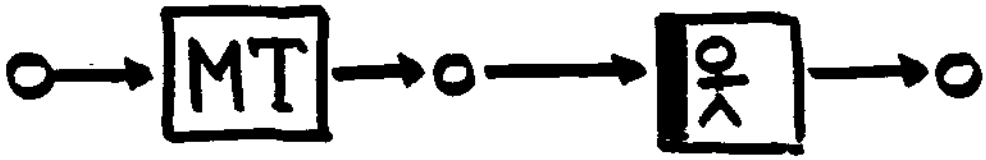
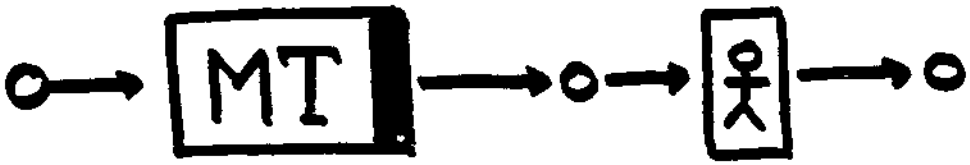
AUTOMATED POSTEDITING OF ENGLISH

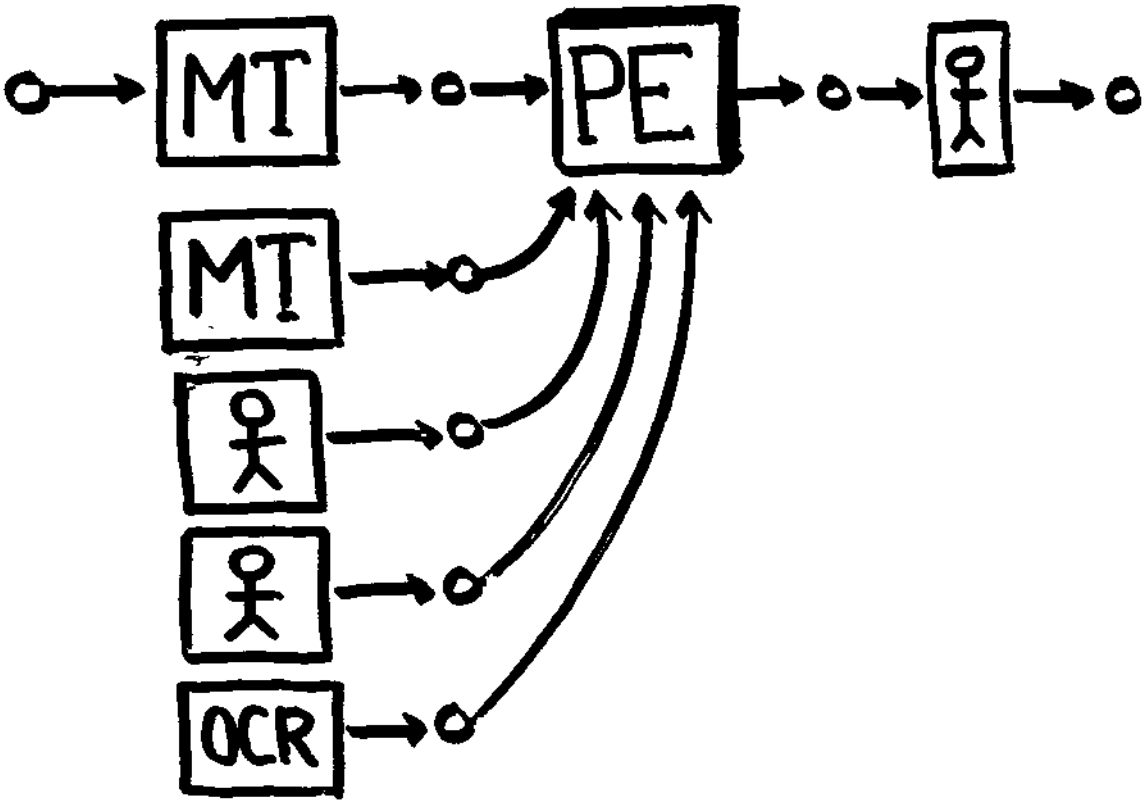
KEVIN KNIGHT, ISI
ISHWAR CHANDER, USC

POST-EDITING





DETACHABLE P.E. MODULE

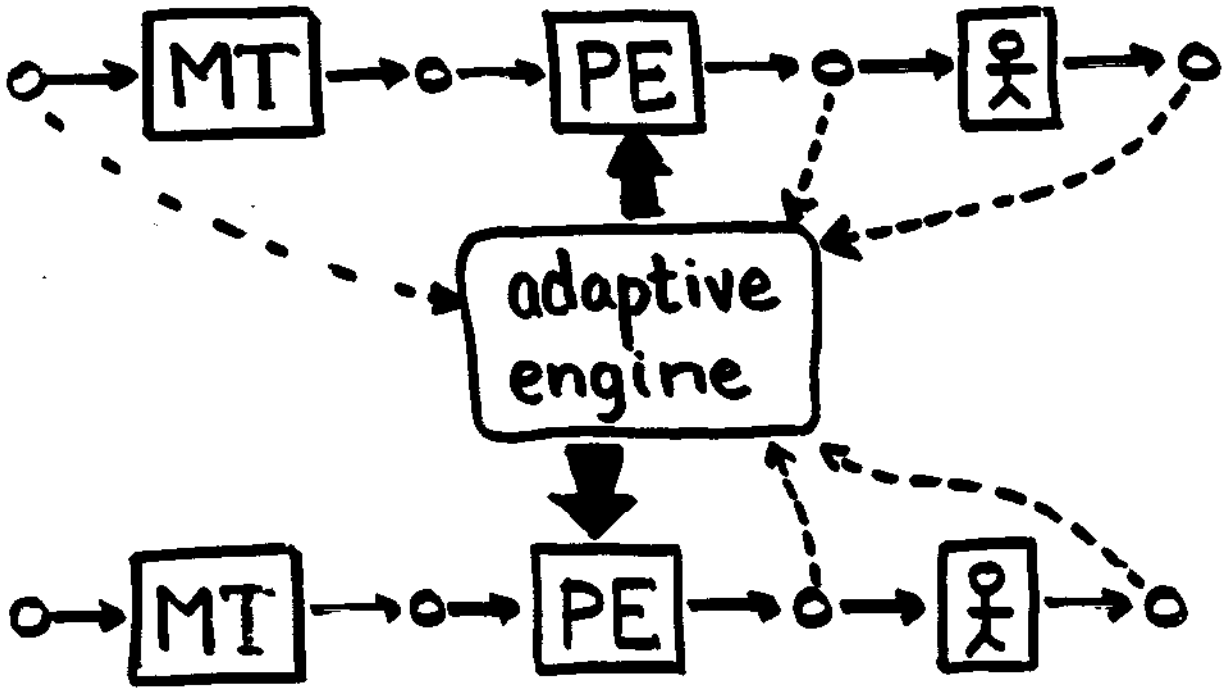


TWO TYPES OF POST-EDITING MODULES

1. ONE-SIZE-FITS-ALL.
   HELPS SOLVE ROUTINE PROBLEMS FACED BY MANY SYSTEMS IN MANY DOMAINS.

2. ADAPTIVE.
   IMPROVES ACCURACY OF A GENERAL-PURPOSE SYSTEM, AS FIELDED IN A PARTICULAR DOMAIN.

ADAPTIVE POST-EDITORS

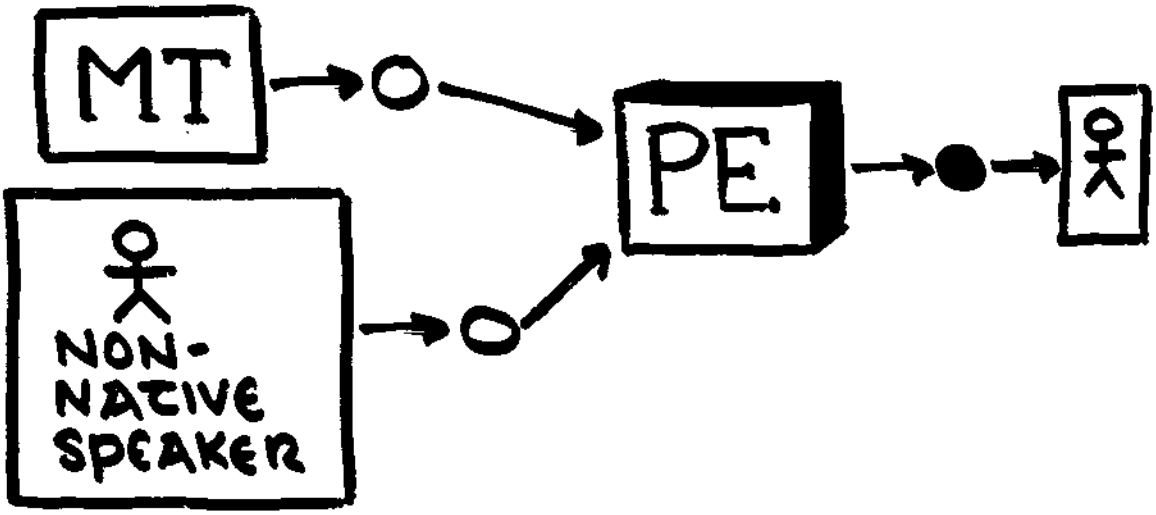


ADAPTATION ALLOWS GENERAL-PURPOSE SYSTEM TO PERFORM IN SPECIAL-PURPOSE SITUATION.

