Potential Use and of Machine Learning Approaches to Extract Adverse Drug Reactions from Product Labels

Rave Harpaz
Senior Research Scientist, Oracle Health Sciences

ORACLE HEALTH SCIENCES

ORACLE

TAC ADR - Task 3

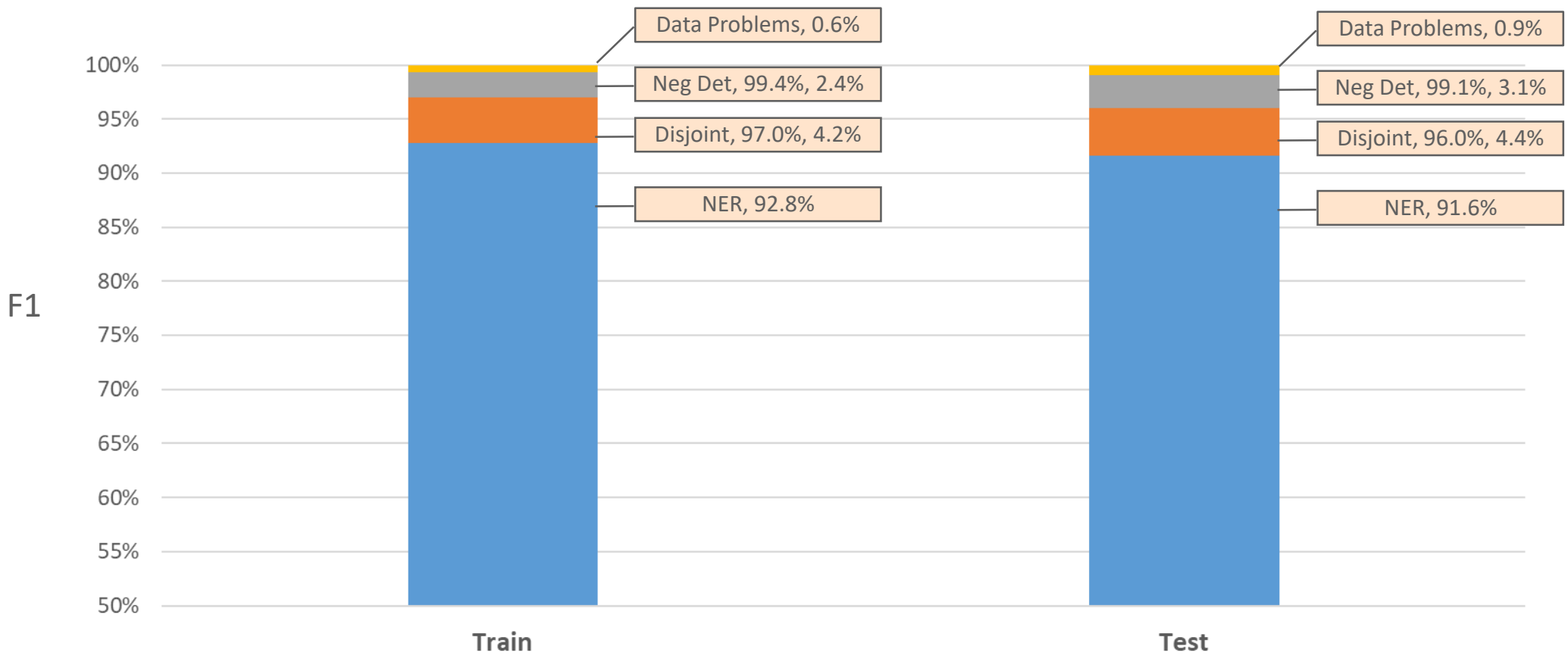
*Identification (*verbatim*) of positive mentions of adverse reactions*

- *positive* - AR that is not *Negated* and is not related by a *Hypothetical* relation to a *Drug Class* or *Animal*

