman et al. (2002)
  - Regular expressions-based
  - Very widely used
    - Easy to adapt to any English dataset
- Comparing rule-based to rule-based*
  - Independent of the dataset?
  - Not too many added heuristics?

*negation detection

Example where NegEx fails



Method



ERG + ACE
Report
Sentences
ARG2/NEQ
/H
ARG2/NEQ
ARG1/EQ
_a_q
_tumor_n_1
neg
_identify_v_1
parg_d
RSTR/H
ARG1/H
UMLS
ontology
Feature 1
Feature 2
Feature 3
Feature 4
tumor
identified
NEG: tumor
NEG: identified
0
0
1
1
LN: POS
Classifier +
Feature selection

Feature Selection

Selection algorithm		with NegEx	with ERG
Variance Threshold (90%) (VT)	total	356	368
	negated	22	6
K-best (K=100)	total	100	100
	negated	17	7
Best percentile (10%)	total	488	408
	negated	88	13
Wrapper: AdaBoost (AB)	total	37	37
	negated	1	2
Wrapper: Random Forest (RF)	total	489	515
	negated	79	24
Wrapper: Logistic Regression (LR)	total	1257	1061
	negated	186	37

(Null? Incremental?) Results

Scikit
Learn

Classifiers

Feature selectors

FS→	VT		AB		RF		K=100		10%		LR	
CL	NE	ERG	NE	ERG	NE	ERG	NE	ERG	NE	ERG	NE	ERG
AB	0.74	**0.76**	0.71	**0.73**	0.69	**0.74**	**0.69**	0.64	0.69	**0.72**	**0.77**	0.69
DT	0.66	**0.72**	0.68	**0.76**	**0.69**	0.69	**0.70**	0.69	0.74	**0.75**	**0.74**	0.66
KNN	**0.67**	0.61	0.73	**0.76**	0.66	**0.69**	**0.71**	0.69	**0.69**	0.68	0.61	0.61
SVM	0.74	**0.74**	0.76	**0.78**	**0.77**	0.74	0.72	**0.74**	0.75	**0.78**	0.74	0.74
NB	0.64	**0.66**	**0.42**	0.39	0.69	0.69	0.34	**0.36**	0.64	**0.68**	0.55	**0.59**
NN	0.70	**0.72**	0.72	**0.80**	**0.77**	0.73	0.72	0.72	0.74	**0.78**	0.73	0.73
RF	0.73	**0.75**	0.76	**0.80**	**0.78**	0.77	0.74	0.74	0.76	**0.77**	0.75	**0.76**

F1 micro-average classification scores

Results on more constrained models

FS→	AB			RF			GB		
CL	BL	NE	ERG	BL	NE	ERG	BL	NE	ERG
AB	0.72	**0.75**	0.73	**0.74**	0.72	0.74	**0.73**	0.72	0.73
RF	**0.77**	0.76	0.74	**0.78**	0.75	0.75	0.72	0.71	**0.73**
GB	**0.77**	0.77	0.77	**0.77**	0.77	0.73	0.74	**0.76**	0.75
VT	**0.77**	0.77	0.76	**0.77**	0.75	0.75	0.74	**0.75**	0.74

Micro-average F1 scores

- No improvement from adding any negated features

Better(?) results

- Same dataset, filtered for "Laterlality Category": Left or Right (lung, breast etc).
  - "Better" (bigger, more balanced) dataset,
    - but still small, reports are still different length etc.

FS→	AB			RF			GB		
CL	BL	NE	ERG	BL	NE	ERG	BL	NE	ERG
AB	**0.87**	0.86	0.86	0.83	0.77	**0.87**	0.84	**0.86**	0.85
RF	0.86	0.87	**0.87**	0.77	0.79	**0.86**	0.85	0.85	**0.86**
GB	**0.87**	0.87	0.87	0.80	0.78	**0.84**	**0.87**	0.87	0.87
Vt	**0.87**	0.87	0.86	0.81	0.79	**0.86**	0.86	**0.86**	0.86

Micro-average F1 scores

Issues

- Sentences are fragmented
  - Easy to make mistakes in tokenization
  - Difficult to parse in a meaningful way
- Dataset is small and unbalanced
  - Few ML algorithms are effective
  - **Very** hard to select features/not overfit