ONE-SIZE-FITS-ALL POST-EDITORS

EXAMPLE:

INSERTING ARTICLES (A/AN/THE) AND PLURALS INTO ENGLISH TEXT.



Stelco Inc. said it plans to shut down three Toronto-area plant', moving their fastener operation. to leased facility in Brantford, Ontario.

company said fastener business "has been under severe cost pressure. for some time." fastener., nut. and bolt', are sold to North American auto market.

company spokesman declined to estimate impact of closure: on earnings. He said new facility will employ 500 of existing 600 employee.. steelmaker employs about 16,000 people.

Stelco said it has option to lease 350,000-square-foot building in Brantford and proposes to spend 24.5 millio Canadian dollar. (US$20.9 million) on facility. three existing plants and their land will be sold.

EVALUATION

TASK DEFINITION:

- TAKE DOCUMENT
- REPLACE ALL "the", "A", "AN" BY DUMMY SYMBOL "XYZ"
- PUT ARTICLES BACK
- HOW MANY DID YOU GET RIGHT?

EASY TO TEST; DON'T NEED FULL MT SYSTEM OR NON-NATIVE TEXTS.

ASSUMPTIONS

- PLURALS ARE GIVEN (AS IN RUSSIAN).
- PLACEHOLDERS ARE GIVEN — NOT YET PREDICTING THE NULL ARTICLE.

EXPERIMENTS CONDUCTED JOINTLY WITH I. CHANDER.

OUR FIRST RESULTS

A RESPECTABLE 65%

THE ALGORITHM

ALWAYS GUESS "THE"!