1. NER of ARs (CRFs)
2. NER of disjoint ARs (rule-based)
3. Negation Detection (rule-based)

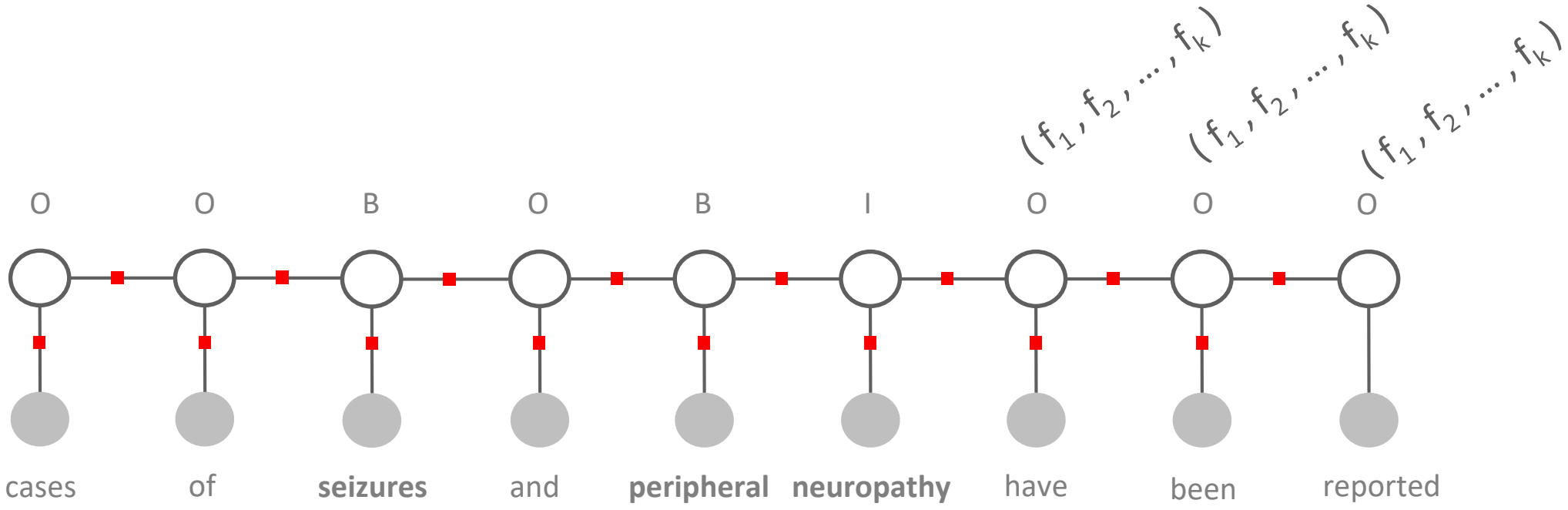
* pre-processing (sentence segmentation, tokenization, POS tagging)
+ post-processing

Problem composition



NER via Conditional Random Fields (CRFs)

- A CRF models text as a sequence labeling joint classification problem



Tag set: B-beginning, I-Inside, O-Outside

Features

- **Word identity**, **contextual** (neighboring words), **POS**, **word shape** (capitalization, suffix, num patterns)
- **Sentence type**:
  - *List* - trigger terms (‘including’) + punctuation (‘:’) + ratio of tokens to commas in a sentence
  - *Table* - ratio of tokens to spaces-between-tokens in a sentence
  - *Header/Bullet* - special charters (‘*’) + numeric string patterns (‘3.1’) at the beginning of a sentence
- **Lexical** (dictionary membership) :
  - *MetaMap* – originally 17K terms, cleaned 15K (‘the of mtc’, ‘5.6’, ‘follow-up examinations’, ‘3 men’)
  - *ADR annotator* – UMLS (med disorders) 10K terms
  - *Dictionary expansion by fuzzy matching*: standard, subset, and token order (MetaMap -> 2K, ADR -> 3K)
    - hypoglycemia episode - hypoglycemic episode
    - thyroid-binding globulin increase - thyroxin binding globulin increased
    - cholesterol elevation - elevated cholesterol