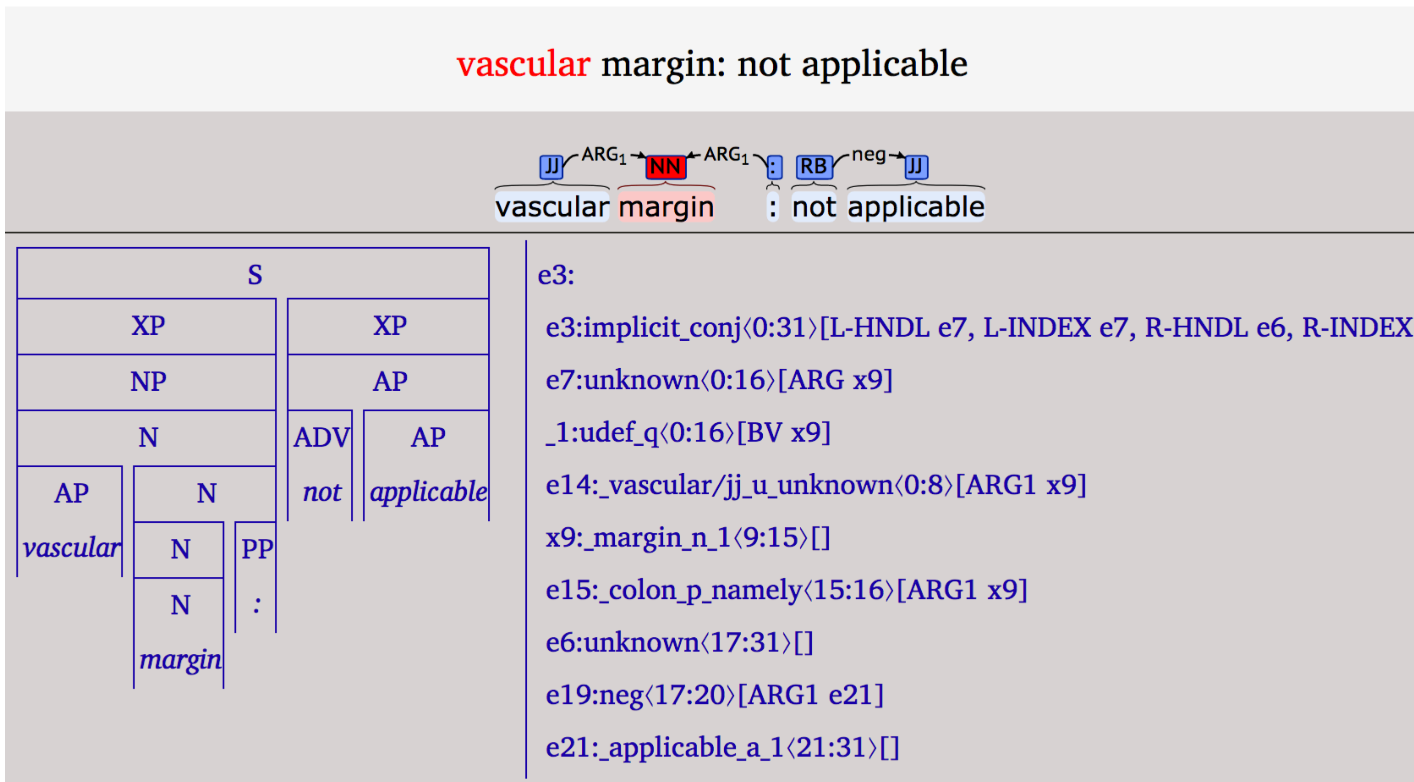
- All changes/tuning that we tried so far did not lead to much change in the results
  - Some negated features are among the selected ones but not many of them

Parse Issues

- Parse coverage
  - 79% (22K/28K)
    - E.g. *Despite an FDA approved scoring guide that classifies 2+ immunohistochemistry results as equivocal, the literature suggests that up to 70% of these cases may be actually the fluorescence in situ hybridization (FISH) negative.*
    - E.g. *Necrosis of invasive component: not identified.*
  - fall back?
    - if fall back to NegEx, it adds more feature tokens than ERG (6K vs. 4K)
  - Compare only sentences for which have a parse?
    - Results still seem null/incremental
- Parse selection
  - Often one of the first parses is good but not the top parse.
    - E.g. *No **X or Y** were identified.*
      - Though, selecting widest scope did not help classification
    - Different top parse with fragmenting on/off
- MRS crawling is basic
  - Just looking at *neg* relation and its ARG1, ARG2 and *no* quantifier for nouns