UPPER BOUNDS

TESTED HUMAN SUBJECT

FULL CONTEXT: 96%

⇒ ARTICLES CONTAIN ALMOST NO INFORMATION (< .25 BITS)

BUT, WE SHOULDN'T EXPECT TO DO BETTER THAN 96%

(A-AN-THE* smoke-filled club)
(A-AN-THE* U. S. discount rate)
(A-AN-THE* utility)
(A-AN-THE* year)
(A-AN-THE* city)
(A-AN-THE* company)
(A-AN-THE* copyright law)
(A-AN-THE* Midwest Financial subsidiary banks
(A-AN-THE* third quarter)
(A-AN-THE* economic slowdown)
(A-AN-THE* benchmark 30-year bond)
(A-AN-THE* food services chain operator)
(A AN-THE* company)
(A-AN-THE* companies)
(A-AN-THE* right)
(A-AN-THE* 30-day simple yield fell)
(A-AN-THE* decision)
(A-AN-THE* government)
(A-AN-THE* White House)
(A-AN-THE* diversification)

LIMITED CONTEXT (NP)

⇒ 80%

```
(("sit" "in") (*A-AN-THE* "smoke-filled" "club") ("," "but"))
(("cut" "in") (*A-AN-THE* "U. S." "discount" "rate") ("in" *A-AN-THE*))
(("Commission" ",") (*A-AN-THE* "utility") ("in" "Paragould"))
(("from" "138") (*A-AN-THE* "year") ("ago" ","))
(("." "And") (*A-AN-THE* "city") ("decided" "to"))
(("\"" "are") (*A-AN-THE* "company") ("'s" "numerous"))
(("to" "enact") (*A-AN-THE* "copyright" "law") ("compatible" "with"))
(("banks" ".") (*A-AN-THE* "Midwest" "Financial" "subsidiary" "banks") ("will" "cor
(("look" "at") (*A-AN-THE* "third" "quarter") ("as" "posting"))
(("landing" "is") (*A-AN-THE* "economic" "slowdown") ("that" "eases"))
(("trading" ".") (*A-AN-THE* "benchmark" "30-year" "bond") ("about" "1/4"))
(("Cara" ",") (*A-AN-THE* "food" "services" "chain" "operator") ("and" "Unicorp"))
(("spokesman" "said") (*A-AN-THE* "company") ("hadn" "'t"))
(("mid-August" ",") (*A-AN-THE* "companies") ("," "through"))
(("Manville" "has") (*A-AN-THE* "right") ("of" "first"))
(("8.14%" ".") (*A-AN-THE* "30-day" "simple" "yield" "fell") ("to" *A-AN-THE*))
(("pay" ",") (*A-AN-THE* "decision") ("that" "could"))
(("would" "owe") (*A-AN-THE* "government") ("after" "his"))
(("leaders" "and") (*A-AN-THE* "White" "House") ("over" *A-AN-THE*))
(("'t" "be") (*A-AN-THE* "diversification") ("." "It"))
```

ND + 2 WORDS TO LEFT AND RIGHT:

⇒ 88%

MEANING...

UNIQUE,
PREVIOUSLY MENTIONED,
GENERIC,
INDIRECT ANAPHORA,
SPORADIC REFERENCE,
NON-REFERRING

⇒ AI-COMPLETE!

⇒ WHO CARES.

LET'S LOOK AT
SOME DATA

E.G., 80 MBYTES
OF ONLINE
WALL STREET
JOURNAL
ARTICLES.

DATA ANALYSIS

- HEAD OF NOUN PHRASE SEEMS IMPORTANT
  - ___ WHITE HOUSE
  - ___ SAND
  - ___ PROBLEMS
- ALSO PREMODIFIERS
  - ___ BIGGEST X
  - ___ NEXT X
- OR BOTH
  - ___ FEDERAL DEFICIT

- AND THE WORD AFTER THE HEAD NOUN

  ___ X AGO

- OR BEFORE

  TRIPLE ___ X

- COMBINATIONS

  REST OF ___ X

  CLEAR ___ WAY

  40% ___ YEAR

  ___ SIGN OF

  . ___ YEAR

  ...