Features

		precision	recall	F1	F1-PER
	Baseline	77.4%	67.2%	71.9%	
A	POS	77.1%	68.5%	72.6%	2%
B	Shape	75.9%	67.6%	71.5%	-1%
C	Sentence type	79.9%	69.1%	74.1%	8%
D	MetaMap	77.0%	69.7%	73.2%	4%
E	ADR annotator	76.9%	70.0%	73.3%	5%
F	MetaMap + derived terms	77.3%	70.7%	73.9%	7%
G	ADR annotator + derived terms	76.8%	70.4%	73.5%	6%
	One word pre/post	85.1%	74.0%	79.1%	26%
	Two words pre/post	86.0%	75.7%	80.5%	31%
	Three words pre/post	86.7%	75.6%	80.8%	31%
	Three words pre/post + A	86.7%	77.0%	81.6%	3%
	Three words pre/post + B	85.5%	77.0%	81.0%	1%
	Three words pre/post + C	86.6%	75.9%	80.9%	0%
	Three words pre/post + D	86.4%	79.0%	82.6%	6%
	Three words pre/post + E	86.3%	79.1%	82.6%	6%
	Three words pre/post + F	86.7%	79.8%	83.1%	8%
	Three words pre/post + G	86.0%	79.8%	82.8%	7%
	Three words pre/post + D + E	86.4%	80.1%	83.1%	8%
	Three words pre/post + F + G	86.6%	80.5%	83.4%	9%
*	Three words pre/post + A + B + C + F + G	86.8%	81.4%	84.0%	11%

performance stats for B/I classification

CRF ensemble

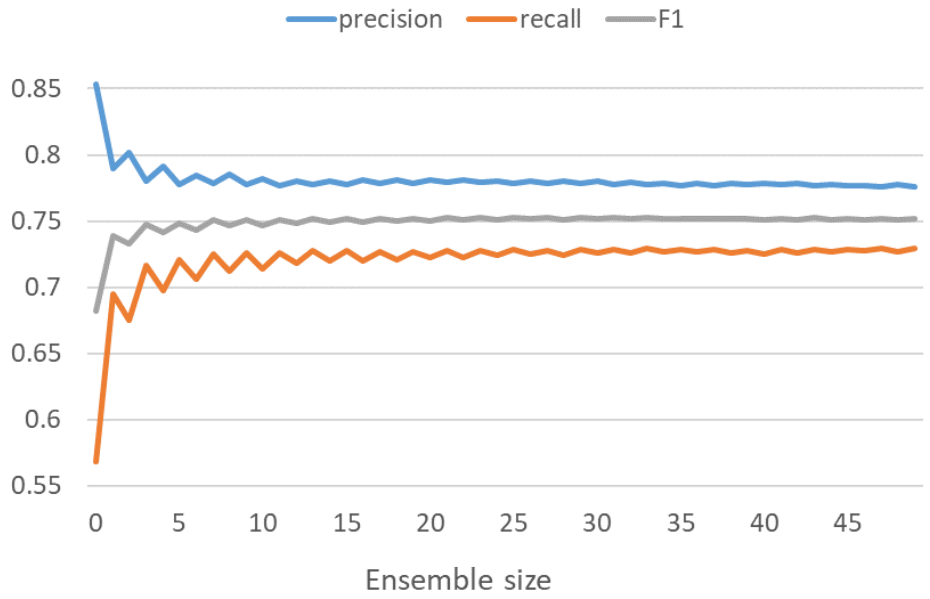
Aim: increase recall (e.g., prec=86% vs rec=79%)

Input: X – training sentences, Y – training sentence labels
Initialize: $C^+ = C^- = (X, Y)$, ensemble $M^* = \emptyset$
For k = 1 to K
  X^+, Y^+ = sample with replacement $(1 - p_{miss}) \times |X|$ sentences and their labels from C^+
  X^-, Y^- = sample with replacement $p_{miss} \times |X|$ sentences and their labels from C^-
  M_k = train base-CRF model on $(X^+ \cup X^-, Y^+ \cup Y^-)$
  apply model M_k to X
  e_k = error of M_k applied to X
  $M^* = M^* \cup (M_k, e_k)$
  C^+ = set of sentences and their labels that have been misclassified by M_k when applied to X
  C^- = set of sentences and their labels that have been correctly classified by M_k when applied to X
Output: M^*