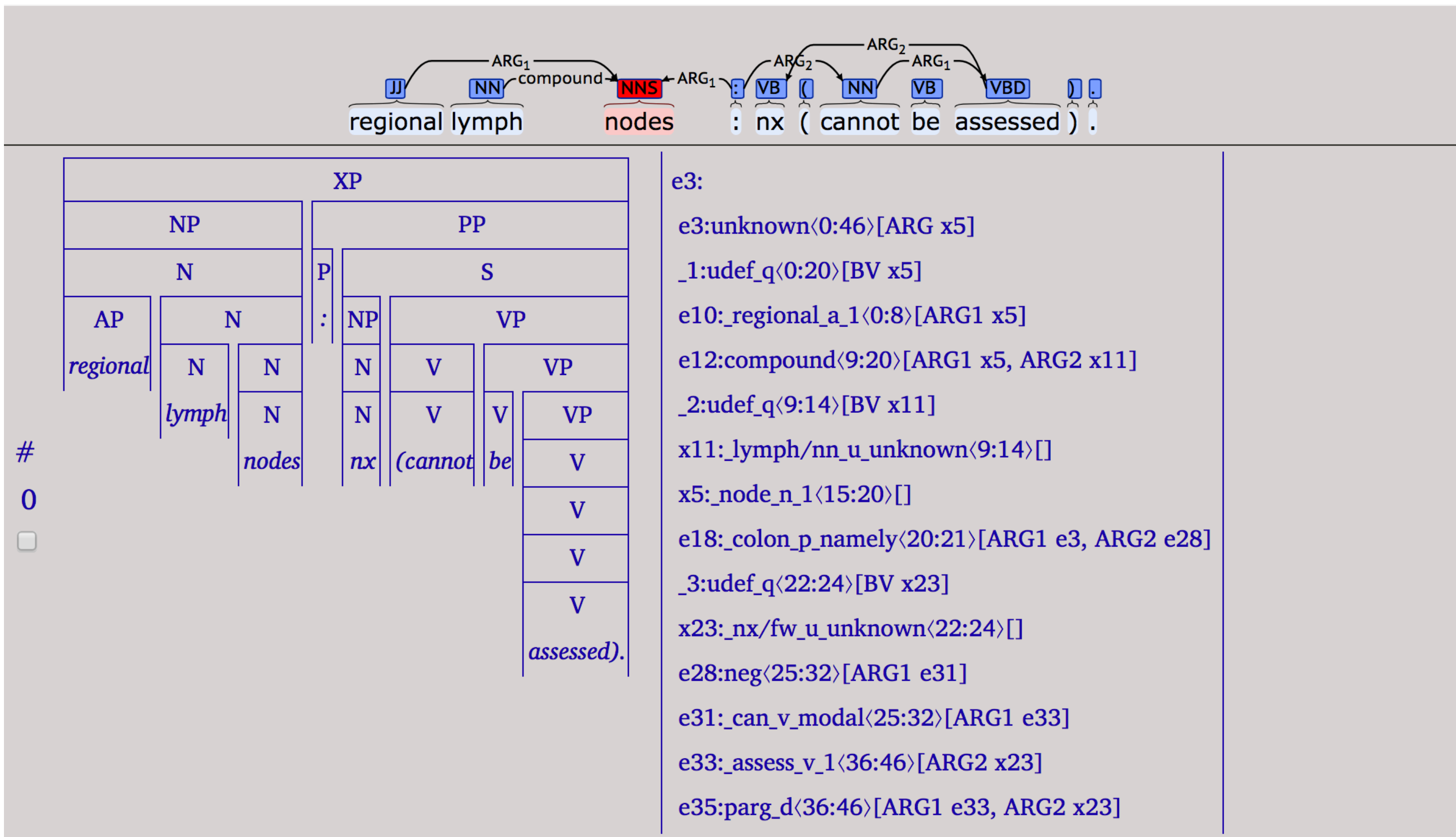
Parse selection issues

- *Tumor: not identified*
  - Often times, *identified* is parsed as a post-head adjective (*not_c*) rather than a predicate (*neg).*