INTEGRATING ALL the PIECES OF EVIDENCE

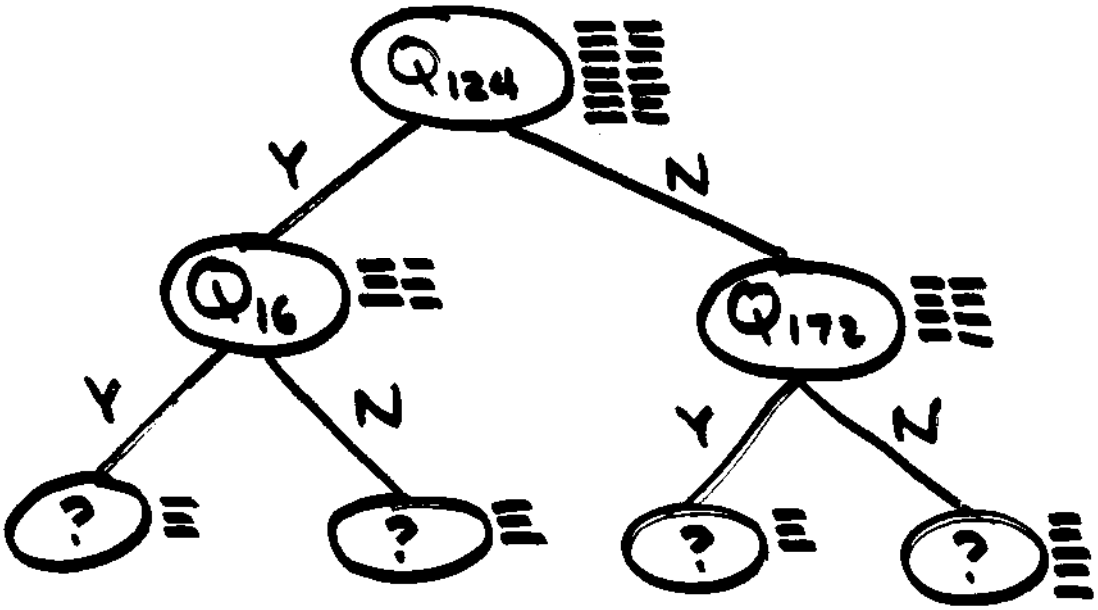
	FREQ	%the
NOUN = "white house"	562	99%
PREMOD ∈ SUPERLATIVE	⋮	⋮
NP-1 = PREPOSITION & NOUN = "YEAR"	2080	73%
NOUN = PLURAL	⋮	⋮
NP-1 = "."		
NP+1 = "AGO"	881	0%

"pay, ___ decision that could"

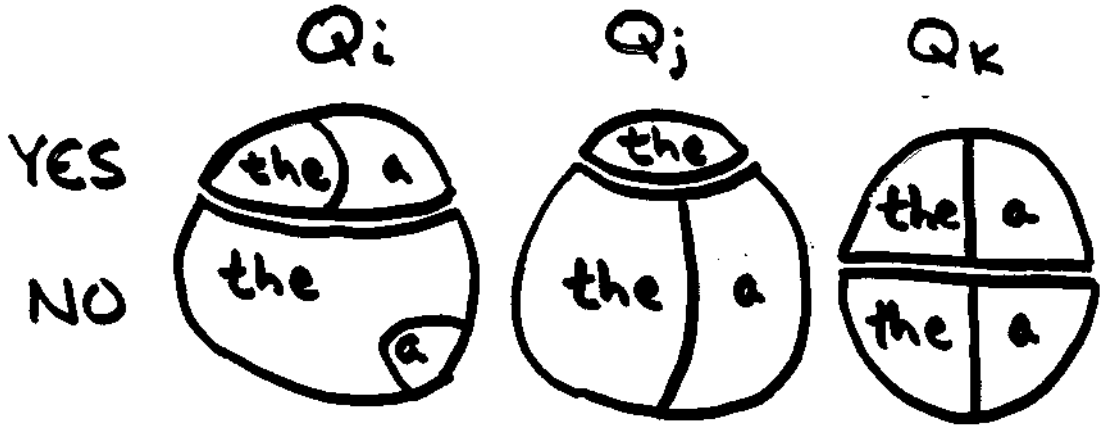
↑
?