	1	2	3	4	5	6	7	8	9	10	
M1	O	O	B	I	I	I	I	O	O	O	e_1
M2	O	O	B	I	O	B	I	O	O	O	e_2
M3	O	O	O	B	O	B	I	O	B	I	e_3
M4	O	O	B	I	O	B	I	O	O	O	e_4
M5	O	O	O	B	I	I	I	O	B	I	e_5
ensemble	O	O	B	I	O	B	I	O	O	O	

A. a sentence is misclassified if any of its predicted token labels is a false positive
B. a sentence is misclassified if any of its predicted token labels is a false negative
C. a sentence is misclassified if either A or B are true

CRF ensemble – NER performance



	training 5-fold CV		
	precision	recall	F1
single CRF model	86.8%	81.4%	83.98%
CRF ensemble - R	85.2%	84.4%	84.80%
CRF ensemble - PR	87.3%	81.1%	84.08%
LSTM	82.0%	84.3%	83.10%

	test		
	precision	recall	F1
single CRF model	83.3%	82.0%	82.67%
CRF ensemble - R	81.6%	83.9%	82.73%
CRF ensemble - PR	84.1%	81.0%	82.53%
LSTM	80.5%	84.9%	82.66%

performance stats for B/I classification

Disjoint NER

Psychiatric Disorders : Suicidal ideation , attempt , behavior , or completion

left part		right part
suicidal	17	[behavior (7), ideation (4), behaviors (4) attempt (1), completion (1)]
injection site	34	[pain (4), irritation (4), bruising (3), erythema (3), pruritus (2), …]

right part		left part
alt	15	[elevations in (4), elevated (2), increase in (2), increased (2), abnormalities in (1), …]
reactions	15	[anaphylactic (7), hypersensitivity (3), anaphylactoid (2), infusion-related (1), infusion (1), …]

'\bsuicidal\b[\s,]+(?:\w+[\s,]+){1,5}\bOTHER_WORD\b'

OTHER_WORD ∈ {*behavior, behaviors, ideation, attempt, completion*}

* 20 different regex, covering 22% of disjoint mentions

Disjoint NER

Top 5

training				test			
mentions		labels		mentions		labels	
suicidal behavior	33	suicidal behavior	7	suicidal behavior	42	suicidal behavior	9
anaphylactic reactions	13	anaphylactic reactions	7	suicidal behaviors	18	injection site pruritus	6
swallowing difficulties	9	injection site irritation	5	suicidal ideation	10	injection site pain	6
suicidal ideation	9	injection site pain	5	injection site pain	10	injection site erythema	5
suicidal behaviors	8	suicidal behaviors	4	injection site pruritus	9	alt elevations	4

	precision	recall	F1
NER non-disjoint (Basline)	95.7%	90.1%	92.8%
Disjoint NER (Alg)	95.0%	91.8%	93.4%
Disjoint NER (Sys)	95.3%	98.9%	97.0%
Alg only	68.2%	86.6%	76.3%

13% distance covered

* No improvement on test set (P=43%, R=66%, F1=52%)

Negation detection

- ConText algorithm:
  - Searches for modifiers (e.g., ‘no’) that appear within a pre-specified token distance from a named entity
  - Requires as input a list of modifiers and their positioning (pre/post named entity)

Top 5

Negation (20)		Factor (11)		Animal (7)		DrugClass (98)	
no	46	placebo	16	rats	16	laba	16
not	14	other than	2	rodents	7	cocs	12
excluding	12	number is too small	2	rabbits	6	tnf blockers	12
without	6	not possible to determine	2	animal	4	corticosteroids	9
not evident	2	reduce the risk	1	mice	4	gadolinium-based contrast agents	9