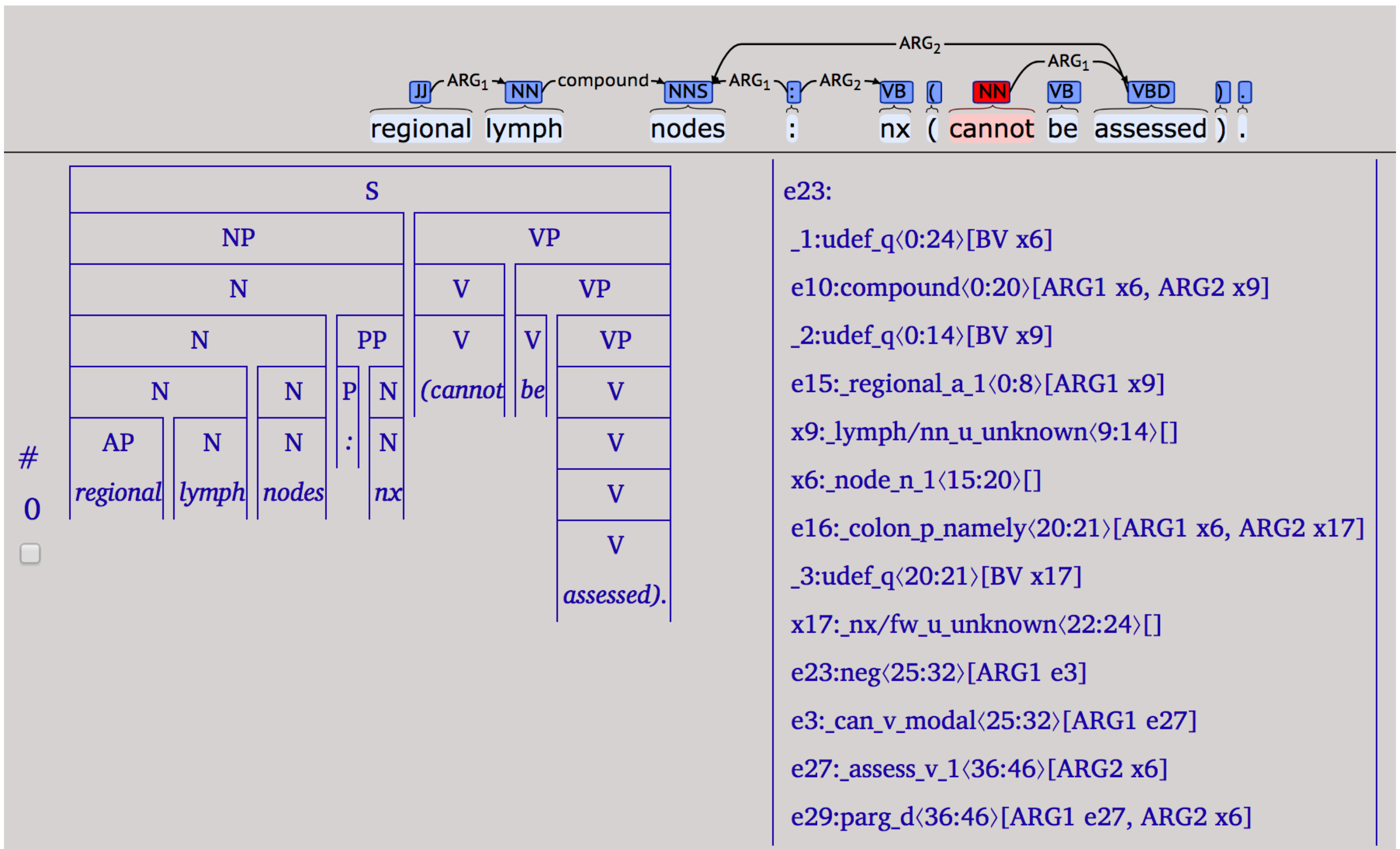
Parse selection issues

- Passive voice with modals: top-ranked parse with fragments on:



Parse selection issues

- Passive voice with modals: top-ranked parse with fragments off:



Conclusion

- ERG should be more useful than something like NegEx for finding negated items.
  - Without too many heuristics for crawling MRS?..
    - Even with heuristics, MRS is more general than surface strings
- How to achieve results that clearly show this?
  - So far, improving parse selection did not help.
    - Need to work more on tuning parameters for classifiers and (especially) feature selection