DECISION TREE

- LIST OF QUESTIONS (FEATURES)
- LIST OF NOUN phrases (TRAINING INSTANCES)



- CHOOSING A QUESTION



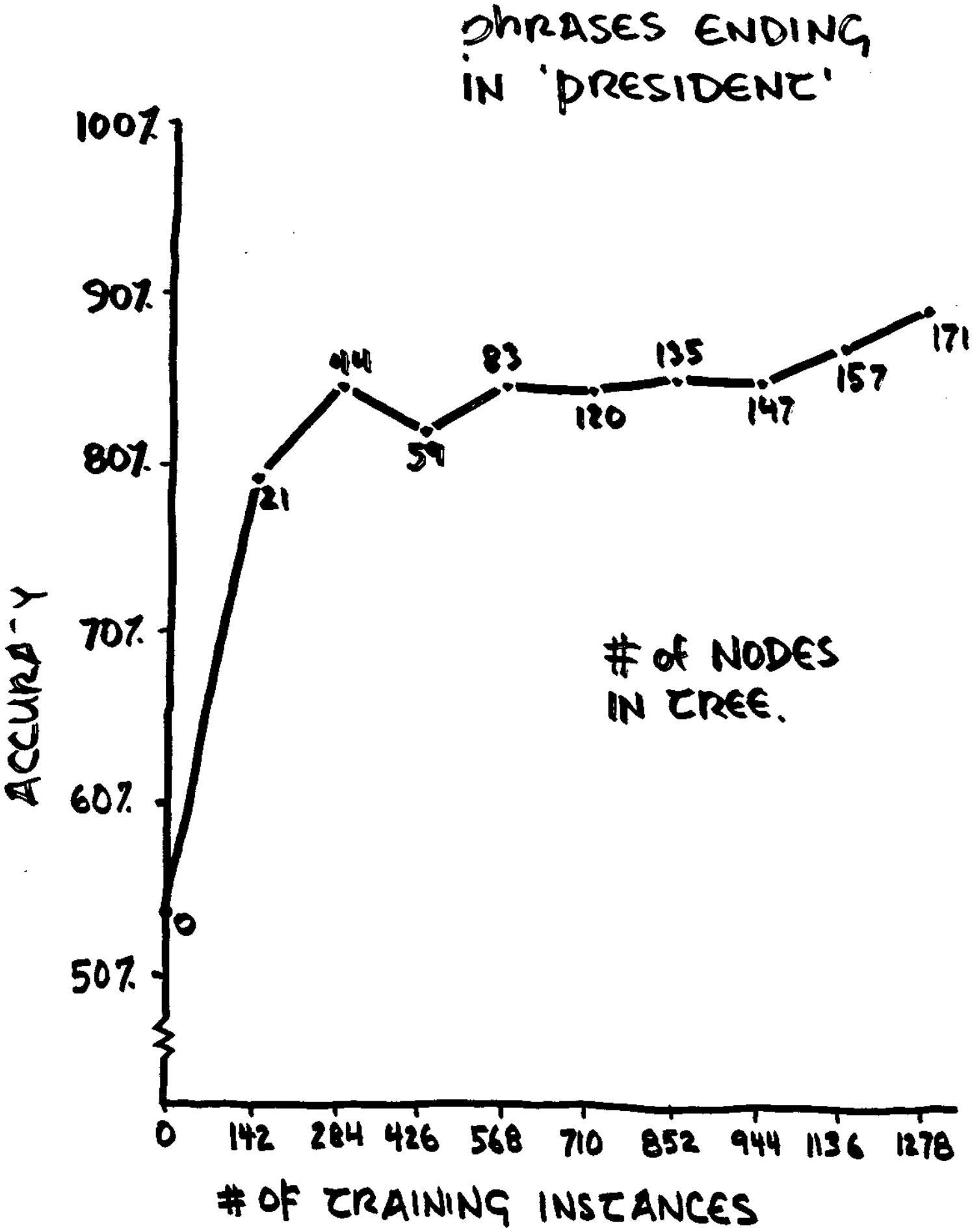
SCALE

400,000 TRAINING EXAMPLES

30,000 FEATURES

NOT FEASIBLE TO DO 400,000 × 30,000 OPERATIONS AT EACH NODE.