Chapman WW et al. Context: An Algorithm For Determining Negation, Experiencer, And Temporal Status From Clinical Reports. J Biomed Inform. 2009 Oct;42(5). 839-51

Negation detection

Modifier performance

	precision	recall	F1	TP	Pred	TPR
Baseline (NER)	95.3%	98.9%	97.0%	6958	7302	
no (Negation)	95.7%	98.8%	97.2%	31	34	91%
without (Negation)	95.3%	98.8%	97.1%	5	7	71%
not (Negation)	95.5%	98.2%	96.8%	61	106	58%
placebo (Factor)	95.3%	97.1%	96.2%	132	255	52%
rats (Animal)	95.6%	98.9%	97.2%	25	25	100%
laba (DrugClass)	95.3%	98.9%	97.1%	2	2	100%
cocs (DrugClass)	95.4%	98.9%	97.1%	10	10	100%
tnf blockers (DrugClass)	95.3%	98.8%	97.0%	3	6	50%

	precision	recall	F1
NER (Basline)	95.3%	98.9%	97.0%
Negation Det (Alg)	97.8%	98.1%	97.9%
Negation Det (Sys)	99.9%	98.9%	99.4%
Alg only	81.8%	72.0%	76.6%

39% distance covered

* 32% improvement on test set (P=80%, R=68%, F1=73%)

Error analysis (training)

- NER (FP)
  - Boundary detection: *hot sensation burning sensation feeling of heaviness*
  - Frequent terms that are not ARs: *disease*, *syndrome*, *outcome*
  - Section sub/heading : * *Hypersensitivity reactions : anaphylaxis, angioedema, urticaria, and …*
  - Measurements: *AST levels > 8 * ULN , WBC counts > 100,000 x 10 6 /L*
  - Extra spaces & characters: *raynaud 's phenomenon*, *anxiety d*
- NER (FN)
  - First/last word & short sentences: * *Thromboembolic events*
  - Table entries: Decreased Potassium 41 (11 %) 23 (6 %)
  - 67% of terms not in lexical resources

Error analysis (training)

- Disjoint NER
  - Relation scope (FP): *injection site pain, and eyelid edema*
  - Relation scope (FN): *application site discomfort or irritation , …7 words… , eyelid irritation and crusting*

- Negation Detection
  - modifier scope/subject (FP):
    - *in SJIA trials no patients discontinued due to hypersensitivity reactions*
    - *QT prolongation was observed on Day 8, … , with no QT prolongation observed on Day 1*