BUT, MOST FEATURES ONLY APPLY TO A FRACTION OF INSTANCES. CAN COMPUTE "NO" DISTRIBUTION BY SUBTRACTION.


PHRASES ENDING IN 'PRESIDENT'
100%
90%
80%
70%
60%
50%
ACCURACY
0
21
44
59
83
120
135
147
157
171
# of NODES IN TREE.
0
142
284
426
568
710
852
944
1136
1278
OF TRAINING INSTANCES

For Head Noun: stock rpt3 ----

Train-Pct	Questions	Decision-Nodes	Correct	Incorrect	Pct
10	150	13	1253 (the=1212/1272, a-an=41/178)	197	0.86
20	267	34	1132 (the=1077/1140, a-an=55/149)	157	0.88
30	370	44	985 (the=931/998, a-an=54/130)	143	0.87
40	478	51	868 (the=814/854, a-an=54/113)	99	0.90
50	564	55	721 (the=680/718, a-an=41/88)	85	0.89
60	619	75	581 (the=539/575, a-an=42/70)	64	0.90
70	677	88	433 (the=400/430, a-an=33/54)	51	0.89
80	741	98	285 (the=264/287, a-an=21/36)	38	0.88
90	775	98	149 (the=138/144, a-an=11/18)	13	0.92

For Head Noun: year rpt3 ----

Train-Pct	Questions	Decision-Nodes	Correct	Incorrect	Pct
10	438	45	4625 (the=1511/1819, a-an=3114/3339)	533	0.90
20	755	78	4159 (the=1453/1659, a-an=2706/2926)	426	0.91
30	1059	97	3704 (the=1318/1449, a-an=2386/2563)	308	0.92
40	1333	131	3206 (the=1103/1205, a-an=2103/2234)	233	0.93
50	1589	172	2681 (the=902/983, a-an=1779/1883)	185	0.94
60	1767	190	2148 (the=690/761, a-an=1458/1532)	145	0.94
70	1922	226	1616 (the=505/558, a-an=1111/1162)	104	0.94
80	2079	241	1068 (the=329/363, a-an=739/784)	79	0.93
90	2140	258	532 (the=157/178, a-an=375/396)	42	0.93

For Head Noun: president rpt3 ----

Train-Pct	Questions	Decision-Nodes	Correct	Incorrect	Pct
10	103	21	1001 (the=564/687, a-an=437/596)	282	0.78
20	187	44	962 (the=532/620, a-an=430/520)	178	0.84
30	268	59	816 (the=456/538, a-an=360/460)	182	0.82
40	327	83	714 (the=407/462, a-an=307/393)	141	0.84
50	410	120	598 (the=329/382, a-an=269/331)	115	0.84
60	449	135	484 (the=270/311, a-an=214/259)	86	0.85
70	488	147	365 (the=215/235, a-an=150/193)	63	0.85
80	536	157	248 (the=137/156, a-an=111/129)	37	0.87
90	562	171	127 (the=69/75, a-an=58/68)	16	0.89

DISCUSSION

- POWER OF "KNOW NOTHING" STATISTICS
- BUT, WE ALREADY HAVE KNOWLEDGE IN THE SYSTEM:
  - NOUN PHRASE PARSER
  - WORD CLASSES (PLURAL, SUPERLATIVE, PREPOSITION, VERB...)
- IMPROVEMENTS WILL NOW COME FROM ADDING KNOWLEDGE