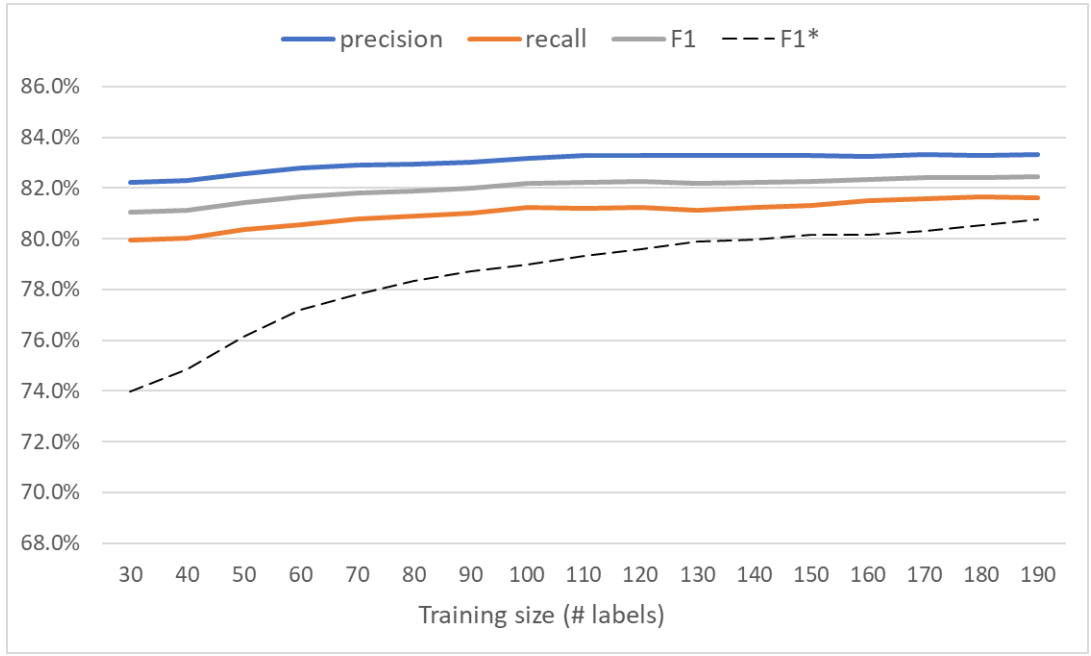
Positive AR identification – steps 1-3

	training 5-fold CV		
	precision	recall	F1
single CRF model	88.9%	79.8%	84.10%
CRF ensemble - R	87.9%	81.3%	84.51%
CRF ensemble - PR	89.4%	79.7%	84.31%

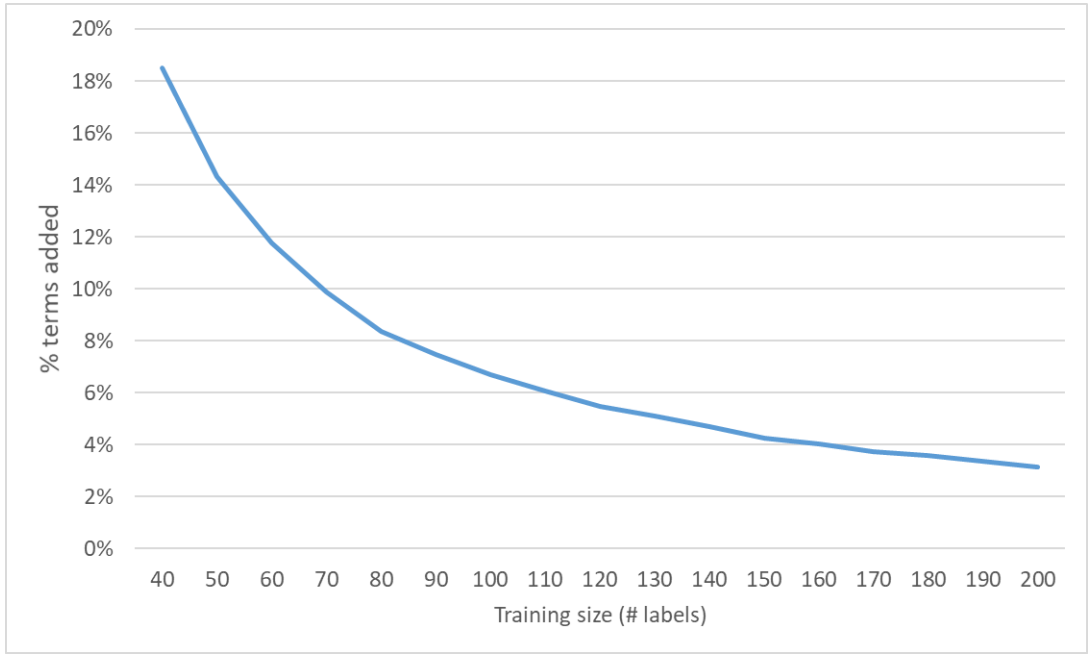
	Micro			Macro		
	precision	recall	F1	precision	recall	F1
single CRF model	81.28%	79.32%	80.28%	81.10%	78.81%	79.20%
CRF ensemble - R	81.18%	79.69%	80.43%	81.47%	79.28%	79.67%
CRF ensemble - PR	82.71%	78.05%	80.31%	82.64%	77.73%	79.42%

- Recall ensemble performed best
- Steps 2-3 added 1% to F1 over the NER step
- 3% lost to NER going from training to testing

Training data size & AR vocabulary growth



Each point 5-fold CV, labels set permuted 20X, no regularization



labels set permuted 500 times